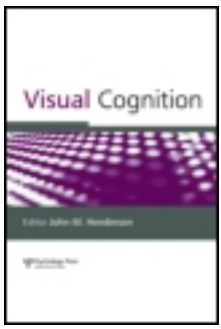

Visual Cognition
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pvis20

Eye movement prediction and variability on natural video data sets

Michael Dorr [a b], Eleonora Vig [a] & Erhardt Barth [a]
[a] Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany
[b] Schepens Eye Research Institute, Department of Ophthalmology, Harvard Medical School, Boston, MA, USA

Available online: 26 Mar 2012

To cite this article: Michael Dorr, Eleonora Vig & Erhardt Barth (2012): Eye movement prediction and variability on natural video data sets, Visual Cognition, DOI:10.1080/13506285.2012.667456

To link to this article: http://dx.doi.org/10.1080/13506285.2012.667456

VISUAL COGNITION, 2012, iFirst, 1–20

Psychology Press
Taylor & Francis Group

Eye movement prediction and variability on natural video data sets

Michael Dorr[1,2], Eleonora Vig[1], and Erhardt Barth[1]

[1]Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany
[2]Schepens Eye Research Institute, Department of Ophthalmology, Harvard Medical School, Boston, MA, USA

We here study the predictability of eye movements when viewing high-resolution natural videos. We use three recently published gaze data sets that contain a wide range of footage, from scenes of almost still-life character to professionally made, fast-paced advertisements and movie trailers. Intersubject gaze variability differs significantly between data sets, with variability being lowest for the professional movies. We then evaluate three state-of-the-art saliency models on these data sets. A model that is based on the invariants of the structure tensor and that combines very generic, sparse video representations with machine learning techniques outperforms the two reference models; performance is further improved for two data sets when the model is extended to a perceptually inspired colour space. Finally, a combined analysis of gaze variability and predictability shows that eye movements on the professionally made movies are the most coherent (due to implicit gaze-guidance strategies of the movie directors), yet the least predictable (presumably due to the frequent cuts). Our results highlight the need for standardized benchmarks to comparatively evaluate eye movement prediction algorithms.

Keywords: Dynamic natural scenes; Eye movement variability; Intrinsic dimension; Saliency; Structure tensor.

Please address all correspondence to Michael Dorr, Schepens Eye Research Institute, Department of Ophthalmology, Harvard Medical School, 20 Staniford St., Boston, MA 02114, USA. E-mail: michael.dorr@schepens.harvard.edu

Our research has received funding from the European Commission within the GazeCom project (IST-C-033816) of the FP6, and was further supported by NIH grants EY018664 and EY019281. All views herein are those of the authors alone; the European Commission is not liable for any use made of the information. The GazeCom data set has been collected in Karl Gegenfurtner's lab at the University of Giessen. We thank the two anonymous reviewers for their comments.

http://www.psypress.com/viscog http://dx.doi.org/10.1080/13506285.2012.667456

Humans constantly move their eyes to sample the visual input with the high-resolution centre of the retina, and where they look is tightly linked to the aspects of a scene they are consciously processing. Despite the obvious importance of oculomotor guidance strategies, we still do not have a full understanding of how we select upcoming saccade targets from the rich visual input in everyday vision. For very simple stimuli, such as coloured letter arrays often used in visual search experiments, the input may adequately be described in terms of simple visual features such as colour or corners versus junctions, and a singleton in one feature dimension may immediately "pop out" and capture attention and gaze (Wolfe, 1998).

Following this observation, a vast body of research has modelled saccade target selection in more complex visual scenes based on low-level "saliency" (Bruce & Tsotsos, 2009; Gao & Vasconcelos, 2009; Itti & Baldi, 2006; Itti, Koch, & Niebur, 1998; Judd, Ehinger, Durand, & Torralba, 2009; Kienzle, Franz, Schölkopf, & Wichmann, 2009; Le Meur, Le Callet, Barba, & Thoreau, 2006; Tatler, Baddeley, & Vincent, 2006; Zhang, Tong, Marks, Shan, & Cottrell, 2008). Typically, a number of easily computable visual features such as luminance, contrast, or orientation are extracted for every location, and those locations that differ in one or more of these features from their neighbourhood are assigned higher saliency values. Then, saccade targets can be picked from the resulting saliency map, e.g., by an iterative winner-takes-all operation with subsequent inhibition of previously fixated locations.

These models are appealing for at least two reasons. First, it has been shown repeatedly that image regions with higher saliency are also more likely to be fixated (Parkhurst & Niebur, 2003; Reinagel & Zador, 1999); in other words, eye movement behaviour can be predicted (to some extent) based on computationally tractable low-level image features. This is particularly true for free viewing experiments, i.e., in the absence of high-level tasks (Einhäuser, Rutishauser, & Koch, 2008). It is important to note, however, that even the best models under the best of circumstances are far from perfect predictors, and the interobserver agreement as a likely upper ceiling for prediction performance is also significantly lower than 100% (Peters, Iyer, Itti, & Koch, 2005).

The second reason for the appeal of saliency models is their conceptual proximity to the biology of the primate visual system, where dedicated circuits in the early stages encode low-level features such as oriented edges, corners, motion, or colour (Wandell, 1995).

The majority of research on saliency has used static images as stimuli. While static real-world scenes are undoubtedly a much more realistic and relevant input to the visual system than traditional, impoverished psychophysical stimuli, they still crucially lack dynamic information. Whenever something or someone moved in the bushes, our ancestors probably were

well advised to quickly determine the source of this motion, whereas static image regions could be examined at leisure. Indeed, eye movements when viewing static images become more idiosyncratic than on dynamic content after only a few seconds of viewing time (Dorr, Martinetz, Gegenfurtner, & Barth, 2010).

