Article published in *Spatial Vision, Volume 22, Number 5, 2009 , pp. 397–408*

Efficient Visual Coding and the Predictability of Eye Movements on Natural Movies

Eleonora Vig, Michael Dorr, and Erhardt Barth[*]
Institute for Neuro- and Bioinformatics, University of Lübeck

Abstract

We deal with the analysis of eye movements made on natural movies in free-viewing conditions. Saccades are detected and used to label two classes of movie patches as attended and non-attended. Machine learning techniques are then used to determine how well the two classes can be separated, i.e. how predictable saccade targets are. Although very simple saliency measures are used and then averaged to obtain just one average value per scale, the two classes can be separated with an ROC score of around 0.7, which is higher than previously reported results. Moreover, predictability is analysed for different representations to obtain indirect evidence for the likelihood of a particular representation. It is shown that the predictability correlates with the local intrinsic dimension in a movie.

Keywords: eye movements, visual attention, saliency, efficient visual coding, intrinsic dimension, curvature, machine learning

1 Introduction

The prediction of where people direct their eyes in natural scenes has long been of interest to the vision community (Itti et al., 1998; Krieger et al., 2000; Yarbus, 1967; Zetzsche et al., 1998), and there are also applications in computer vision (Paletta and Rome, 2007) and image compression (Geisler and Perry, 1998). The

[*] Corresponding author. Postal address: Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany. E-mail address: barth@inb.uni-luebeck.de

standard approach to saliency prediction is to design computational models of bottom-up attention (Bruce et al., 2007; Guo et al., 2008; Itti et al., 1998). For each location in the visual field, simple features such as edges, contrast, or colour are stored in feature-activation maps; these maps are then combined to form a saliency map, where each location is assigned a value of how salient or interesting it is to an observer. Finally, the most salient locations are taken to predict human fixations. Only more recently, saliency-based interest operators have been derived from human fixation data by using machine learning techniques (Kienzle et al., 2007).

In this paper, we analyse eye movements made on natural dynamic scenes and, in particular, the extent to which these eye movements can be predicted by using machine learning techniques that operate on different visual representations. We use a data set in which local movie patches have been classified as attended and non-attended locations by measuring eye movements from a large number of people (see Section 2). We then use different representations of the movie patches and machine learning techniques that operate on these representations to learn to distinguish between these two classes. Finally, we argue that a representation that yields better classification results is more likely to be used by the visual system than a representation where the classification results are worse. Additional motivation for our research comes from work on gaze guidance – see Section 5.2.

1.1 Remarks related to this special issue

Some years ago a few vision scientists, among them Terry Caelli, were interested in understanding the spatio-temporal structure of the visual input, and the problems of efficient coding and object recognition, by the use of differential geometry, e.g. Barth et al. (1993). Differential geometry was initially used in vision research to describe invariant properties of object surfaces. Some principles of differential geometry have been generalised later by the use of signal processing and could then be applied to the problem of efficient coding. The insights gained from a fusion of vision research, signal processing, and differential geometry have led to the concept of intrinsic dimension (Zetzsche and Barth, 1990), the relationship to higher-order statistical dependencies (Zetzsche et al., 1993), and to the uniqueness proof (see Section 3.1.2).

2 Experimental setup

We collected data from 54 subjects freely viewing 18 high-resolution movie clips of natural outdoor scenes of about 20 s duration each. Videos had a spatial resolution of 1280 by 720 pixels and a temporal resolution of 29.97 Hz (NTSC HDTV standard); they were displayed at a visual angle of 48 by 27 degrees, so that the maximum spatial frequency of the display was 13.3 cycles per degree.

The commercially available videographic eye tracker EyeLink II produced by SR Research was used to record gaze data at 250 Hz. Trials where more than 5% of all samples were invalid (typically because the subject excessively blinked) were discarded, leaving 844 recordings (between 37 and 52 per video sequence). From these recordings, about 40000 saccades were extracted by a velocity-based two-step procedure (Böhme et al., 2006): to improve noise resilience, gaze velocity had to exceed a high threshold θ_1=137.5 deg/s to initiate saccade detection; saccade onset and offset then were determined by the first samples where gaze velocity rose above or fell below a lower threshold θ_2=17.5 deg/s, respectively. Finally, several checks were performed for biological plausibility: minimal and maximal saccade duration and average and maximal saccade velocity (the reason being that impulse noise might lead to high sample-to-sample velocities).

