Variability of eye movements when viewing dynamic natural scenes

Michael Dorr   Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany

Thomas Martinetz   Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany

Karl R. Gegenfurtner   Department of Psychology, Justus Liebig University, Gießen, Germany

Erhardt Barth   Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany

How similar are the eye movement patterns of different subjects when free viewing dynamic natural scenes? We collected a large database of eye movements from 54 subjects on 18 high-resolution videos of outdoor scenes and measured their variability using the Normalized Scanpath Saliency, which we extended to the temporal domain. Even though up to about 80% of subjects looked at the same image region in some video parts, variability usually was much greater. Eye movements on natural movies were then compared with eye movements in several control conditions. "Stop-motion" movies had almost identical semantic content as the original videos but lacked continuous motion. Hollywood action movie trailers were used to probe the upper limit of eye movement coherence that can be achieved by deliberate camera work, scene cuts, etc. In a "repetitive" condition, subjects viewed the same movies ten times each over the course of 2 days. Results show several systematic differences between conditions both for general eye movement parameters such as saccade amplitude and fixation duration and for eye movement variability. Most importantly, eye movements on static images are initially driven by stimulus onset effects and later, more so than on continuous videos, by subject-specific idiosyncrasies; eye movements on Hollywood movies are significantly more coherent than those on natural movies. We conclude that the stimuli types often used in laboratory experiments, static images and professionally cut material, are not very representative of natural viewing behavior. All stimuli and gaze data are publicly available at http://www.inb.uni-luebeck.de/tools-demos/gaze.

Keywords: eye movements, active vision, structure of natural images

Introduction

Eye movements while watching moving images

Humans make several eye movements per second, and where they look ultimately determines what they perceive. Consequently, much research over several decades has been devoted to the study of eye movements, but for technical reasons, this research has mostly been limited to the use of static images as stimuli. More recently, however, an increasing body of research on eye movements on dynamic content has evolved. Blackmon, Ho, Chernyak, Azzariti, and Stark (1999) reported some evidence for certain aspects of the "scanpath theory" (Noton & Stark, 1971) on very simple, synthetic dynamic scenes. Several studies were concerned with modeling saliency, i.e., the contribution of low-level features to gaze control (e.g., Itti, 2005, Le Meur, Le Callet, & Barba, 2007), and found, not surprisingly, that motion and temporal change are strong predictors for eye movements. Tseng, Carmi, Cameron, Munoz, and Itti (2009) quantified the bias of gaze toward the center of the screen and linked this center bias to the photographer's bias to place structured and interesting objects in the center of the stimulus. Carmi and Itti (2006) investigated the role of scene cuts in "MTV-style" video clips and showed that perceptual memory has an effect of eye movements across scene cuts. Cristino and Baddeley (2009) recorded videos with a head-mounted camera while walking down the street; they then compared gaze on these original with that on a set of filtered movies to assess the impact of image features vs. "world salience," i.e., behavioral relevance. Stimuli obtained with a similar setup, a head-mounted and gaze-controlled camera (Schneider et al., 2009), were used by 't Hart et al. (2009). These authors presented the natural video material to subjects either as the original,

continuous movie or in a shuffled, random sequence of 1-s still shots. The distribution of gaze on the continuous stimuli was wider than for the static sequence and also a better predictor of gaze during the original natural behavior. In other studies, the variability of eye movements of different observers was analyzed with an emphasis on how large the most-attended region must be to encompass the majority of fixations in the context of video compression (Stelmach & Tam, 1994; Stelmach, Tam, & Hearty, 1991) and enhancement for the visually impaired (Goldstein, Woods, & Peli, 2007). Marat et al. (2009) evaluated eye movement variability on short TV clips using the Normalized Scanpath Saliency (Peters, Iyer, Itti, & Koch, 2005). Comparing the viewing behavior of humans and monkeys, Berg, Boehnke, Marino, Munoz, and Itti (2009) found that monkeys' eye movements were less consistent with each other than those of humans. Hasson, Landesman et al. (2008) presented clips from Hollywood movies and everyday street scenes to observers while simultaneously recording brain activation and eye movements; both measures showed more similarity across observers on the Hollywood movies (particularly by Alfred Hitchcock) than on the street scenes. However, when playing the movies backward, eye movements remained coherent whereas brain activation did not.