Therefore, the number of studies that model eye movement behaviour using dynamic natural scenes has been growing recently, comprising bioinspired, information-theoretic, machine-learning, and signal processing based approaches (Guo & Zhang, 2010; Itti, 2005; Kienzle, Schölkopf, Wichmann, & Franz, 2007; Mahadevan & Vasconcelos, 2010; Le Meur et al., 2006; Mital, Smith, Hill, & Henderson, 2011; Tatler, Baddeley, & Gilchrist, 2005; Vig, Dorr, & Barth, 2009; Zhang, Tong, & Cottrell, 2009). Unfortunately, the problem of what constitutes "naturalness" of a scene is only exacerbated in the dynamic case relative to static images. The space of possible image sequences is even larger due to the combinatorial explosion of adding another dimension. Furthermore, where databases of hundreds of calibrated images exist that can be displayed in quick succession, movies are typically longer and harder to come by, so fewer can be presented in one experimental session. Several studies have used professionally produced stimuli such as Hollywood movies or TV shows, but these are often carefully arranged in order to direct attention to specific objects of interest, and contain cuts that do not typically occur under normal viewing conditions.

The choice of the optimal stimulus set notwithstanding, the concept of saliency has been also challenged fundamentally; for a recent review, see Tatler, Hayhoe, Land, and Ballard (2011). Even though prediction performance is significantly above chance in virtually all studies, current low-level models seem to hit a ceiling at an ROC score of around 0.7, so that a large part of eye movement selection remains unexplained; when high-level task demands or socially meaningful stimuli such as faces are introduced, prediction performance gets even worse (Einhäuser, Rutishauser, & Koch, 2008; Einhäuser, Spain, & Perona, 2008). Because blank surfaces contain no information at all, but reacting to sudden looming onsets might be crucial for survival, it is intuitively plausible that image features *can* guide attention. However, it simply may be that under many conditions, saliency is not causal for eye movements, but merely correlated with the presence of semantically meaningful objects. Without low-level properties such as edges or a texture gradient, objects cannot be distinguished from their surround, but the magnitude of these properties—once above a certain threshold—may be less crucial.

More recent, complex saliency models may also suffer from too many free parameters that were introduced in the attempt to cover all possible factors or low-level image features that might potentially influence saccade target selection. On the other end of the complexity spectrum, Vig et al. (2009;

Vig, Dorr, Martinetz, & Barth, in press) used the geometric invariants of the structure tensor to predict eye movements, and outperformed state-of-the-art saliency models. The invariants simply encode the amount of local change in a signal and thus yield very generic video representations. Based on these representations, prediction performance was improved upon even further by employing machine learning algorithms.

Recent results by Vig, Dorr, Martinetz, and Barth (2011) also indicate that the contribution of saliency to fixation selection is not entirely straightforward even in naturalistic videos. These authors cross-correlated in time analytical dynamic saliency maps with "empirical" saliency maps that were based on observed eye movements. For less natural footage, such as video games or professionally cut material, the peak of the correlation function occurred at a shift of about 133 ms between a dynamic event and a gaze response, similar to classical laboratory experiments where observers can react to unpredictable events (such as the sudden appearance of a saccade target marker) only with a latency of 150–250 ms. In more natural, uncut outdoor scenes, however, the peak of the correlation function occurred at around 0 ms, which implies that observers have an internal model of natural environments that allows them to predict where informative image regions will be after the next saccade, and that truly unpredictable events are rare in the real world. Predictive gaze behaviour becomes even more prominently visible when subjects truly interact with an environment, e.g., in everyday tasks such as tea- or sandwich making, or sports (Land & Hayhoe, 2001; Land & McLeod, 2000; Land, Mennie, & Rusted, 1999).

Eye movement behaviour is further affected by oculomotor constraints and peripheral resolution limits. Horizontal saccades are more frequent than vertical ones, which in turn are more frequent than oblique saccades, independent of the visual input, and the amplitude distribution is heavily skewed towards medium-sized saccades (Tatler & Vincent, 2009; see also Foulsham and Kingstone, forthcoming).

Even if the contribution of saliency to saccade target selection is mainly of a correlative rather than causal nature, saliency models can still be of practical value. For example, more than a million images and three million video frames are uploaded every minute to two particular web sites alone, and to evaluate them all with human observers is impossible. Knowledge of where observers will look, however, can be beneficial for, e.g., video compression (Itti, 2004; Li, Qin, & Itti, 2011; Nyström & Holmqvist, 2010), or determining what message will ultimately be conveyed by visual material.

In this paper, we shall investigate three important aspects of modelling eye movements in dynamic natural scenes. First, we will look at several recently made available data sets of eye movements that were recorded while subjects watched high-resolution videos, and we will compare the interobserver

agreement of gaze patterns in each of these data sets. Because of the vast dimensionality of the space of natural movies, even large video collections cannot be representative, and consequently we find large differences between the data sets. These differences indicate that a fair comparison of eye movement prediction methods requires standardized benchmark protocols, similar to the ones used in the machine learning and computer vision communities (e.g., Everingham, van Gool, Williams, Winn, & Zisserman, 2010; Laptev, Marszalek, Schmid, & Rozenfeld, 2008).

As a first step towards such a goal, we evaluate three classes of state-of-the-art saliency models on all data sets. The range of eye movement predictability for individual movies spans cases where performance is essentially random to almost perfect performance. We find that even though prediction performance generally is highly correlated for the different models, some models still consistently outperform others.

Finally, we investigate the relationship of eye movements variability and predictability. Interobserver agreement of eye movement patterns has been studied for different stimulus types (Dorr et al., 2010; Mital et al., 2011; Peters et al., 2005), and some types such as engaging Hollywood movie trailers evoke particularly high agreement (i.e., low variability). This could be due to two different effects: Strong low-level saliency, such as rapid motion or a very shallow depth of field, may attract attention; alternatively, gaze coherence may be induced by the semantic meaning of a scene. In the former, but not in the latter case, we would expect a positive correlation of eye movement coherence and feature-based prediction performance.