3 Eye-movement modelling

3.1 Saliency computation

There are numerous methods for deriving measures of saliency for static images – for a review, see Itti (2005) – and fewer methods for movies (Carmi and Itti, 2006; Kienzle et al., 2007; Meur et al., 2007). In our study we opted for a simple measure, because we (a) wanted to introduce minimal bias, and (b) need to make saliency-based predictions on high-resolution movies in real time. It is well known that the visual input needs to change over space and in time in order to attract eye movements (we seem not to like blank walls). Therefore, a simple assumption one can make is that the more the visual signal changes, the more salient it is. The degree to which a spatio-temporal signal changes is qualitatively well described by the intrinsic dimension of the signal and we use this concept as a simple measure of saliency. In case of images, the reasoning would be that edges are more salient than uniform regions and curved edges more salient than straight edges, and there is some evidence for that already (Zetzsche et al., 1998).

In case of movies, features that change over space and in time would be more salient (see below). Such simple assumptions seem reasonable even though they obviously cannot be sufficient to explain the complex nature of eye movements, since top-down factors such as task demands clearly influence where people will look. Nevertheless, we will show that eye movements are quite predictable even with such simple saliency measures. In the following, we first motivate our approach to saliency and then describe how we compute saliency and how we define the feature space in which we then apply the machine learning techniques.

3.1.1 Intrinsic dimension

A movie can be described by a function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$. Given an (open) region Ω, for all $\mathbf{p} = (x, y, t) \in \Omega$, the movie f is said to locally have intrinsic dimension $0, 1, 2$, or 3 (i0D, i1D, i2D, i3D for short) depending on the number of directions in which f changes. More formally, for a given region Ω, we choose a linear subspace $E \subset \mathbb{R}^3$, of highest dimension, such that

$$f(\mathbf{p} + \mathbf{v}) = f(\mathbf{p}) \quad \text{for all } \mathbf{p}, \mathbf{v} \text{ such that } \mathbf{p}, \mathbf{p} + \mathbf{v} \in \Omega,\ \mathbf{v} \in E. \tag{1}$$

The intrinsic dimension denotes the number of degrees of freedom that are locally used and it is relevant to image coding due to the predominance of i0D and i1D regions in natural images (Zetzsche et al., 1993) and the fact that images and movies are fully determined by their i2D regions (see Section 3.1.2).

The intrinsic dimension of f is $3 - \dim(E)$ for movies (and $n - \dim(E)$ for n-dimensional signals).

3.1.2 Curvature and uniqueness of 2D regions

We here briefly discuss the geometrical intuition behind the concept of intrinsic dimension and the uniqueness proof that creates a link between the geometry of the visual input and its information content.

A movie $f(x, y, t)$ can be embedded to obtain a movie surface $\{x, y, t, f\}$. Curvature is then a prominent local geometric invariant of the surface. It is important to first mention the difference between the different types of curvature. In case of $n = 1$ (curves), there is only one type of curvature and curves are either straight or curved. In case of $n = 2$ (here image surfaces), we have mean and Gaussian curvature (two very different concepts). A cylinder has mean curvature but zero Gaussian curvature because it is flat, i.e. it can be mapped to a plane. In case of images, straight edges have mean but no Gaussian curvature, only corners and line ends etc. have Gaussian curvature. In case of $n = 3$ (here movie surfaces), we still have mean and Gaussian curvature but here the "true" curvature is measured by the Riemann tensor – see Barth and Watson (2000).

The tools of differential geometry are quite useful since one can show that the curved (i2D) regions of a surface S and thus of a movie f are unique, i.e. the complete information in the movie is contained in the i2D regions (Barth et al., 1993; Mota and Barth, 2000). We thus have a geometrical proof that i0D and i1D regions of a movie (and of an image) are redundant; we would expect i2D and i3D regions to be more informative and therefore more predictive of eye movements.

3.1.3 Invariants of the structure tensor

The intrinsic dimension can be estimated with different differential methods; here, we will use the differential method based on the structure tensor. More general non-differential approaches are based on the compensation principle (Zetzsche and Barth, 1990) and the Volterra-Wiener theory of nonlinear systems (Krieger et al., 1995).

