Vision Research 49 (2009) 2918-2926

The contribution of low-level features at the centre of gaze to saccade target selection

Michael Dorr*[,a], Karl R. Gegenfurtner[b], Erhardt Barth[a]

[a]*Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany*
[b]*Abteilung Allgemeine Psychologie, Justus-Liebig-Universität, Otto-Behaghel-Str. 10F, 35394 Gießen, Germany*

Abstract

Does it matter what observers are looking at right now to determine where they will look next? We recorded eye movements and computed colour, local orientation, motion, and geometrical invariants on dynamic natural scenes. The distributions of differences between features at successive fixations were compared with those from random scanpaths of varying similarity to natural scanpaths. Although distributions show significant differences, these feature correlations are mainly due to spatio-temporal correlations in natural scenes and a target selection bias, e.g. towards moving objects. Our results indicate that low-level features at fixation contribute little to the choice of the next saccade target.

Key words: eye movements, natural scene statistics

1. Introduction

After several decades of intensive research, it still remains an open question what the exact factors are in driving eye movements on natural scenes. There is a general consensus that eye movements are guided by both bottom-up, image-driven properties as well as top-down, cognitive processes, but the

*Corresponding author
Email addresses: dorr@inb.uni-luebeck.de (Michael Dorr), Karl.R.Gegenfurtner@psychol.uni-giessen.de (Karl R. Gegenfurtner), barth@inb.uni-luebeck.de (Erhardt Barth)

relative importance of these two mechanisms is still under debate. Recently, Dragoi and Sur (2006) introduced a further mechanism that does not fall neatly in either category and rests on the relationship of low-level features at the current centre of gaze and low-level features at potential saccade targets. Because information about the observer, the current gaze position, needs to be taken into account, a pure bottom-up model does not suffice to describe this mechanism; on the other hand, the mechanism seems to work at a pre-attentive stage, so a description as top-down would also be inadequate. Dragoi et al. based their work on measurements of rhesus monkeys watching still images. In this paper, we will systematically investigate whether the proposed mechanism also can be found, for a variety of low-level features, in human observers watching videos, i.e. whether low-level features at the current centre of gaze contribute to saccade target selection under natural viewing conditions.

1.1. Bottom-up and top-down mechanisms

An influence of the task at hand on gaze behaviour was already found by Yarbus (1967), a finding that was corroborated also for real-life activities (Land and Hayhoe, 2001; Ballard and Hayhoe, 2009). Because of the complexity of modelling cognitive factors, however, much research has focused on bottom-up, low-level factors that can be computed from the stimuli alone. This was further facilitated by the finding that the distribution of image features at the centre of fixation differs significantly from that at random control locations (Mannan et al., 1997; Reinagel and Zador, 1999; Parkhurst et al., 2002; Tatler et al., 2005; Baddeley and Tatler, 2006; Tatler et al., 2006), which can be interpreted as a preference of the human visual system for highly structured image regions (but see below). A common approach to model such low-level factors is that of a *saliency map* (e.g. Privitera and Stark (2000); Itti and Koch (2001); Itti (2005) on static images; Carmi and Itti (2006); Böhme et al. (2006) on videos). Every feature under consideration, such as contrast, motion, or edge density, is stored in a feature map that assigns a certain value to every location in the image. These feature maps are combined by a weighting scheme to obtain relative saliency values for image locations. Simple models might always pick the image location with the maximum saliency value as the next saccade target; see Tatler et al. (2005) for a discussion of how top-down strategies could select targets from a set of candidate points that were determined on the saliency map.

Despite some success in predicting eye movements based on low-level features alone during free viewing (Meur et al., 2007; Vig et al., 2009), it has also been shown that task demands can overrule image-based saliency (Henderson et al., 2007; Einhäuser et al., 2008a). It is argued that it is not low-level features per se, but semantically meaningful objects (the presence of which is correlated with image structure) that drive attention (Foulsham and Underwood, 2008; Einhäuser et al., 2008b); however, it is also still under debate whether low-level features are merely correlated with objects or give rise to their perception (Elazary and Itti, 2008).

1.2. Correlations of low-level features at successive fixations

Based on psychophysical and electrophysiological evidence, a further mechanism for the selection of saccade targets was put forward by Dragoi and Sur (2006). They showed that when V1 neurons were adapted to gratings of a certain orientation for 400 ms, subsequent discrimination performance improved for both iso-orientation and orthogonal gratings; discrimination of gratings with an intermediate orientation difference, on the other hand, did not change significantly. Dragoi and Sur (2006) related these findings to eye movement recordings from rhesus monkeys viewing still images that showed that fixations of an image patch were likely to be followed by either a small saccade to a patch with similar orientation or by a large saccade to a patch with largely dissimilar orientation. The proposed explanation was that eye movements exploit the improved discrimination performance and steer gaze towards either iso-oriented or orthogonally oriented image patches. A schematic illustration of this analysis can be found in Fig. 1, which depicts a putative scanpath on a synthetic scene: from each fixation patch, dominant local orientation ϕ is extracted (e.g. $\phi_1 = 90°$, $\phi_2 = 135°$, etc.). The differences of orientation at successive fixations then can be computed (e.g. $\Delta\phi_1 = |\phi_2 - \phi_1| = 45°$) and their distribution compared with a distribution of differences obtained on randomly generated control scanpaths. In the case of Dragoi et al., the distribution of differences in orientation was more U-shaped for measured than for random baseline scanpaths because both very small and very large differences occured more often.

Looking at these differences of low-level features can also be interpreted as evaluating the correlation of such features introduced by the visual system's target selection process. At this point, however, it is important to note that natural scenes are highly correlated both in space and time (Zetzsche et al.,

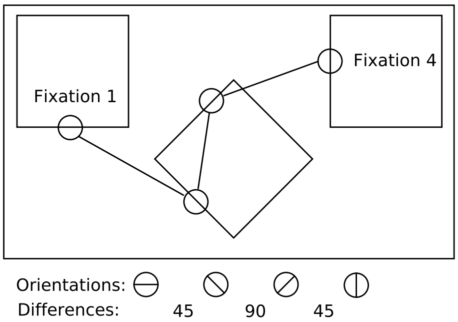


Figure 1: Schematic illustration of the analysis for a synthetic scene; real data was measured on natural videos. Low-level features (here: orientation) are extracted from each fixated image patch and their differences along the scanpath are computed.

1993; Simoncelli, 1997); it is therefore crucial to carefully discriminate these image-inherent correlations from those that are due to eye movements.