METHODS

Eye movement data sets

We analysed three recent, publicly available data sets of high-resolution video material with accompanying eye movement recordings. Our own data set (Dorr et al., 2010) comprises 18 videos of about 20 s duration each at HDTV resolution (1280×720 pixels, 29.97 frames per second). Clips were taken in outdoor settings around Lübeck with a camera that was fixed in all but two recordings (where the camera followed animals). Fifty-four observers watched these clips while their eye movements were tracked with an EyeLink II system running at 250 Hz.

The VAGBA database (Li et al., 2011) contains data from 14 observers watching 50 stimuli, in- and outdoor scenes captured with a static camera, of 10 s duration each while their gaze was recorded at 240 Hz with an iScan eyetracker. For our analysis, we downsampled videos from their original 1920×1080 pixels resolution (30 fps) to 1280×720 pixels.

Video stimuli in the DIEM database (Mital et al., 2011) have different resolution and aspect ratio, but a fair comparison between stimuli requires a common coordinate system. Therefore, only videos that differed no more than 5% in either dimension from a resolution of 1280 pixels horizontally and 720 pixels vertically were used in our analysis and centred in a 1280×720 frame; border pixels were taken from the nearest stimulus pixel to avoid hard image transients. From this set of 47 movies, two were excluded from the analysis because they had less than 40 eye movement recordings (mean 69.6, $\sigma = 48.5$ recordings); for the variability analysis, a constant number of training recordings was required. Two further movies were discarded because their video content was contained in the database twice (with and without soundtrack). At 30 frames per second, videos had an average duration of 102.6 s ($\sigma = 39.6$ s). In the original recordings, eye movements were monitored at 1000 Hz with an EyeLink 1000 tracker, but the publicly available data was reduced to one sample per video frame (i.e., at 30 Hz). Contrary to the other two data sets, DIEM contains professionally cut material such as movie trailers, documentary footage, TV shows, etc.

Gaze analysis

Saccadic landing points were used to analyse those image features that triggered an eye movement. The use of saccadic landing points rather than raw gaze samples potentially ignores the effect of different fixation durations (e.g., see Nuthmann & Henderson, forthcoming), but is computationally more tractable. The DIEM and VAGBA data sets already contain information on detected saccades. For the GazeCom data set, saccades were extracted using a dual-threshold velocity-based algorithm (Dorr et al., 2010) where raw gaze velocity had to exceed a fairly high threshold (150 deg/s) first for noise robustness, and on- and offsets were then detected using a lower threshold (19 deg/s). Because the different temporal resolution of the three data sets could influence the analysis of gaze variability, raw data were lowpass-filtered and downsampled to one sample per video frame in VAGBA and GazeCom to match the resolution of DIEM.

Eye movement prediction

For the prediction of eye movements, we used three different saliency models. The source code for the Itti (Itti & Baldi, 2006; Itti et al., 1998) and the SUNDAy (Zhang et al., 2009) saliency models is publicly available (at http://ilab.usc.edu/toolkit and http://mplab.ucsd.edu/~nick/NMPT, respectively) and we will only briefly describe their function here. Default parameters were used for these models. Our own model, which is based on the geometrical invariants of the structure tensor, has previously been

described for greyscale videos in (Vig et al., 2009, in press); we here use it to operate in two different colour spaces.

Itti Maxnorm model

Itti's Maxnorm model is an implementation of the classical saliency map of Koch and Ullman (Koch & Ullman, 1985). In its first, preattentive phase, various low-level features, such as intensity, orientation, colour, flicker, and motion, are extracted in parallel on multiple scales. Then, from the combination of centre-surround feature maps (so-called conspicuity maps), a master saliency map is generated, on which biological mechanisms, such as winner-takes-all competition and inhibition of return, bias the selection of the next location to be fixated. Different fusing schemes of the individual feature-specific saliency maps have been proposed. Here we use the Maxnorm normalization scheme, in which the fusion of conspicuity maps is based on normalized summation.

SUNDAy model

In SUNDAy, saliency is computed as the self-information of low-level visual features. In the context of eye movement prediction, self-information quantifies the intuitive assumption that novel items draw human gaze. In the formulation of (Zhang et al., 2009), self-information and the probability of the presence of a visual feature are inversely proportional, i.e., rarer features are more informative. As opposed to existing approaches, feature statistics are here learned from a large collection of natural videos, and are not based on movies in the three data sets analysed here.

Geometric invariants of the structure tensor

Greyscale videos are signals that can change in any of their three dimensions. Due to the spatiotemporal correlations in natural scenes, however, natural videos often locally change little or at all: Neighbouring pixels (in space and time) typically have the same or similar intensity. The *intrinsic dimensionality* of an image region (Zetzsche & Barth, 1990) formalizes and quantifies this observation and describes the number of degrees of freedom that are used locally (Figure 1). It can be shown that image regions with an intrinsic dimension (*i*n*D*) of less than two, such as uniform regions and straight edges, are redundant (Barth, Caelli, & Zetzsche, 1993; Mota & Barth, 2000), and that regions with higher *iD* are not only more informative, but also less frequent. As a consequence, encoding only image regions of higher *iD* is an example of efficient sparse codes, which might be a generic mechanism in neural systems (Barth, Dorr, Vig, Pomarjanschi, & Mota, 2010; Field, 1987; Olshausen & Field, 1996).