A straightforward method to estimate the intrinsic dimension is based on the equivalence in Ω of Eq. (1) and the constraint

$$\frac{\partial f}{\partial \mathbf{v}} = 0 \quad \text{for all } \mathbf{v} \in E \ . \tag{2}$$

The subspace E can be estimated as the subspace spanned by the set of unity vectors that minimise the energy functional

$$\mathbf{v} = \int_{\boldsymbol{\Omega}} \left| \frac{\boldsymbol{\partial f}}{\boldsymbol{\partial} \mathbf{v}} \right|^{\mathbf{2}} \mathbf{d\Omega} = \mathbf{v}^{\boldsymbol{T}} \boldsymbol{J} \mathbf{v} \ , \tag{3}$$

where the structure tensor $\mathbf{J}$ is given by

$$\mathbf{J} = \int_{\Omega} \nabla f \otimes \nabla f \ \mathrm{d}\Omega = \int_{\Omega} \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega \ . \tag{4}$$

In the above equation, the symbol $\otimes$ denotes the tensor product, and f_x, f_y, f_t are short notations for the partial derivatives $\partial f/\partial x$, $\partial f/\partial y$, $\partial f/\partial t$. Therefore, E is the eigenspace associated with the smallest eigenvalue of $\mathbf{J}$, and the intrinsic dimension of f corresponds to the rank of $\mathbf{J}$ and may be obtained from the eigenvalue analysis of $\mathbf{J}$ or, equivalently, from its symmetric invariants H, S, and K (Mota et al., 2001):

$$\begin{array}{lll} H & = 1/3\,\mathrm{trace}(J) & = \lambda_1 + \lambda_2 + \lambda_3 \\ S & = |M_{11}| + |M_{22}| + |M_{33}| & = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 \\ K & = |J| & = \lambda_1\lambda_2\lambda_3 \end{array} \tag{5}$$

If $K \neq 0$, the intrinsic dimension is 3; if $S \neq 0$ it is at least 2; and if $H \neq 0$ it is at least 1.

The structure tensor and its eigenvalue analysis are widely used in image processing to estimate orientation and motion (Granlund and Knutsson, 1995; Jähne et al., 1999). The method has later been extended to multiple orientations and multiple motions (Mota et al., 2001; Mota et al., 2004) and to multispectral images (Mota et al., 2006).

3.1.4 Multi-scale features

To extract salient features on different spatio-temporal scales, we constructed a 5-level spatio-temporal isotropic Gaussian multiresolution pyramid, by successively filtering the video sequence with a 5-tap binomial filter and downsampling it by a factor of two in all three dimensions.

To obtain the structure tensor on each scale, we computed the spatio-temporal derivatives using [-1, 0, 1] kernels after smoothing the input sequence with 5-tap binomials in space and time. The lowpass filter that integrates over the neighbourhood Ω was another spatio-temporal 5-tap binomial. The invariants H, S, and K were raised to the power of six, three, and two, respectively, to map them to the same dynamic range (since they contain products of one, two, and three eigenvalues, respectively) (Mota et al., 2001).

To make the H, S, and K saliency measures equally sparse, we defined a threshold θ that was 5 percent of the maximum value in K (the sparsest representation); all pixel values below this threshold were set to zero, so that a certain amount of non-zero pixels remained in K; H and S were also thresholded so that the same number of non-zero pixels remained as well. Note that without this adaptive thresholding, the differences between the representations as reported below are even larger.

3.2 Machine learning and classification

The analysis of eye movements and their predictability was embedded into a classification framework. This required a set of positive, attended, i.e. salient video locations and a set of negative, non-attended locations. The set of salient locations consisted of the landing points of the saccades that preceded each fixation. Additionally, a temporal offset of 70 ms (prior to the fixation) was introduced to account for the delay the human visual system exhibits until it can react to an event in the world. This time lag was determined by cross-correlating the analytical saliency measures based on the geometric invariants with an 'empirical saliency' measure based on real gaze data. The offset with the maximal correlation denoted the average latency in our natural outdoor scenes.

A common method used for the selection of non-attended movie regions is shuffling the scanpaths and the videos among each other. This means that the non-attended locations of a selected video clip are determined using the scanpaths of a different, randomly selected video. This method has been proposed to remove artifacts due to non-uniform scene viewing, such as the central fixation bias (Reinagel and Zador, 1999; Tatler et al., 2005), and in the meantime, assures that the spatial and temporal distribution of the non-attended locations (over all movies) is identical to the distribution of the fixations.