If we found such eye movement–induced correlations indeed, we could also understand them as a contribution of low-level features at the current centre of fixation to the selection of the next saccade target. This is of particular interest to the prediction of eye movements: here, it were not sufficient anymore to look at a saliency map that is independent of current eye position. On the contrary, information from the current eye position would be required to determine where the eye will look next. Similar analyses of oculomotor tendencies such as saccadic amplitude and direction, fixation duration, and the bias towards the centre of the stimulus have shown that such factors can significantly improve feature-based models of eye guidance (Tatler and Vincent, 2009b,a).

In the remainder of this paper, we will apply the technique of looking at feature differences at successive fixations to a large set of eye movement data from human subjects watching high-resolution video clips. To use video clips instead of still images has the advantage that viewing conditions are more natural; on still images, a few fixations might suffice to capture all relevant scene information, after which image sampling might become idiosyncratic. For further work on eye movements on dynamic natural scenes, we refer to e.g. Tseng et al. (2009); Carmi and Itti (2006); Stelmach and Tam (1994).

To extend the analysis, we did not only look at local orientation, but systematically investigated other low-level features as well. In particular, these were brightness, colour, and motion. Even though the choice of these features might be arbitrary to a certain extent, there seems to be a general consensus that these features are extracted at an early stage in visual information processing (Adelson and Bergen, 1991). Furthermore, we analysed the correlation of geometrical invariants, which are basic dynamic features from a computational perspective and have been shown to be useful in understanding various phenomena in biological vision (Zetzsche and Barth, 1990; Zetzsche et al., 1993; Barth and Watson, 2000). The invariant H (see below) can also be interpreted as spatio-temporal contrast.

Finally, our analysis was performed on a spatio-temporal multiresolution pyramid (Burt and Adelson, 1983) in order to capture any effect that might be limited to a certain spatio-temporal scale.

2. Methods

2.1. Experimental setup

18 colour video clips of 20 s duration each were recorded using a JVC JY-HD10 HDTV video camera. Clips showed outdoor real-world scenes in and around Lübeck; they had a spatial resolution of 1280 by 720 pixels and a frame rate of 29.97 frames per second (progressive scan). These image sequences were displayed at 90 Hz refresh rate on an Iiyama MA203DT 22" screen covering an area of 40 by 22.5 cm. At a viewing distance of 45 cm, the videos thus spanned a horizontal field of view of about 48 degrees and had an angular resolution of 13.4 cycles/degree. 54 subjects took part in the experiment. They were instructed to watch the movies attentively; no other specific task was given. Their eye movements were recorded using an SR Research EyeLink II tracker at 250 Hz. An initial calibration was performed prior to the experiment; after each movie presentation, an additional drift correction was performed.

2.2. Eye movement analysis

The eye tracker flags invalid samples (for example, during blinks); in a first analysis step, trials where more than 5% of samples were invalid were discarded, leaving between 37 and 52 recordings per video sequence and 844 recordings (with 42331 fixations) overall. The extraction of fixations made on dynamic scenes from such raw data is not trivial due to the occurrence of smooth pursuit eye movements (Munn et al., 2008), and our investigation of successive fixations obviously hinges crucially on a faithful detection of fixations. Therefore, we chose to implement two different fixation identification algorithms, namely the GUIDe algorithm developed by Kumar (2007) and our own algorithm that combines velocity- and dispersion-based approaches: first, saccades were extracted in a two-fold procedure. To initialize the search for a saccade onset, gaze velocity had to exceed a high threshold $\theta_1 = 138°/\mathrm{s}$ first; then, saccade onset and offset were defined as the points in time where gaze velocity passed a lower threshold $\theta_0 = 17°/\mathrm{s}$ that is biologically more plausible but less robust to noise. From intersaccadic samples, a fixation was detected if gaze remained stationary (within 0.35°) for at least 100 ms; the (x, y) coordinates of the fixation were then computed as the mean of the stationary samples. Performance of both algorithms was validated against a randomly sampled set of 550 hand-labelled fixations; the GUIDe algorithm yielded a slightly better agreement and was therefore used for all results presented in this paper. Nevertheless, we ran the same analyses using the second algorithm and obtained qualitatively similar results. For the GUIDe algorithm, we also computed the extent to which gaze samples remained unlabelled as either fixation or saccade, which might indicate a smooth pursuit movement. About 9% of gaze samples could not be labelled reliably; however, average duration of such unlabelled episodes was 37 ms, which would be fairly short for phases of smooth pursuit, so that it was possibly often rather the transitions between (high-velocity) saccades and (low-velocity) fixations that caused problems for the algorithm. Manual inspection further revealed that some clear episodes of smooth pursuit, e.g. when a flock of birds flies by in one of the videos, were broken into a series of fixations and 'undefined' samples. However, the depicted objects are not translated rigidly, change course, etc., so that even a manual labelling would be difficult. In the context of the present study, it is not clear at any rate how smooth pursuit should be treated, since e.g. catch-up saccades would keep fixation on the same object.

2.3. Low-level features

All low-level features were computed on a multiresolution pyramid constructed from the image sequence by successive blurring and sub-sampling in both the spatial and the temporal domain. In our implementation, we created 5 spatial (13.4, 6.7, 3.3, 1.7, and 0.8 cycles/degree) and 3 temporal (30, 15, 7.5 fps) scales. Except for colour, all features were determined on the luma channel (see Methods below) of the video.

Timing of feature extraction with regard to fixation onset. For each fixation, we extracted features from that video frame that was shown on the screen at the onset of fixation. The human visual system, however, has to base its decision where to move the eyes next on information that was available earlier already because of its sensory-motor latency. Therefore, we additionally ran all analyses again with features that were extracted at up to 200 ms (in steps of 25 ms) before fixation onset, respectively; due to the temporal correlations in the videos, results were qualitatively similar (data not shown).