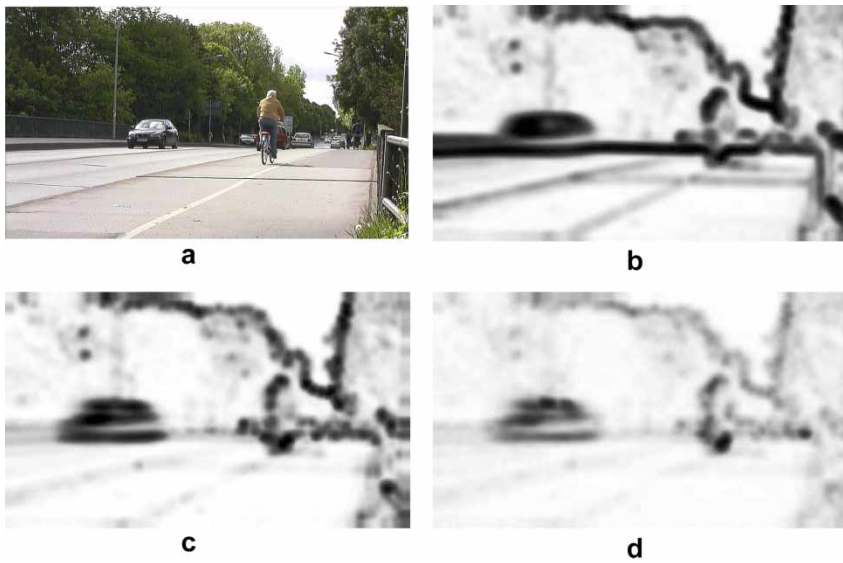

Figure 1. Example of geometrical invariants, which denote the locally used number of degrees n of freedom of a signal, i.e., the *intrinsic dimensionality* inD. See main text for an in-depth description. Contrast inverted and enhanced for better legibility, with same parameters for b–d. (a) Stillshot from original movie. (b) Invariant H, which represents regions with $iD \geq 1$. Regions with $iD = 0$, i.e., uniform regions, are suppressed. (c) Invariant S ($iD \geq 2$) that encodes static corners. Note the reduced response to straight edges. (d) Invariant K ($i3D$) that encodes spatiotemporal corners, i.e., regions where the signal changes in all three dimensions. To view this figure in colour, please see the online issue of the Journal.

The intrinsic dimensionality of an image region can be computed by several methods, and we will here describe the approach based on the structure tensor, which is a common tool in computer vision (Jähne, 1999). For every pixel, the partial spatiotemporal derivatives are taken (typically, after an appropriate low-pass filter operation to increase noise robustness; in this paper, we used a five-tap binomial filter), and the products of all possible pairs of the derivatives f_x, f_y, f_t are computed. Because natural videos contain not only greyscale information, but q colour channels, this first requires the definition of a suitable scalar product for vectorial pixels, and we here use

$$\vec{y} \cdot \vec{z} = \sum_{k=1}^{q} a_k y_k z_k$$

for vectors

$$\vec{y} = \left(y_q, ..., y_q\right), \quad \vec{z} = \left(z_q, ..., z_q\right)$$

The a_k are weights that can be used to adjust the relative importance of individual colour channels. We here perform the saliency computations in two different colour spaces that both have one luminance channel and two channels of colour opponency information. The first, straightforward choice is the canonical colour space for video files, $Y'C_bC_r$; the second, *Lab*, is optimized towards the human visual system and is perceptually uniform (Poynton, 2003). In both colour spaces, the two colour opponency channels have much smaller dynamic range than the luminance channel, and larger a_k for these channels can compensate for this difference; we here used the inverse of the channel's mean energy.

Finally, the products of derivatives are smoothed with a spatiotemporal low-pass filter ω, here chosen to be five-tap spatiotemporal binomials, and we arrive at the structure tensor **J**:

$$J = \omega(x, y, t) * \begin{pmatrix} ||\vec{f}_x||^2 & \vec{f}_x \cdot \vec{f}_y & \vec{f}_x \cdot \vec{f}_t \\ \vec{f}_x \cdot \vec{f}_y & ||\vec{f}_y||^2 & \vec{f}_y \cdot \vec{f}_t \\ \vec{f}_x \cdot \vec{f}_t & \vec{f}_y \cdot \vec{f}_t & ||\vec{f}_t||^2 \end{pmatrix}$$

The intrinsic dimensionality can now be computed based on the *eigenvalues* λ_i of **J**, or, alternatively, on its minors. Those pixels where the geometric invariants H, S, or K are nonzero denote regions of at least iD one, two, or three, respectively:

$$H = \lambda_1 + \lambda_2 + \lambda_3 (iD \geq 1)$$
$$S = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3 (iD \geq 2)$$
$$K = \lambda_1\lambda_2\lambda_3 (iD = 3)$$

Put in less mathematical and more intuitive terms, uniform regions that do not change in any direction are intrinsically zero-dimensional ($i0D$, no response in H, S, or K); at a static edge, the signal changes only in one direction orthogonal to the edge and is thus $i1D$ (response in H), and static corners are $i2D$ (response in H and S). Temporally transient corners change in all possible directions and are $i3D$ (response in H, S, and K). It should be noted, however, that these correspondences are not necessarily perfect in complex natural scenes because of geometric distortions, camera noise, and small camera motions (Figure 1).

The bandwidth of the derivatives constrains feature extraction to a narrowband spatiotemporal scale, but natural scenes are characterized by a very wide distribution of spatiotemporal frequencies. We therefore computed the invariants H, S, and K on each level of a spatiotemporal Gaussian pyramid that was created by iteratively low-pass filtering and downsampling the input video. For the DIEM and the GazeCom data sets, we created a pyramid with five spatial and five temporal scales; for the shorter movie clips

in the VAGBA data set, the temporal border effects would have rendered most of the data unusable, so that we computed a five-by-three pyramid only.

Even the invariants computed on only a single scale can be used as the equivalent of a saliency map. However, better performance can be achieved by learning the structural differences of fixated and nonfixated image patches using machine learning techniques. In principle, one could directly feed the pixels of image patches (or movie subvolumes) into an automatic classifier, where patches that were fixated by human observers are assigned a positive class label, and randomly selected, nonfixated patches are assigned a negative label. In practice, however, the curse of dimensionality renders this approach impossible, because the required number of dimensions grows with the number of pixels. We therefore pooled the feature energy $e_{s,t}$ on each spatiotemporal scale s,t in a spatial 2.4×2.4 degree (for GazeCom and VAGBA) or 1.2×1.2 degree (for DIEM) neighbourhood around fixation; these optimal window sizes relate to the dominant scale and the amount of clutter in each data set and were thus inferred separately by cross-validation:

$$e_{s,t} = \sqrt{\frac{1}{W_s H_s} \sum_{i=-W_s/2}^{W_s/2} \sum_{i=-H_s/2}^{H_s/2} I_{s,t}^2(x_s - i, y_s - j)}$$

with W_s, H_s the window size in pixels at spatial scale s, t the temporal scale, and I the feature map (i.e., one of the geometrical invariants H, S, or K).