To compensate for the possible inaccuracies inherent in both the human visual system and the eye-tracking devices, local feature-based methods of saliency prediction also need to take into consideration a neighbourhood of a certain size around each saccade target. However, using raw pixel information of a reasonably-sized image patch is computationally not feasible as it results in dimensionality explosion: e.g. for a window with a size of 64×64 pixels (i.e. less than 3×3 degrees) the feature space already has more than 4000 dimensions. We therefore reduced the features to just one value per movie patch and scale. At each fixated and non-fixated location we extracted a square patch, from the current frame, centred around the saccade landing point, and computed the mean energy per pixel (i.e. the square root of averaged squares) in that window. The computations were performed on each spatio-temporal level of the above saliency representations, but in a window whose size was decreased by a factor of two per level in the spatial domain so that the spatial envelope was kept constant. In time, one frame of a lower level corresponded to several frames on the original level, so that the time window was about half a second. The feature energy was thus computed as follows:

$$e_l = \sqrt{\frac{1}{W_l^2} \sum_{i,j}^{W_l} I_l^2(i,j)} \ , \qquad (6)$$

where W_l is the window size and I_l is a video frame of the l-th spatio-temporal level of the saliency pyramid.

For each location, we thus obtained a feature vector containing the energy information (in a specific neighbourhood of the location) on the different spatio-temporal levels of the invariants H, S, and K. Together with the associated labels (attended or non-attended), these vectors comprised the input data of the classifier. Note that the dimensionality of the feature space is determined only by the number of spatio-temporal levels the Gaussian pyramid has (in our case, 5), and is independent of the patch size.

We divided the set of all attended and non-attended locations into a training set that contained the fixations of two-thirds of all subjects, and a test set, containing the fixations of the remaining one-third of the subjects. Gaze data from all 18 movies were used both for training and testing, but for the sake of generality, the fixations of a subject on a given movie were only present in one of the two data sets.

For the classification we used an efficient learning tool from machine learning, namely a soft-margin Support Vector Machine (Schölkopf and Smola, 2002). The kernel function was a Gaussian whose gamma parameter and the penalty term C, which controlled the softness, were found by 5-fold cross-validation. For our analysis, we used the publicly available LIBSVM implementation of SVMs (Chang and Lin, 2001). Hypothesis testing requires a large number of realisations to obtain stable results; we therefore ran the algorithm on 30 random subdivisions of the data into a training and a test set.

4 Results

One of the standard metrics used for saliency prediction is the ROC (Receiver Operator Characteristic) score, also referred to as the Area Under the Curve (AUC) (Tatler et al., 2005). This measure illustrates the discriminating ability of the classifier over the entire range of prediction rates. An ROC score of 0.5 corresponds to a random classifier, whereas for perfect discrimination the score will be 1.0.

Quantitative differences in the distribution of prediction rates (ROC scores) for saliency measures based on the invariants H, S, and K are plotted in Fig. 1. Due to the spatial and temporal uncertainty of both the target locations (see e.g. Becker (1991) for saccadic accuracy) and of the eye tracking system, and taking into consideration the fact that receptive field sizes are in the order of up to several degrees, we ran our prediction algorithm on multiple window sizes, ranging from one degree up to a window that covered the full screen (at the screen boundaries, if required, windows exceeding the dimensions of the screen were clipped to the size of the overlap). Figure 1 shows the results obtained for a patch size of about