Orientation. Local orientation was extracted using a standard technique based on the eigenvalues of the two-dimensional structure tensor J_2 (for a textbook coverage, see e.g. Jähne, 1999).

$$J_2 = \omega * \begin{pmatrix} f_x f_x & f_x f_y \\ f_x f_y & f_y f_y \end{pmatrix},$$

where $f(x, y)$ is the image-intensity function, subscripts indicate partial derivatives, and ω is a spatial smoothing kernel applied to their products (here, an 11-tap binomial kernel was used). If the rank of J_2 is zero (both eigenvalues $\lambda_1, \lambda_2 = 0$), the image patch is homogeneous. A rank of two ($\lambda_1 > 0, \lambda_2 > 0$) indicates a 2D feature, e.g. a corner. An ideal orientation corresponds to a rank of one ($\lambda_1 > 0, \lambda_2 = 0$), with a direction given by the eigenvector corresponding to the zero eigenvalue. To increase robustness in the presence of noise, however, eigenvalues were not checked against zero, but against a threshold defined by the maximum eigenvalue over all image patches extracted from a video: $\lambda_1 + \lambda_2 < \theta_1 \cdot \lambda_{max}$; relative size of the eigenvalues was controlled by a second threshold $\theta_2 < \frac{\lambda_2 - \lambda_1}{\lambda_1 + \lambda_2}$. We varied θ_1, θ_2 systematically in the range 0.01–0.1 and 0.1–0.9, respectively.

Colour. MPEG-2 video as recorded by our camera stores colour in the $Y'C_rC_b$ format with one channel corresponding to brightness and two to colour-opponency information (Poynton, 2003). We directly used the intensity values from all channels.

Velocity. Motion estimation followed the algorithm presented in Barth (2000). It is based on the three-dimensional structure tensor J defined as

$$J = \omega * \begin{pmatrix} f_x f_x & f_x f_y & f_x f_t \\ f_x f_y & f_y f_y & f_y f_t \\ f_x f_t & f_y f_t & f_t f_t \end{pmatrix},$$

where ω is a spatio-temporal smoothing kernel (here 5-tap binomials in both space and time). Analogously to two-dimensional orientation, motion can now be computed from the eigenvalues of J (Haußecker and Spies, 1999), but a more robust method is to use the minors M_{ij} of J, obtained by eliminating row $4 - i$ and column $4 - j$ of J. Four different expressions of the form $\vec{v}_1 = (M_{31}, -M_{21})/M11$ can be derived and compared against each other for improved noise resilience. Velocity was computed as $v = \sqrt{v_x^2 + v_y^2}$ and locations where v was less than 1% of the maximum velocity in that video frame were discarded. Finally, results were smoothed with a Gaussian kernel with length 15, $\sigma = 3$ pixels.

Geometrical invariants. We also computed geometrical invariants that have been shown to be useful in understanding biological vision (Barth and Watson, 2000); they have also been used to predict eye movements before (Zetzsche et al., 1998; Böhme et al., 2006; Vig et al., 2009):

$$\begin{aligned} H &= 1/3 \text{ trace}(J) && = \lambda_1 + \lambda_2 + \lambda_3 \\ S &= |M_{11}| + |M_{22}| + |M_{33}| && = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 \\ K &= |J| && = \lambda_1\lambda_2\lambda_3. \end{aligned}$$

From these invariants, the *intrinsic dimensionality* of the image sequence can be inferred (Zetzsche and Barth, 1990). The intrinsic dimension of a signal at a particular location is the number of directions in which the signal is locally non-constant. $H > 0, S = K = 0$ indicates an $i1D$ feature such as an edge, $S > 0, K = 0$ indicates an $i2D$ feature such as a corner or a transient edge, and transient corners are $i3D$ and have $K > 0$. An example image for invariant S is shown in Fig. 2.

2.4. Artificial scanpaths as baseline measure

To be able to compare our results against a baseline measure, we created random sequences of fixations, or scanpaths. However, real scanpaths have certain characteristics that need to be taken into account. For example, the distribution of saccadic amplitudes that subjects made on our stimuli (see Fig. 3(a)) is heavily skewed (mean amplitude is 7.4°, median is 5.6°). Because natural scenes show spatio-temporal correlations that vary with distance (Zetzsche et al., 1993; Simoncelli, 1997), see also Fig. 3, any correlations found along the scanpath might be due to these image-inherent correlations alone. Furthermore, it is a well-known fact that human gaze prefers image patches with high local structure, such as edges, corners, or motion. This

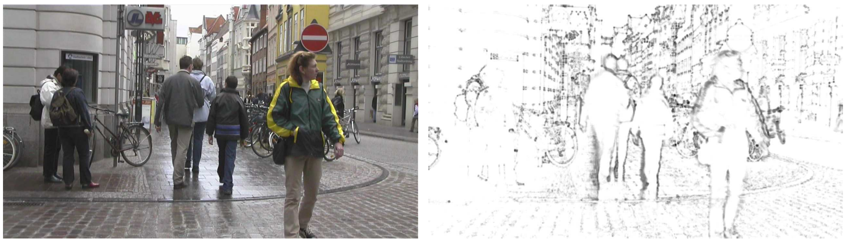


Figure 2: Left: Still shot from one of the movies used in our experiment. Right: Corresponding image of geometrical invariant S. Non-white locations change in at least two spatio-temporal directions (brightness thresholded and inverted for better legibility).

repulsion from homogeneous areas is of particular importance in the context of the orientation feature since orientation cannot be reasonably extracted from such areas.

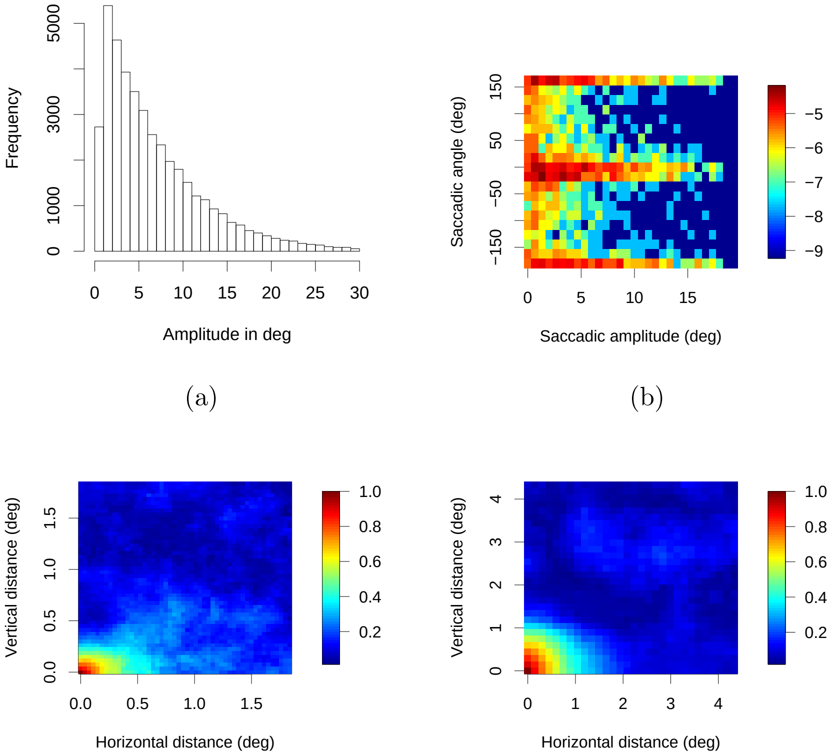
In order to disambiguate these effects, we created four different sets of baseline scanpaths with a different similarity to the recorded scanpaths. A graphical illustration of these control conditions is given in Fig. 4.