Furthermore, randomly selecting negative examples overestimates predictability because of the centre bias of both the photographer and the subjects. We shuffled scanpaths from other movies to generate negative examples, thereby maintaining a constant spatiotemporal distribution of fixations in the two classes.

Finally, the set of feature energies on all spatiotemporal scales at video location p_i together with a class label l_i

$$(e_{0,0}, e_{0,1} \ldots, e_{S-1,T-1}, l_i)$$

was fed into a standard classification algorithm. We used a soft-margin Support Vector Machine (Chang & Lin, 2001) that fits the optimally separating hyperplane through the training data in a feature space whose dimensionality was reduced to the number of spatiotemporal scales by the energy pooling. Optimal parameters were found by cross-validation on a training set (two thirds of the data for the larger sets VAGBA and DIEM; all movies but one for GazeCom); prediction performance is reported as the area under the curve (AUC) for the receiver-operating characteristic for the remaining, *test* data that was previously unseen by the classifier. This

procedure was repeated until each movie had been part of the test set once and thus received a prediction score.

Variability analysis

Unless observers look at the same set of locations in exactly the same order, it is difficult to assess the (dis-)similarity of two scanpaths. Depending on the task at hand and the visual stimulus, the minimum spatiotemporal distance or change in order of fixations that gives rise to a meaningful difference may not be the same. We here use the Normalized Scanpath Saliency (Peters et al., 2005) extended to the dynamic domain (Dorr et al., 2010), which is based on a superposition of spatiotemporal Gaussians at each gaze sample (note that for this analysis, raw gaze samples instead of saccades are used).

For each movie, N observers $i = 1,...,N$ were chosen randomly as a training set, and a spatiotemporal Gaussian (s_x,$s_y = 64$ pixels, $s_t = 33$ ms) centred around each of their gaze samples $\vec{x}_i^j (j = 1, ..., M_i)$ was placed in a fixation map F:

$$F(\vec{x}) = \sum_{i=1}^{N} \sum_{j=1}^{M_i} G_i^j(\vec{x})$$

with

$$G_i^j(\vec{x}) = e^{\frac{\left(\vec{x}-\vec{x}_i^j\right)^2}{2\left(\sigma_x^2+\sigma_y^2+\sigma_t^2\right)}}$$

F was subsequently normalized to zero mean and unit standard deviation to obtain an NSS map N, and the NSS score was computed as the mean of the NSS map values at the gaze samples of a test observer k,

$$NSS = \frac{1}{M_k} \sum_{j=1}^{M_k} N\left(\vec{x}_k^j\right)$$

This was repeated for 500 randomly drawn realizations of subjects into training and test sets.

RESULTS

Prediction performance for the different saliency models on all data sets is listed in Table 1, where subscripts Y and L denote the colour spaces $Y'C_bC_r$ and Lab, respectively. Several patterns clearly emerge. First, the predictor based on the geometric invariant K performs best on all three data sets; a paired Wilcoxon's test over individual movies shows the difference to be statistically significant at $p < .01$ for all intermodel comparisons except for S (both S_Y and S_L) and SUNDAy, which are not significantly different.

TABLE 1
ROC scores for the different saliency models on the three eye movement data sets

Model	*GazeCom data set*	*VAGBA data set*	*DIEM data set*
H_Y	0.661	0.719	0.630
H_L	0.673	0.733	0.623
S_Y	0.670	0.730	0.640
S_L	0.675	0.735	0.639
K_Y	0.687	0.778	0.653
K_L	0.689	0.781	0.646
Itti	0.623	0.704	0.628
SUNDAy	0.640	0.746	0.645

The predictor based on the geometric invariant K, regardless of colour space, consistently outperforms the other predictors; performance is poorest for the DIEM data set.

Second, the choice of colour space has a significant impact on predictability. Even though $Y'C_bC_r$ and *Lab* are conceptually similar, the perceptually uniform *Lab* space improves results for GazeCom and VAGBA, but decreases predictive power for DIEM. This effect is in line with the overall pattern that performance differs significantly for the three data sets. On the professionally made footage of the DIEM data set, predictability based on low-level features is worst. Better prediction is achieved for the GazeCom and VAGBA data sets, where movies were recorded with a static camera in essentially random in- and outdoor scenes.

In Figure 2, we plot prediction performance for all movies from the three data sets combined for selected pairs of saliency models. We do not show all possible pairs of models because the invariants H, S, and K form sub- and superset relationships, respectively. For example, S_Y responds to intrinsically at least two-dimensional features and therefore is a superset of K_Y, which responds to *i3D* features; as a consequence, their correlation is almost perfect (correlation coefficient .923).

Despite different absolute performance for the whole data set, performance on the individual movies is correlated also across different classes of saliency models. The Itti model is modestly correlated with, e.g., SUNDAy and K (correlation coefficient .491 and .513, respectively), but SUNDAy and K are highly correlated with a coefficient of .901.