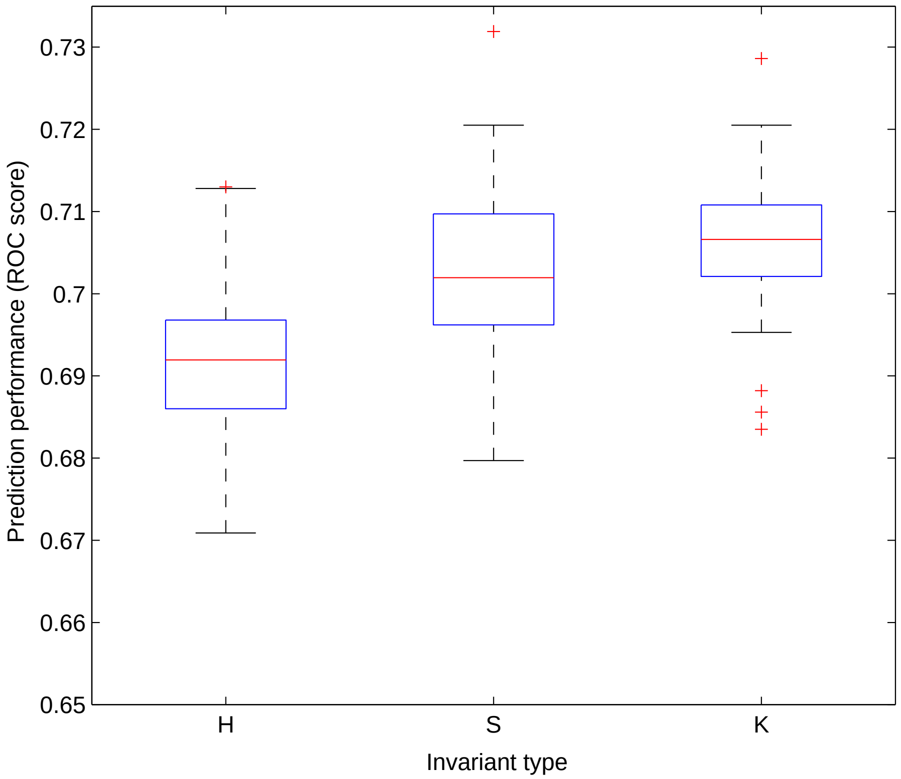

Figure 1: Box plot comparing ROC scores of the invariants H, S, and K over all 30 training and test set realisations (for a window size of about 5 degrees). Horizontal lines indicate the median, the lower and upper quartiles, and the minimum and maximum values of the specific result set. Crosses represent outliers. Comparison of the prediction performances was done by Wilcoxon's signed rank test. Statistical significance obtained at $p << 0.001$ for H–S and H–K, and $p < 0.0314$ for S–K.

5 degrees (128 pixels, that is 4.8 degrees).

For all three invariants H, S, and K, ROC scores were significantly higher than chance, indicating that any of these representations could be a reasonable measure for bottom-up visual saliency. Interestingly, for a large window covering the full screen the prediction rate was still above chance, which means that the global energy measure is predictive by its temporal sequence alone (data not shown).

In case of invariant K, fixated and non-fixated locations were the most discriminable at about 5 degrees window size reaching a median AUC of 0.71, followed by S with a median AUC of 0.70, whereas the worst performing was H having a median ROC score of 0.69. To test the significance of the above ranking assumptions, we performed a paired non-parametric Wilcoxon signed rank test on the ROC results of each invariant pair H–S, H–K, and S–K. We found that in all three cases the differences in saliency predictability were statistically significant (see Fig. 1).

For window sizes up to 5 degrees, invariant K was consistently the most predictive, and it was followed by invariant S, which performed better than invariant H. With larger windows, in the sparser representations such as K the signal-to-noise ratio decreased significantly, hence their performance dropped more suddenly, and no consistent ranking could be observed between the predictability of H and S.

5 Discussion and conclusion

5.1 Predictability of eye movements

In the context of existing models of saliency prediction, the predictability that we obtain is higher than previously reported results on both static and dynamic scenes. The state-of-the-art model of Itti et al. (1998) for still images yields an ROC score of 0.65 (Peters et al., 2005); however, the use of a more complex model incorporating detailed physiological mechanisms (e.g. eccentricity-dependent processing) improves its accuracy to about 0.70 AUC. A more recent study by Kienzle et al. (2007) makes also use of machine learning algorithms to predict where an observer would look in a video. Their learned interest point detector achieves an ROC score of 0.63.

Most of the existing computational models of saliency operate on video patches of a certain size in the spatio-temporal pixel domain. Therefore, their prediction performance depends on the dimensionality of these patches; especially when multiple scales are used, the number of dimensions can easily become very large. In order to overcome the limitations of the existing machine learning methods (i.e. that their performance drops significantly when the dimensionality of the data is high relative to the number of training samples), we have opted for a low-dimensional representation. We reduced the number of dimensions to only one value per movie patch, namely the mean feature energy, computed on each scale. Even though we are aware of the amount of information loss we introduce by such a notable reduction, we argue that this is a good trade-off between the incapabilities of learning algorithms to cope with dimensionality explosion and the information present in the video patches.