"Random". Fixation durations were copied from real scanpaths, but image coordinates of fixations were uniformly sampled across the whole scene, resulting in a mean saccadic amplitude of 19°. Thus, in this condition neither saccadic amplitude nor the set of fixated patches remained the same as in the real scanpaths.

"Same lengths". Saccade lengths were copied from real scanpaths, but direction was randomized; most correlations inherent in natural scenes were therefore conserved, but the image patches from which features were extracted were random.

"Scrambled". In this condition, the order in which a subject fixated a series of image patches was shuffled. This yielded a different distribution of saccadic amplitudes (mean 13°, almost twice as large as that of the original distribution), but the set of fixation coordinates (x, y) remained constant. Note that this does not imply that fixated image patches were exactly the same; because of moving objects and illumination changes over time, features at (x, y, t_1) and (x, y, t_2) might differ for $t_1 \neq t_2$.

"Synthetic". All the above conditions are based on data from a single trial (combination of one subject and one movie) per output scanpath. Using



(a)

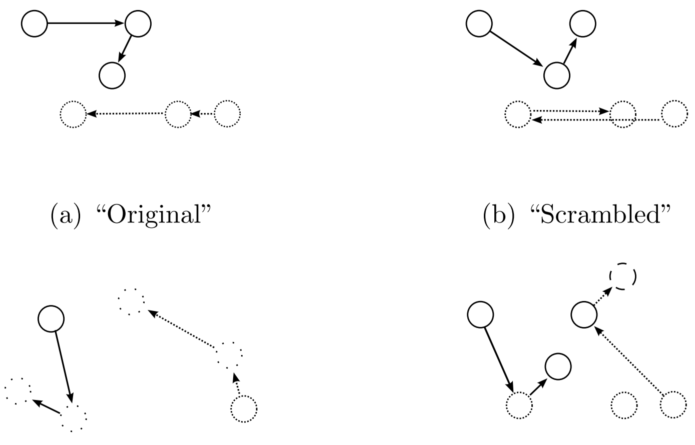
(b)

(c)

(d)

Figure 3: a) Distribution of saccadic amplitudes on our set of videos that were displayed at a size of 48 by 27 degrees visual angle. About 1% of saccades had an amplitude of more than 30 degrees (not shown here). b) Log-plot of joint distribution of saccadic amplitudes and angles. There is a strong bias towards horizontal and vertical saccades. c) Image-based correlation of local orientations on the highest spatial scale (13.4 cycles/deg). The bottom-left corner corresponds to the correlation of a pixel with itself, which is 1.0 by definition. At longer distances (above 0.5 to 1 degree), correlations drop to chance level; notable is the anisotropy that correlations decay more slowly along the horizontal axis. d) Image-based correlation of local orientations for a middle spatial scale (3.3 cycles/deg). Again, correlations are anisotropic.

data only from a single subject, it is impossible to change the scanpath (i.e. generate an artifical scanpath) while keeping constant both the set of fixated patches and the spatio-temporal distances between these fixations. However, by mixing scanpaths made by different observers on the same video,


(a) “Original” (b) “Scrambled”

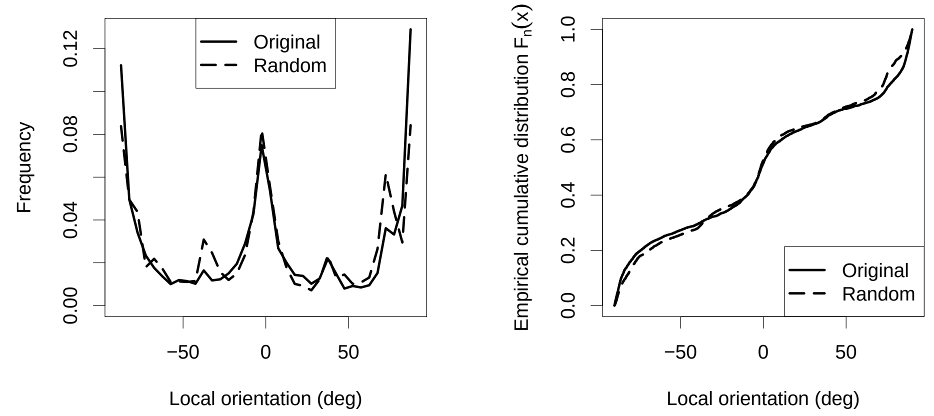
(c) “Same lengths” (d) “Synthetic”

Figure 4: Illustration of the control conditions (“random” not shown). a) Measured scanpaths from two subjects (solid line / dashed line). b) “Scrambled”: fixations are the same, but their order is randomized. c) “Same lengths”: the fixation locations (except for the start position) are random, but the connecting saccades have the same amplitudes as in the “original” condition. d) “Synthetic”: using scanpaths from several subjects, both the set of fixated locations and the joint distribution of saccadic angles and amplitudes are approximated (in this small sketch, only amplitudes are similar); no saccadic segment occurs in the “original” scanpaths. Note the fixation from a putative third subject in the top right corner.

both these characteristics can be approximated simultaneously. Consider a sequence of two fixations made by subject A: $f_A(n) = (x_A(n), y_A(n)), f_A(n+1) = (x_A(n+1), y_A(n+1))$ with a distance $\Delta_A(n) = (x_A(n+1)-x_A(n), y_A(n+1) - y_A(n))$ (for simplicity, we ignore time in this example). In an artificial scanpath S, we would then want to model a pair of fixations with the same distance (since Δ is a vector-valued function, this also includes the angle between the two fixations), $f_S(n) = (x_S(n), y_S(n)), f_S(n+1) = f_S(n) + \Delta_A(n)$. Furthermore, $f_S(n)$ and $f_S(n+1)$ should not be random points, but real fixation points. Given a sufficient number of scanpaths from other subjects, it is not unlikely to find (at least approximately) such a pair of fixations, e.g. from subjects B and C: f_B, $f_C = f_B + \Delta_A(n) + \epsilon$, that we can use for our “synthetic” scanpath: $f_S(n) := f_B$, $f_S(n+1) := f_C$. Care has to be taken, however, that the artificial scanpath does not coincidentally become a mere copy of original scanpath segments, i.e. that there is no subject X with fixations $f_X(n) = f_B$, $f_X(n+1) = f_C$.