Empirical cumulative distribution functions of the Normalized Scanpath Saliency in the three data sets are shown in Figure 3. Intersubject gaze variability is highest, i.e., NSS is lowest, in the GazeCom data set that contains both some relatively inanimate as well as very cluttered scenes, e.g., a boat in the distance or a busy roundabout seen from a church tower. Videos in the VAGBA set were also recorded with a static camera in everyday situations, but occasionally show people interacting with, i.e.,

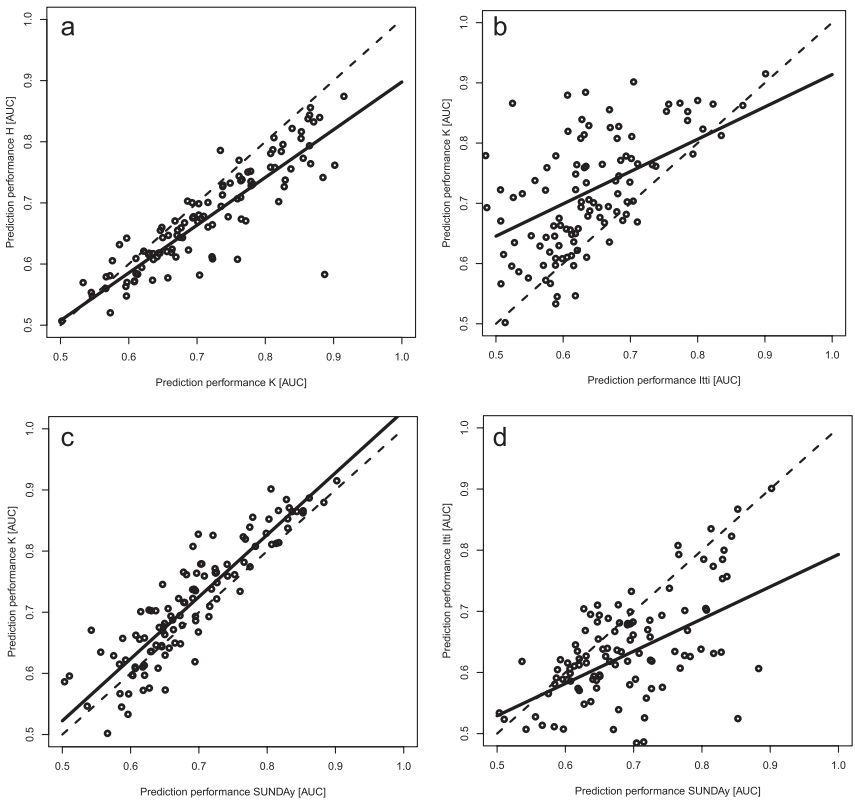

Figure 2. Prediction performance of the three saliency models on individual movies is highly correlated. Dashed line indicates equal performance, solid line shows best linear fit of the data. (a) Invariant K vs. invariant H. (b) Itti model vs. invariant K. (c) SUNDAy model vs. invariant K. (d) SUNDAy model vs. Itti model.

posing in front of, the camera, which leads to a very high gaze coherence. On average, the highest gaze coherence can be found in the DIEM data set with its professionally made advertisements and Hollywood-style material. All pairwise comparisons with a Kolmogorov-Smirnov test show statistically significant differences at $p < .001$.

Finally, Figure 4 shows a scatterplot of gaze predictability for individual movies, based on K_Y, against gaze coherence. The solid lines depict separate linear fits for the three data sets. For the two non-professional data sets (GazeCom and VAGBA, black and red symbols), gaze coherence is positively correlated with predictability, i.e., less variable eye movements are easier to predict based on low-level image features alone. Removal of those six movies from the VAGBA data set that show faces directly turned towards and interacting with the camera, and which have the highest NSS

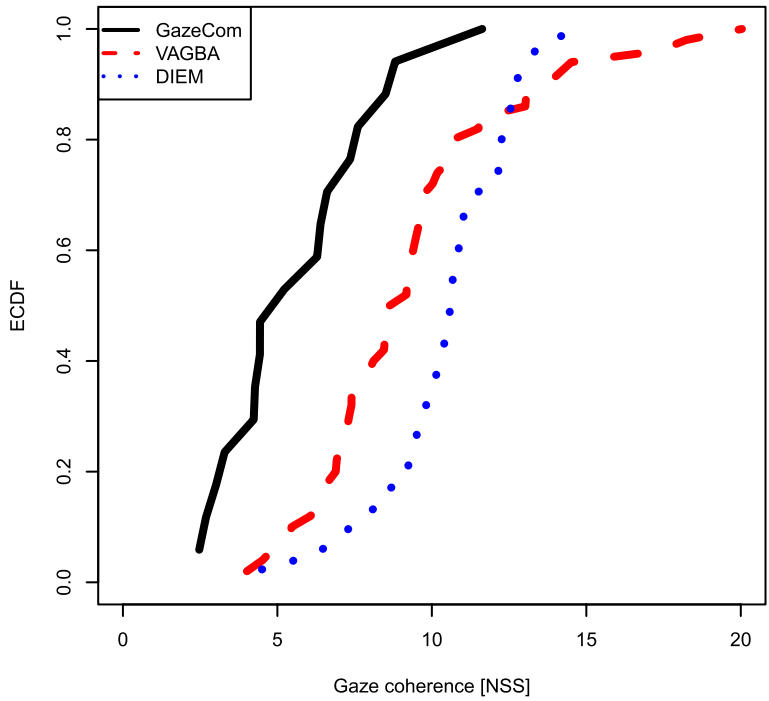

Figure 3. Intersubject gaze variability differs significantly in the three data sets under study (Kolmogorov-Smirnov test, all pairwise comparisons $p < .001$). Variability is lowest (NSS coherence is highest) for the DIEM data set of professionally directed and cut material such as TV shows, movie trailers, and documentaries. The VAGBA and GazeCom data sets were recorded by quasirandomly placing a camera on a tripod in public places such as at crossroads, in parks, etc., and gaze variability is higher. In some VAGBA movies, however, recorded persons interact with the camera/the viewer. To view this figure in colour, please see the online issue of the Journal.

scores, leads to a remarkable result (dashed line in Figure 4). The slopes of the fits for the two nonprofessional data sets are now almost identical (0.00176 GazeCom, 0.00179, VAGBA without outliers; $p < .08$ and $p < .02$, respectively). For DIEM, however, eye movements on movies with greater gaze coherence are less predictable; the negative slope is not statistically significant, though.

DISCUSSION

In this paper, we have investigated some aspects of eye movement predictability and variability on natural video data sets. We analysed three data sets of high-resolution videos that recently were made publicly available and we found large differences in eye movement behaviour between data sets.