However, we do not claim that the low-dimensional representation presented here is optimal. We expect even higher prediction rates for representations that incorporate more information at the cost of a moderate increase in number of dimensions. Indeed, preliminary results obtained on an anisotropic Gaussian pyramid (with 25 instead of 5 levels at which the invariants H, S, and K have been extracted) already indicate a significant improvement in prediction, reaching an AUC of almost 0.80 (compared to the here presented best result of 0.71).

The invariants S and K can account for the fact that moving objects are more salient since they encode spatio-temporal changes. However, we have ignored a number of image attributes such as colour and contrast (divisive normalisation) that certainly do influence the saliency of a movie patch.

Overall, there are a number of reasons why the predictability that we obtain is higher than one would expect from earlier results. First, our input data are high-resolution natural movies as opposed to static images or small movies. Second, it seems that our spatio-temporal features capture the saliency quite well, and third, we use state-of-the-art machine learning techniques in combination with a very large data set.

5.2 Gaze guidance

Our interest in the low-level features contributing to saccade target selection derives from our goal to integrate gaze into future visual communication systems by measuring and guiding eye movements.

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. In an attempt to overcome these limitations, we propose to guide the observer's gaze by designing gaze-contingent interactive displays that change, in real time, the saliency distribution of the visual scene (Barth et al., 2006). To perform gaze guidance, we first predict a limited number of candidate locations that would attract the user's gaze, increase the saliency at the location that is selected for being attended and, at the same time, decrease saliency at all possible distractors.

To derive such transformations between the two classes of fixated and non-fixated image patches (making an image region more or less salient), we first need to find an appropriate feature space with good inter-class separability. Our low-dimensional representation of the energy vectors computed on a number of spatio-temporal scales of the invariants seems to be a good candidate for such a feature space. Using again algorithms from machine learning, rules can now be derived to manipulate local feature energy at candidate locations in such a manner that would alter the saliency of that area[1].

5.3 Efficient coding and intrinsic dimension

The existence of a linear subspace E (see Eq. 2) already implies the redundancy of signals with low intrinsic dimension, because the signal is constant and thus predictable in that subspace. Moreover, the intrinsic dimension of a movie f is closely related to the curvature of the movie surface $\{x, y, t, f\}$. If a signal is curved it is at least i2D (it cannot be i0D or i1D) and therefore signals with intrinsic dimensions less than 2 are redundant, because the curved image regions are unique (see Section 3.1.2).

5.4 Efficient coding and saliency

Our prediction is that the brain may have been able to learn that those regions in the visual input that are curved, and where the intrinsic dimension is higher, are more informative and may have used that to obtain a more efficient code. Previous work has argued in favour of the above argument based on the existence of end-stopped and motion-selective neurons, and on a number of psychophysical results – see for example Zetzsche and Barth (1990). Here we have presented some indirect evidence based on the analysis of eye movements.

[1] For further information see Barth et al. (2006) or visit http://www.gazecom.eu and http://www.inb.uni-luebeck.de/research/itap.

5.5 Conclusions

We have shown that eye movements on natural dynamic scenes are quite predictable even with a very simple low-dimensional feature space. Moreover, we have differentiated among different types of features based on the intrinsic dimension of the visual input and have shown that the predictability correlates with the intrinsic dimension: the higher the intrinsic dimension, the higher the saliency. Finally, we have discussed how the results relate to the geometry of the input signal and the principle of efficient coding. We can conclude that the principle of efficient coding seems to be reflected in the way our observers have decided to direct their gaze.

6 Acknowledgement

Our research has received funding from the European Commission within the project GazeCom (contract no. IST-C-033816) of the 6th Framework Programme. All views expressed herein are those of the authors alone; the European Community is not liable for any use made of the information.

References

Barth, E., Caelli, T. and Zetzsche, C. (1993). Image encoding, labeling, and reconstruction from differential geometry. *CVGIP: Graphical Models and Image Processing* **55**(6), 428–446.

Barth, E., Dorr, M., Böhme, M., Gegenfurtner, K. R. and Martinetz, T. (2006). Guiding the mind's eye: improving communication and vision by external control of the scanpath, in: *Human Vision and Electronic Imaging*, B. E. Rogowitz, T. N. Pappas and S. J. Daly (Eds), Proc. SPIE, Vol. 6057. Invited contribution for a special session on Eye Movements, Visual Search, and Attention: a Tribute to Larry Stark.

Barth, E. and Watson, A. B. (2000). A geometric framework for nonlinear visual coding. *Optics Express* **7**(4), 155–165.