In practice, "synthetic" scanpaths were created as follows. An output scanpath was initialized with the first fixation of an original scanpath. Then, the same number of fixations as in the input scanpath was generated by sampling pairs of angles and amplitudes from the joint distribution over the original scanpaths (see Fig. 3(b)); for each sample, we searched among all observers' fixations for one with a similar distance at a similar angle to the current fixation (tolerances were 0.2 degrees of amplitude and 10 angular degrees). Because of moving objects, the image patch around one fixation point might look different over time, and therefore we initially searched only among those fixations that had been made at a similar point in time (tolerance 0.5 s). As mentioned above, theoretically it would be possible to end up with an exact copy of the input scanpath, since that copy trivially mimicks both saccadic amplitudes and angles and the set of fixation points. Therefore, a further constraint was that no sampled pair of saccade onset and offset was also part of any of all subjects' original scanpaths (again with a tolerance of 0.2 degrees). Obviously, these conditions could not always be fulfilled: even a large data set of fixation points is relatively sparse on the screen (the screen measures about 1300 deg^2; at a spatial tolerance of 0.2 deg, a single fixation point covers only 0.01% of this area), and certain combinations of angles and amplitudes might take a scanpath outside the borders of the video, which is clearly nonsensical. In these cases, sampling from the joint distribution was repeated up to 10 times and the tolerance for "similar" time points was gradually relaxed until a matching fixation patch could be found. The C++ source code for this algorithm is available upon request.

To assess how closely the original distribution of saccade length and direction was approximated, we computed the Kullback-Leibler divergence between the original distribution and those generated by the baseline conditions. For a reference point, we also computed the KLD of one half of the original data set to the other half. Results were 1.19, 0.4, 0.08, 0.07, and 0.05, respectively (for "random", "scrambled", "lengths", "synthetic", and "original"). These results show that the "synthetic" scanpaths are only an approximation to the original scanpaths, but model the saccade characteristics of original scanpaths more closely than those in the "lengths" condition, even though they consist only of real fixation points (in the "lengths" condition, fixation points are random).



(a) (b)

Figure 5: Example of probability and empirical cumulative distribution functions (ECDF); here, the distribution of orientations is plotted. $F_n(x)$ denotes the proportion of samples having a value of less than or equal to x, e.g. 50% of samples have an orientation between -90 and 0 degrees. Peaks in the probability distribution (left panel) correspond to a steep slope in the ECDF (right), e.g. at -90, 0, +90 degrees.

In summary, by introducing the concept of synthetic scanpaths, we can avoid the shortcomings of random and scrambled scanpaths and, in addition, match the natural distribution of saccade length and direction.

3. Results

To see whether the features along scanpaths made by human observers are correlated beyond the level that is to be expected from image-inherent spatio-temporal correlations alone, we have to compare the distributions of feature differences along the "original" scanpaths with distributions based on the control scanpaths. Because of random fluctuations, finding subtle differences in raw distributions is quite hard; we therefore look at the empirical cumulative distribution functions $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$, where $I_{X_i \leq x} = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases}$, which integrate over difference magnitude.

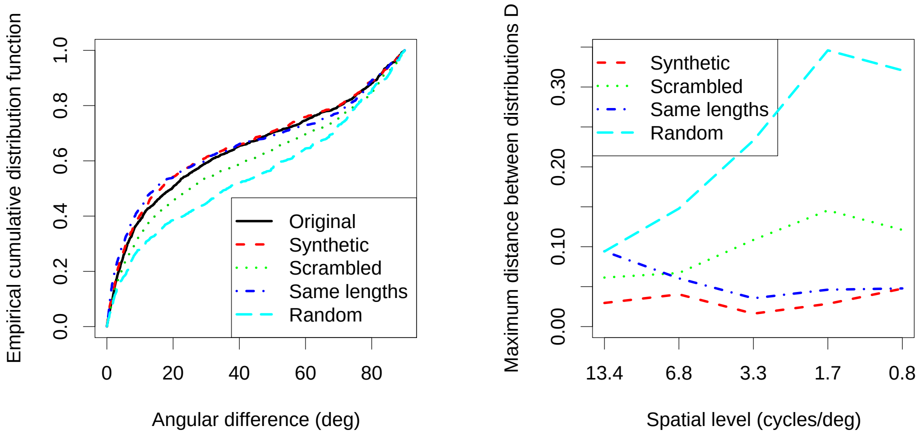
As an example, consider the two distributions of orientation values in Fig. 5. The solid line depicts the distribution of orientations at human fixation points and the dashed line those at random control points; the dominance of the horizontal ($\phi = 0°$) and vertical ($\phi = \pm 90°$) axes is a well-known property of natural scenes and can therefore be found both in human and random data. The ECDF (shown in the right panel) at x tells us what proportion of samples have a value of less than or equal to x, e.g. about 50% of samples have an orientation between -90 and 0 degrees. Peaks in the probability distribution (left panel) correspond to a steep slope in the ECDF (e.g. for the cardinal axes); low $p(x)$ values correspond to plateaus (e.g. for oblique orientations). Based on the ECDF, the Kolmogorov-Smirnov test statistic $D_{ij} = \sup_x |F_i(x) - F_j(x)|$ denotes the maximum distance of two cumulative distributions on the y-axis. In our example in Fig. 5, this maximum distance is 6.3%: around 82% of samples in the "original" distribution have an orientation of less than $x = 80$ degrees, but the dashed "random" curve has reached more than 88% at this point already.

Depending on the number of samples in the distributions, every such distance D_{ij} is then assigned a probability p to test for statistical significance. Since the Kolmogorov-Smirnov test is valid only for continuous distributions, but the low-level features colour and invariants are represented by discrete values, we performed a 1000-fold bootstrap test and report 95% confidence interval values.

We should take statistical tests with a grain of salt, though. Overall, we have almost 500 conditions ($5 \cdot 3$ spatio-temporal levels, 8 different features with varying parameters, 4 types of control scanpaths). Even at a significance level of $p = 0.01$, this implies that we have to expect around 5 conditions with presumably significant results, even if there was no underlying effect. Therefore, we carefully have to look out for systematic effects, i.e. those that are robust against scale or parameter changes. Also, because of the high number of samples, even miniscule effects can show up as highly significant.