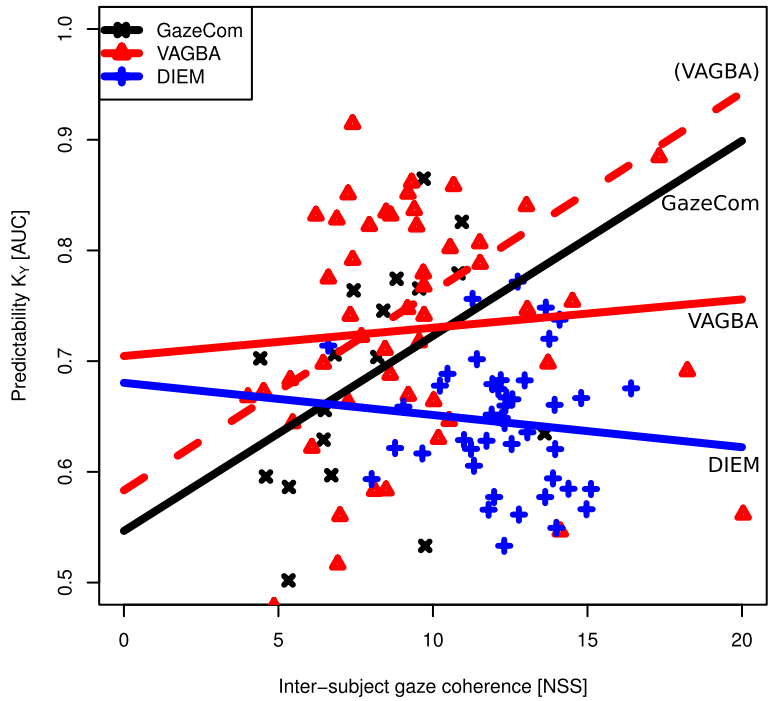

Figure 4. Correlation of eye movement coherence and predictability. In the two naturalistic data sets GazeCom (black crosses) and VAGBA (red triangles), prediction performance is better for movies where eye movements are more similar between observers. Out of VAGBA's 50 movies, six show people directly posing for and interacting with the camera; when these movies, which also have the highest coherence scores, are removed from the analysis, the fits for GazeCom and VAGBA become almost parallel (dashed line). Blue plusses show data for DIEM with professional material; the negative slope is not statistically significant. To view this figure in colour, please see the online issue of the Journal.

This finding replicates previous research that professionally edited material elicits different gaze patterns than more naturalistic stimuli without cuts or deliberate gaze-guidance techniques, such as placing single objects of interest free of distractors at shallow depth of field (Dorr et al., 2010; Mital et al., 2011; Vig et al., 2011).

We then evaluated the predictability of eye movements, using several state-of-the-art models of saliency and two different colour spaces. Doing so on several data sets is especially useful because this approach tests how well a model generalizes to arbitrary input data; repeatedly running and fine-tuning a model on one particular data set might lead to overfitting. Indeed, we found that a conceptually very simple model based on the invariant K of the structure tensor consistently outperformed the more complex Itti and SUNDAy models. Even the invariant H, which very generically describes the

amount of signal change in any dimension, performs better than the Itti saliency model and comparably to the SUNDAy model. However, the absolute differences between models are relatively small, and their performance is highly correlated, i.e., videos that can be predicted well with one saliency model will also be predicted reasonably well with another model. Of particular interest is the almost perfect correlation between SUNDAy and *K* (correlation coefficient .901). This indicates that SUNDAy uses good features to predict eye movements already, but *K* in combination with machine learning achieves superior performance by optimizing parameters such as thresholds to the particular data set. Overall, these results question the assumption that more complex models of low-level saliency can close the gap between the current and perfect prediction performance.

The causal link between image saliency and eye movements that has been put forward previously has recently also come under more fundamental criticism (Tatler et al., 2011). For many practical purposes, however, the distinction between low-level features and high-level, semantic factors might be a moot point. For example, the remarkable ability of human observers to very rapidly and preconsciously saccade towards static images of animals (Drewes, Trommershäuser, & Gegenfurtner, 2011; Kirchner & Thorpe, 2005) indicates that reliable relationships between low-level features and ecologically relevant objects exist. Even more so, under natural, i.e., dynamic conditions, motion is tightly linked to semantically meaningful objects precisely because the ability to move makes predator or prey relevant. Conversely, the absence of low-level features under natural conditions typically also implies the absence of coherent objects. We would therefore hypothesize that the low-level features at a given time determine a set of potential saccade targets from which the actual saccade target is selected based on the history of previous saccades and on simple mechanisms such as inhibition of return and with a task-specific bias (for a population averaging account of saccade selection in visual search, also see Zelinsky, forthcoming).

In a further analysis, we studied the relationship between predictability and variability of eye movements. Previous work by Mital et al. (2011) showed that gaze samples in often-fixated regions (dense gaze clusters) were better predictable. Here, we computed the correlation of predictability and gaze coherence not at the level of individual gaze samples, but for entire video clips. For the two more natural data sets, GazeCom and VAGBA, we also found a positive correlation; in other words, image regions that draw many fixations are more distinct from nonfixated control regions in their low-level features than image regions that are fixated less often. This result would imply that eye movements are at least partially determined by low-level saliency. Remarkably, the slope of a linear fit of predictability versus gaze coherence was almost identical for GazeCom and VAGBA after those movies were removed from VAGBA that are "staged". These movies show

faces directly interacting with the camera, and these very strong high-level cues override any bottom-up saliency (Einhäuser, Spain, & Perona, 2008).

However, contrary to the finding of Mital et al. (2011) at the level of individual gaze samples, we found no positive correlation of predictability and coherence for the DIEM data set at the movie level; we also found no such correlation for a global comparison of all data sets. Surprisingly, the professionally made movies are the least predictable, but also the most coherent. We believe this reduced variability is due to the explicit and implicit gaze-guidance strategies of the movie directors (Dorr et al., 2010; Hasson et al., 2008; Mital et al., 2011). At the same time, this type of movie cannot be predicted well. One possible explanation for this result might be the frequent occurrence of scene cuts. Cuts induce spatiotemporal transients in the saliency map; because they are often ignored by the subjects (Smith & Henderson, 2008), they increase the number of image regions falsely identified as salient. Moreover, we have shown previously that professional videos do not exhibit the near zero average time lag between salient events and eye movements (Vig et al., 2011) and it may well be that the predictions that we make are based on the wrong time lag more often than in other movies.