Becker, W. (1991). Saccades, in: *Vision & Visual Dysfunction Vol 8: Eye Movements*, R. H. S. Carpenter (Ed.), pp. 95–137, CRC Press.

Böhme, M., Dorr, M., Krause, C., Martinetz, T. and Barth, E. (2006). Eye movement predictions on natural videos. *Neurocomputing* **69**(16–18), 1996–2004.

Bruce, N. D. B., Loach, D. P. and Tsotsos, J. K. (2007). Visual correlates of fixation selection: a look at the spatial frequency domain, in: *ICIP 2007*, pp. 289–292.

Carmi, R. and Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research* **46**, 4333–4345.

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

Geisler, W. S. and Perry, J. S. (1998). A real-time foveated multiresolution system for low-bandwidth video communication, in: *Human Vision and Electronic Imaging: SPIE Proceedings*, B. E. Rogowitz and T. N. Pappas (Eds), pp. 294–305.

Granlund, G. H. and Knutsson, H. (1995). *Signal Processing for Computer Vision*. Kluwer, Dordrecht.

Guo, C., Ma, Q. and Zhang, L. (2008). Spatio-temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform, in: *Proceedings of IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8.

Itti, L. (2005). Models of bottom-up attention and saliency, in: *Neurobiology of Attention*, L. Itti, G. Rees and J. K. Tsotsos (Eds), pp. 576–582, Elsevier, San Diego, CA.

Itti, L., Koch, C. and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(11), 1254–1259.

Jähne, B., Haußecker, H. and Geißler, P. (Eds) (1999). *Handbook of Computer Vision and Applications*. Academic Press.

Kienzle, W., Schölkopf, B., Wichmann, F. A. and Franz, M. O. (2007). 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)*, pp. 405–414, Springer Verlag, Berlin, Germany.

Krieger, G., Rentschler, I., Hauske, G., Schill, K. and Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision* **13**, 201–214.

Krieger, G., Zetzsche, C. and Barth, E. (1995). Nonlinear image operators for the detection of local intrinsic dimensionality, in: *Proc. IEEE Workshop Nonlinear Signal and Image Processing*, pp. 182–185.

Meur, O. L., Callet, P. L. and Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research* **47**(19), 2483–2498.

Mota, C. and Barth, E. (2000). On the uniqueness of curvature features, in: *Dynamische Perzeption*, G. Baratoff and H. Neumann (Eds), Proceedings in Artificial Intelligence, Vol. 9, pp. 175–178, Infix Verlag, Köln.

Mota, C., Dorr, M., Stuke, I. and Barth, E. (2004). Categorization of Transparent-Motion Patterns Using the Projective Plane. *International Journal of Computer and Information Science* **5**(2), 129–140.

Mota, C., Stuke, I. and Barth, E. (2001). Analytic solutions for multiple motions, in: *Proc. IEEE Int. Conf. Image Processing*, Vol. 2, pp. 917–920, IEEE Signal Processing Soc., Thessaloniki, Greece.

Mota, C., Stuke, I. and Barth, E. (2006). The Intrinsic Dimension of Multispectral Images, in: *MICCAI Workshop on Biophotonics Imaging for Diagnostics and Treatment*, pp. 93–100.

Paletta, L. and Rome, E. (Eds) (2007). *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*. Lecture Notes in Computer Science, Vol. 4840, Springer.

Peters, R. J., Iyer, A., Koch, C. and Itti, L. (2005). Components of Bottom-Up Gaze Allocation in Natural Scenes, in: *Proc. Vision Science Society Annual Meeting (VSS05)*.

Reinagel, P. and Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Comput Neural Syst* **10**, 341–350.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels*. MIT University Press, Cambridge.

Tatler, B. W., Baddeley, R. J. and Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research* **45**, 643–659.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York.

Zetzsche, C. and Barth, E. (1990). Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research* **30**, 1111–1117.

Zetzsche, C., Barth, E. and Wegmann, B. (1993). The importance of intrinsically two-dimensional image features in biological vision and picture coding, in: *Digital Images and Human Vision*, A. B. Watson (Ed.), pp. 109–138, MIT Press, Cambridge.

Zetzsche, C., Schill, K., Deubel, H., Krieger, G., Umkehrer, E. and Beinlich, S. (1998). Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach, in: *From animals to animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, R. Pfeifer et al. (Ed.), Vol. 5, pp. 120–126, MIT Press, Cambridge.