In the following, we will present and discuss some representative findings. We will start out with orientation and colour because here the analysis is straightforward; results for motion and the geometrical invariants need more consideration.

The evaluation of local orientation poses the problem that the threshold θ_1, θ_2, which separate oriented from homogeneous patches, have to be defined. We systematically varied these parameters and found, not surprisingly, that for low orientation specificity ($\theta_1 = 0.02, \theta_2 < 0.4$), random noise dominates the measurements and the control conditions cannot be distinguished from



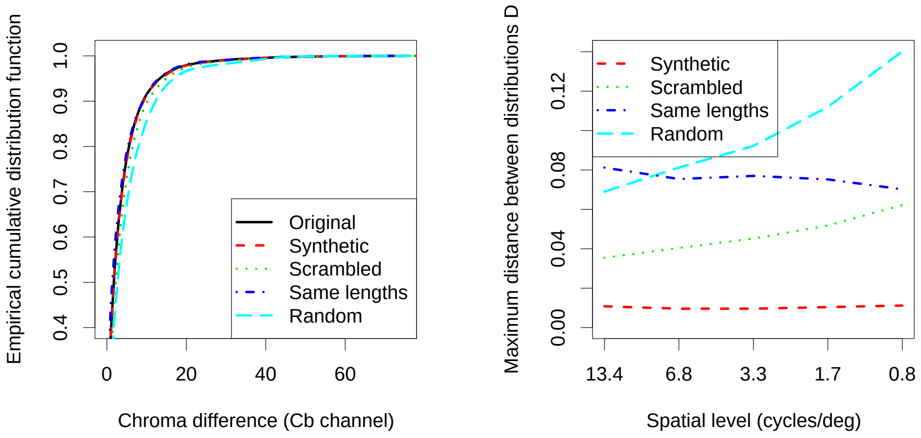
(a)

(b)

Figure 6: Results for local orientation. a) ECDF for differences of orientations on the second spatial, first temporal level. The "random" and "scrambled" conditions strongly differ from the original data in their saccadic amplitudes and therefore also differ in their orientation differences. The "same lengths" condition is closer, but still different at around significance level ($D = 6.0\%$, $p < 0.017$); "synthetic" scanpaths show no such difference ($D = 4.0\%$, $p > 0.18$). b) Maximum distance between original data distribution and control conditions for different spatial scales (first temporal scale; results are similar for other temporal scales). The "synthetic" condition is always closest; "random" is particularly different on the lower-frequency scales.

the "original" condition. At e.g. $\theta_1 = 0.05, \theta_2 = 0.8$, however, reliability of orientation estimation is high; at only about 12% of image patches can orientation be extracted then (nevertheless, the following also holds true for moderate parameter variation). Because human fixations are drawn to structured image regions, the number of strongly oriented patches decreases slightly for the "same lengths" and the "random" conditions (to about 9%).

In Fig. 6(a), the distributions of orientation differences along the scanpaths are plotted for one exemplary spatio-temporal scale. Clearly, the "scrambled" and the "random" conditions are very different from the original data. In these conditions, the saccadic amplitudes changed drastically and hence, also the distance-determined correlations of the image patches changed. The "same lengths" condition mimicks the original data more closely, but is still different almost at significance level ($D = 6.0\%, p < 0.017$); however, only when the image-based correlations are fully modelled in the



(a)

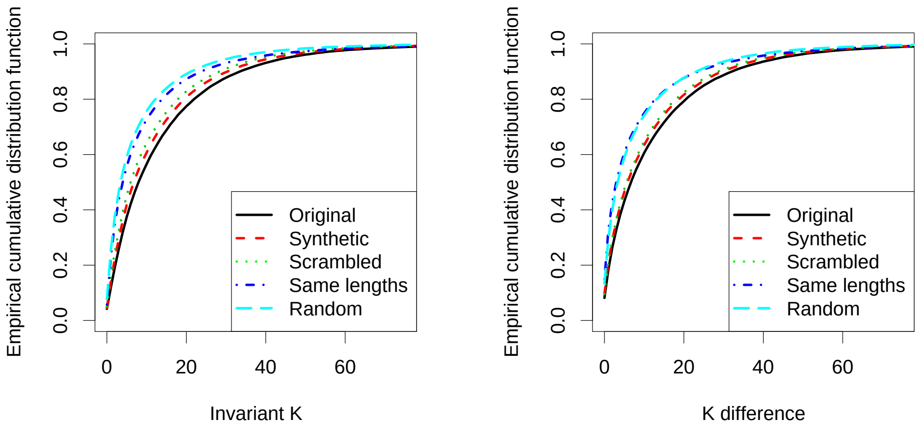
(b)

Figure 7: a) ECDFs of colour differences on the fourth spatial, first temporal level. The "synthetic" condition shows no significant difference to the original data ($D = 0.7\%$, $p > 0.07$). b) Maximum distance between original data distribution and controls for different spatial scales.

"synthetic" condition and even the angular distribution of saccades is taken into account, the difference to the human data vanishes. Compared to the "synthetic" artificial scanpaths, human subjects did not show a preference for certain orientation differences from one fixation to the next ($D = 4.0\%$, $p > 0.18$). The same pattern can be seen in Fig. 6(b), where the test statistic D is plotted for all spatial scales. The "synthetic" condition is always closest to "original", and "random" is particularly bad on the lower spatial scales.

Because Dragoi and Sur (2006) found different effects for saccades of different sizes, we also evaluated subsets of our data based on saccadic amplitude: following Dragoi and Sur, we binned saccades into small ($< 1°$), medium ($1 - 3°$), and large ($> 3°$); since the stimuli in our data set were much larger, we also partitioned the saccades along the median of roughly 6°. No significant differences between "synthetic" and "original" could be found in any of these subsets (data not shown).

Proceeding to the next low-level feature, colour, Fig. 7(a) shows exemplary data for the blue-difference chroma channel C_b on the fourth spatial level, but the following applies also to luma (Y) and red-difference chroma (C_r). Here, all those artifical scanpath models with different sac-



(a) (b)

Figure 8: a) Cumulative distribution of K values at fixated image patches. The "original" condition shows a small bias towards larger K values compared to "synthetic" and a large bias compared to "same lengths". b) ECDF of differences of geometrical invariant K on the third spatial and the first temporal level. There is a statistically significant difference ($D = 2.5\%$, $p < 10^{-5}$) between the "original" and the "synthetic" condition, but this difference can be explained by the difference in the underlying feature distributions (see a)).

cadic amplitudes ("scrambled" and "random") or different fixation locations ("same lengths") lead to very different colour differences along the scanpath ($p < 10^{-5}$ on almost all spatio-temporal levels). Only the "synthetic" condition shows no significant difference to the original data ($D = 0.7\%$, $p > 0.07$); for this condition, no such difference can be found for any spatio-temporal level (see Fig. 7(b)) or colour or brightness channel.