Overall we have given a comprehensive overview of the state of the art in analysing eye movements made on natural videos and have provided novel results on the generalization performance of saliency models and the impact of different colour representations. Further, we have systematically analysed the predictability and variability of eye movements relative to different saliency measures and have found significant correlations both between eye movements and saliency and between the predictability and the variability of eye movements. Despite the emergence of some global patterns, the various differences we found for different data sets and subtle parameter choices indicate that the large field of research on saliency models should agree on well-defined benchmark data sets that have become standard in the machine learning and computer vision communities. However, the ultimate and objective approach to quantify our understanding of eye movements and saliency will be to evaluate the models in terms of how well one can deliberately change eye movement patterns through low-level feature modifications (Barth, Dorr, Böhme, Gegenfurtner, & Martinetz, 2006).

REFERENCES

Barth, E., Caelli, T., & 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., & Martinetz, T. (2006). Guiding the mind's eye: Improving communication and vision by external control of the scanpath. In B. E. Rogowitz, T. N. Pappas, & S. J. Daly (Eds.), *Human vision and electronic imaging*

XI (Vol. 6057, pp. 116–123). Bellingham, WA: Society of Photo-Optical Instrumentation Engineers.

Barth, E., Dorr, M., Vig, E., Pomarjanschi, L., & Mota, C. (2010). Efficient coding and multiple motions. *Vision Research*, *50*(22), 2190–2199.

Bruce, N., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 1–24.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines [Software]. Retrieved from http://www.csie.ntu.edu.tw/cjlin/libsvm

Dorr, M., Martinetz, T., Gegenfurtner, K., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), 1–17.

Drewes, J., Trommershäuser, J., & Gegenfurtner, K. R. (2011). Parallel visual search and rapid animal detection in natural scenes. *Journal of Vision*, *11*(2), 1–21.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 1–19.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 11–26.

Everingham, M., van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Field, D. J. (1987). Relations between the statistics of natural images and the response profiles of cortical cells. *Journal of the Optical Society of America*, *4A*, 2379–2394.

Foulsham, T., & Kingstone, A. (forthcoming). Modeling the influence of central and peripheral information on saccade biases in gaze-contingent scene viewing. *Visual Cognition*.

Gao, D., & Vasconcelos, N. (2009). Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, *21*(1), 239–271.

Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, *19*(1), 185–198.

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience of film. *Projections*, *2*(1), 1–26.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*(10), 1304–1318.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*(6), 1093–1123.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, Vol. 19 *(NIPS 2005)* (pp. 547–554). Cambridge, MA: MIT Press.

Itti, L., Koch, C., & 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. (1999). Local structure. In B. Jähne & H. Haußecker (Eds.), *Handbook of computer vision and applications* (Vol. 2, pp. 209–238). San Diego, CA: Academic Press.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *Proceedings of IEEE international conference on Computer Vision (ICCV)* (pp. 2106–2113). Los Alamitos, CA: IEEE Computer Society.

Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*(5), 1–15.

Kienzle, W., Schölkopf, B., Wichmann, F. A., & 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). Berlin, Germany: Springer Verlag.

Kirchner, H., & Thorpe, S. (2005). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*, 1762–1776.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–227.

Land, M., & McLeod, P. (2000). From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience*, *3*(12), 1340–1345.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559–3565.

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*(11), 1311–1328.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *IEEE conference on Computer Vision and Pattern Recognition* (pp. 1–8). Los Alamitos, CA: IEEE Computer Society.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817.

Li, Z., Qin, S., & Itti, L. (2011). Visual attention guided bit allocation in video compression. *Image and Vision Computing*, *29*, 1–14.

Mahadevan, V., & Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 171–177.

Mital, P. K., Smith, T. J., Hill, R., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, *3*(1), 5–24.

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

Nuthmann, A., & Henderson, J. (forthcoming). Using CRISP to model global characteristics of fixation durations in scene viewing and reading with a common mechanism. *Visual Cognition*.

Nyström, M., & Holmqvist, K. (2010). Effect of compressed off-line foveated video on viewing behavior and subjective quality. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *6*(1), 1–14.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125–154.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(8), 2397–2416.

Poynton, C. (2003). *Digital video and HDTV*. San Francisco, CA: Morgan Kaufmann Publishers.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*, 341–350.

Smith, T. J., & Henderson, J. M. (2008). Edit blindness: The relationship between attention and global change blindness in dynamic scenes. *Journal of Eye Movement Research*, *2*(2), 1–17.

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

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*, 1857–1862.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 1–23.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6–7), 1029–1054.

Vig, E., Dorr, M., & Barth, E. (2009). Efficient visual coding and the predictability of eye movements on natural movies. *Spatial Vision*, *22*(5), 397–408.

Vig, E., Dorr, M., Martinetz, T., & Barth, E. (2011). Eye movements show optimal average anticipation with natural dynamic scenes. *Cognitive Computation*, *3*(1), 79–88.

Vig, E., Dorr, M., Martinetz, T., & Barth, E. (in press). Intrinsic dimensionality predicts the saliency of natural dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wandell, B. A. (1995). *Foundations of vision*. Sunderland, MA: Sinauer Associates.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–73). Hove, UK: Psychology Press.

Zelinsky, G. (forthcoming). TAM: Explaining off-object fixations and central fixation tendencies as effects of population averaging during search. *Visual Cognition*.

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

Zhang, L., Tong, M. H., & Cottrell, G. W. (2009). SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of the 31st annual Cognitive Science conference, Amsterdam, The Netherlands* (pp. 2944–2949). Mahwah, NJ: Lawrence Erlbaum.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 1–20.