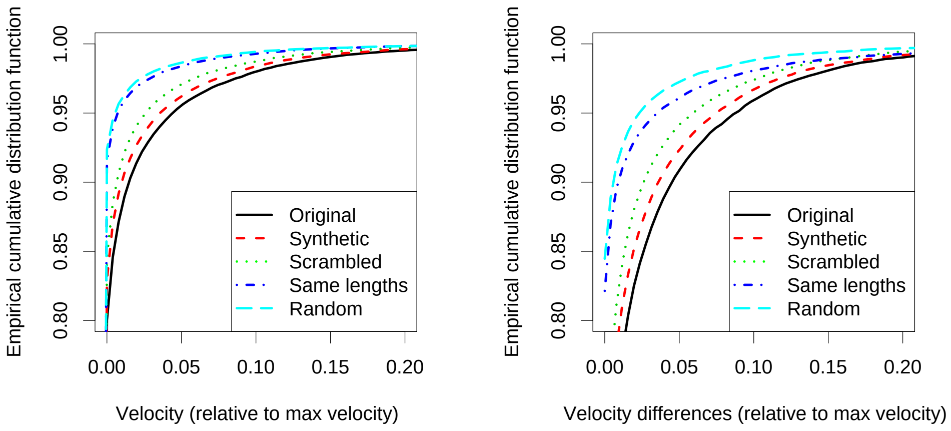
In Fig. 8(a), results are plotted for the geometrical invariant K, which describes the intrinsically three-dimensional video patches such as transient corners. Similar effects could be found on several spatio-temporal levels, and we will here describe one exemplary case (third spatial, first temporal level). Statistically significant differences could not be found for invariants H and S, which correspond to intrinsically one- and two-dimensional features; these features are less sparsely distributed than K and the following discussion therefore does apply only loosely to them.

The black solid curve in Fig. 8(b), which represents the "original" data, saturates later than the other curves; they, in turn, have a steeper slope near $\Delta K = 0$. This means that in the original scanpaths, K values showed larger absolute differences. This effect is particularly strong when comparing the original scanpaths with the conditions “same lengths” and “random”, which are those conditions where image patches were drawn (quasi-)randomly. The difference for the “synthetic” and “scrambled” conditions is less pronounced, but still is statistically significant ($D = 0.9\%, p < 0.017$).

Let us now turn to Fig. 8(b) for an explanation. Shown here are the cumulative distributions of raw K values at fixated image patches. When comparing the “original” condition with “same lengths”, we can see that there is a strong bias towards higher K values—which is in line with the observation that humans prefer to look at highly structured image regions. The image patch selection in the “same lengths” condition, on the other hand, was random and therefore showed no such bias. Although the set of image patches in the “synthetic” condition approximates the measured set of fixated image patches, some spatio-temporal uncertainty is introduced (see section 2.4), so the raw K values for this condition are slightly smaller than for the recorded data ($D = 3.6\%$; although these numbers cannot be compared directly, this is at least in the same order of magnitude as the distance of the distributions of ΔK, $D_{\Delta K} = 2.5\%$).

Thus, we can state that the distribution of K at the centre of gaze is wider for human data than for artifical scanpaths; therefore, the distribution of differences along the scanpath also becomes wider. This bias of the human visual system towards image regions with higher K values, i.e. regions of changes in all spatio-temporal dimensions, can be used to reliably predict eye movements (Vig et al., 2009), regardless of the question whether this observed bias is merely a correlate of other, top-down factors such as a preference for (moving) objects. However, there is no strong evidence for a particular bias in selecting the next saccade target based on the K value at the current centre of fixation.

A similar effect could be found for the motion feature. On almost all spatio-temporal levels, there are significant differences between the original scanpaths and all control conditions. For an example, the distribution of velocity differences on the third spatial and first temporal level is shown in Fig. 9(b). Again, human subjects show a bias towards larger absolute feature differences compared to random processes, but as in the case of K, the underlying distribution is also different. As can be seen in Fig. 9(a), humans tend to fixate moving objects more often (in practice, moving objects are



(a)

(b)

Figure 9: a) Cumulative distribution of velocities. Subjects exhibit a clear bias towards image patches with high velocities. b) ECDF of differences of velocity on the third spatial and the first temporal level. There is a statistically significant ($D = 2.9\%$, $p < 10^{-5}$) difference between the "original" and the "synthetic" condition.

often followed with a smooth pursuit eye movement; see the Methods section for a discussion). Note that the difference between the "original" and the "scrambled" condition is fairly large here even though the spatial locations of the image patches stay the same. Their temporal order changes, and by definition, a moving object will be at a different place at a different time.

Summarizing our results, we can conclude that for orientation, as well as for the other low-level features, there is no significant contribution of the feature at the current centre of gaze to saccade target selection.

4. Discussion

The study presented here was motivated by our research on gaze prediction and gaze guidance. Previously, we had found that low-level features such as the geometrical invariants can be successfully used to predict where observers will direct their gaze in natural movies. In order to potentially improve our prediction algorithm, we investigated the correlation of a variety of low-level features across consecutive fixations. In line with earlier findings by Dragoi and Sur (2006), we found that such correlations are not random and feature differences along the scanpath exhibit systematic characteristics. However, our data does not support Dragoi et al.'s hypothesis that neural adaptation plays a crucial role in forming these characteristics; in other words, that low-level features at the centre of gaze contribute to saccade target selection.

On the contrary, we find that the correlations of features along the scanpath can be explained by two factors. First, natural scenes themselves show strong spatio-temporal correlations, and any distribution of saccadic amplitudes and angles will reproduce these correlations to a varying degree. Second, there exists a general bias in saccade target selection, e.g. the preference of human observers to look at image regions with spatio-temporal structure, which in natural scenes often corresponds to object locations.

For geometrical invariants, which describe the number of spatio-temporal dimensions that change locally, and motion, this preference resulted in a wider distribution of raw feature values at fixated patches; therefore, differences of those features at successive fixations also differed from those in control conditions. For colour and local orientation, we could find an effect only for some of the control scanpath models; when we matched saccade statistics and therefore matched the scene-inherent spatio-temporal correlations, the effect vanished.

Nevertheless, we should stress that our findings do not rule out that low-level features at fixation contribute to saccade target selection at all; it is possible that the human visual system might have learned to make use of a specific distribution of saccadic amplitudes and angles, which induces correlations in the sequence of fixated low-level features that may be beneficial in terms of neural adaptation. However, such a putative mechanism would require no direct knowledge of the relationship of features at fixation and potential saccade targets.

If we had found strong evidence that the visual system does indeed evaluate and compare low-level features at fixation and in the periphery, this would have been a strong argument in the ongoing debate whether top-down or bottom-up factors are more important in the control of eye movements on natural scenes. We here find no indicator that low-level features are explicitly represented and used in oculomotor control. Nonetheless, the opposite conclusion that low-level features are irrelevant is also not supported by our data, since here we investigated exclusively the role of features along the scanpath, not at single fixations.

Finally, we developed and compared several methods to generate artificial scanpaths. The “scrambled” and the “lengths” condition focus on the characteristics of saccade target selection and of oculomotor tendencies, respectively; the “synthetic” condition accurately models both these processes and should thus be preferred, but requires a larger data set to sample from. The highly different results we obtained for these different control conditions emphasize the importance of precisely modelling saccade statistics when comparing human subjects with random processes. In general, this helps to disentangle the properties of the visual input and those of the human visual system.

Acknowledgements

Our research has received funding from the European Commission within the GazeCom project (IST-C-033816) of the FP6, see http://www.gazecom.eu. All views herein are those of the authors alone; the European Commission is not liable for any use made of the information.

References

Adelson, E. H., Bergen, J. R., 1991. The plenoptic function and the elements of early vision. In: Landy, M. S., Movshon, J. A. (Eds.), Computational Models of Visual Processing. MIT Press, Cambridge, MA, pp. 3–20.

Baddeley, R. J., Tatler, B. W., 2006. High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. Vision Research 46, 2824–33.

Ballard, D. H., Hayhoe, M. M., 2009. Modelling the role of task in the control of gaze. Visual Cognition 17 (6-7), 1185–204.

Barth, E., 2000. The minors of the structure tensor. In: Sommer, G. (Ed.), Mustererkennung 2000. Springer, Berlin, pp. 221–228.

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

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

Burt, P. J., Adelson, E. H., 1983. The Laplacian pyramid as a compact image code. IEEE Transactions on Communications 31 (4), 532–540.

Carmi, R., Itti, L., 2006. The role of memory in guiding attention during natural vision. Journal of Vision 6 (9), 898–914.

Dragoi, V., Sur, M., 2006. Image structure at the center of gaze during free viewing. Journal of Cognitive Neuroscience 18 (5), 737–48.

Einhäuser, W., Rutishauser, U., Koch, C., 2008a. 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., 2008b. Objects predict fixations better than early saliency. Journal of Vision 8 (14), 11–26.

Elazary, L., Itti, L., 2008. Interesting objects are visually salient. Journal of Vision 8 (3), 1–15.

Foulsham, T., Underwood, G., 2 2008. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision 8 (2), 1–17.

Haußecker, H., Spies, H., 1999. Motion. In: Jähne, B., Haußecker, H., Geißler, P. (Eds.), Handbook of Computer Vision and Applications. Vol. 2. Academic Press, Ch. 13, pp. 309–96.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., Mack, M., 2007. Visual saliency does not account for eye movements during visual search in real-world scenes. In: van Gompel, R., Fischer, M., Murray, W., Hill, R. (Eds.), Eye Movement Research: Insights into Mind and Brain. Elsevier, pp. 537–562.

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., Koch, C., 2001. Computational modelling of visual attention. Nature Reviews Neuroscience 2 (3), 194–203.

Jähne, B., 1999. Local structure. In: Jähne, B., Haußecker, H. (Eds.), Handbook of Computer Vision and Applications. Vol. 2. Academic Press, Ch. 10, pp. 209–38.

Kumar, M., 2007. GUIDe saccade detection and smoothing algorithm. Tech. Rep. CSTR 2007-03, Stanford.

Land, M. F., Hayhoe, M., 2001. In what ways do eye movements contribute to everyday activities? Vision Research 41, 3559–65.

Mannan, S. K., Ruddock, K. H., Wooding, D. S., 1997. Fixation sequences made during visual examination of briefly presented 2D images. Spatial Vision 11 (2), 157–78.

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

Munn, S. M., Stefano, L., Pelz, J. B., 2008. Fixation-identification in dynamic scenes: Comparing an automated algorithm to manual coding. In: APGV '08: Proceedings of the 5th symposium on Applied perception in graphics and visualization. ACM, New York, NY, USA, pp. 33–42.

Parkhurst, D., Law, K., Niebur, E., 2002. Modeling the role of salience in the allocation of overt visual attention. Vision Research 42, 107–23.

Poynton, C., 2003. Digital Video and HDTV. Morgan Kaufmann Publishers, San Francisco, CA, USA.

Privitera, C. M., Stark, L. W., 2000. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (9), 970–982.

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

Simoncelli, E. P., 1997. Statistical models for images: Compression, restoration and synthesis. In: Proc 31st Asilomar Conf on Signals, Systems and Computers. Vol. 1. IEEE Computer Press, pp. 673–678.

Stelmach, L. B., Tam, W. J., 1994. Processing image sequences based on eye movements. In: Human Vision, Visual processing and Digital Display. Vol. 2179 of Proceedings of the SPIE. IEEE Computer Press, pp. 90–98.

Tatler, B. W., Baddeley, R. J., Gilchrist, I. D., 2005. Visual correlates of fixation selection: effects of scale and time. Vision Research 45, 643–59.

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, 1857–62.

Tatler, B. W., Vincent, B. T., 2009a. The prominence of behavioural biases in eye guidance. Visual Cognition 17 (6-7), 1029–54.

Tatler, B. W., Vincent, B. T., 2009b. Systematic tendencies in scene viewing. Journal of Eye Movement Research 2 (2), 1–18.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., Itti, L., 7 2009. Quantifying center bias of observers in free viewing of dynamic natural scenes. J. Vis. 9 (7), 1–16.

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.

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

Zetzsche, C., 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., Wegmann, B., Oct. 1993. The importance of intrinsically two-dimensional image features in biological vision and picture coding. In: Watson, A. B. (Ed.), Digital Images and Human Vision. MIT Press, pp. 109–38.

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