With a few exceptions, these studies used professionally recorded and cut stimulus material such as TV shows or Hollywood movies. Arguably, such stimuli are not representative of the typical input to a primate visual system. Other authors therefore have also studied gaze behavior in real-world tasks, such as driving (Land & Lee, 1994; Land & Tatler, 2001), food preparation (Land & Hayhoe, 2001), and walking around indoors (Munn, Stefano, & Pelz, 2008) and outdoors (Cristino & Baddeley, 2009, Schneider et al., 2009). We here set out to study viewing behavior on natural, everyday outdoor videos.

Purpose of this study

How do people watch dynamic natural scenes? So far, much research on eye movements has either used static natural images or easily accessible, professionally cut video material as stimuli. In this exploratory study, we recorded eye movements from a large number of subjects to investigate various facets of the free viewing of truly naturalistic, uncut scenes. Besides general eye movement parameters such as saccadic amplitude, fixation duration, and the central bias, we also analyze gaze variability, i.e., the similarity between eye movements of different observers. This was motivated by our work on gaze guidance to aid observers in following optimal gaze patterns (Barth, Dorr, Böhme, Gegenfurtner, & Martinetz, 2006). More specifically, we aimed to understand the limits of variability in eye movements observers make on dynamic natural scenes. Intuitively, a very low variability, i.e., a scene on which all observers follow the same gaze

pattern, offers little room to guide the observer's attention; at the same time, a very high variability might indicate a dominance of idiosyncratic viewing strategies that would also be hard to influence. The measurement of gaze variability further allows us to empirically evaluate one particular prediction of the scanpath theory (Noton & Stark, 1971), namely that there exists a hierarchy of eye movement similarity for comparisons among and across subjects and stimuli.

However, most of these observations are merely descriptive and not meaningful per se. To see whether natural movies are a special class of stimuli, we therefore repeated the same analyses with different stimulus types. Each of these control conditions differs from natural movies in one specific aspect, and any differences in viewing behavior can then be attributed to this change.

The obvious defining difference between natural movies and static images is the absence of continuous temporal change in the latter, and we therefore explore the effect of such temporal change on viewing behavior. A straightforward approach would be to follow the common psychophysical paradigm for the collection of eye movements on static images and to present a "random" series of images for several seconds each, and indeed we collect such data as a baseline. However, the comparison of such stimuli with natural movies poses two problems. First, it is not clear what the optimal presentation time should be for the individual static images. If it is too short, obviously not much information can be extracted beyond the very first few fixations; if it is too long, on the other hand, observers might lose interest and resort to idiosyncratic top-down viewing strategies in the absence of sufficient bottom-up stimulation. Second, random series of images are typically used to avoid any potential bias introduced by prior knowledge of the stimulus, i.e., any upcoming stimulus image should be unpredictable by the observer. Contrary to this, natural movies usually are highly predictable. To avoid these two problems, we therefore designed "stop-motion" movies without continuous temporal change that resembled natural movies as closely as possible: they consisted of a sequence of interleaved frames taken from a natural movie in their chronological order (note, however, that a small semantic difference between the movies remains because very short events are not necessarily depicted in the stop-motion stimuli). A similar study to compare static and continuous image presentations was recently undertaken by 't Hart et al. (2009), who took 1-s-long still shots from a set of natural videos and reassembled them into random sequences. However, in their experiment, depicted scenes were not predictable by the previous images, whereas in the present study, most of the scene (the static background, but not moving objects) stayed the same across image transitions.

Another common class of stimuli comprises professionally cut material, such as TV recordings or Hollywood movies. We therefore study the effect of cuts and deliberate camera work by comparing eye movements on

natural, everyday scenes with those on Hollywood action movie trailers.

Finally, we make our stimuli and gaze recordings publicly available to provide a large data set of eye movements on high-resolution natural videos at http://www.inb.uni-luebeck.de/tools-demos/gaze.

Methods

Natural movies

A JVC JY-HD10 HDTV video camera was used to record 18 high-resolution movies of a variety of real-world scenes in and around Lübeck. Eight movies depicted people in pedestrian areas, on the beach, playing mini golf in a park, etc.; three movies each mainly showed either cars passing by or animals; a further three movies showed relatively static scenes, e.g., a ship passing by in the distance; and one movie was taken from a church tower, giving a bird's-eye view of buildings and cars. All movie clips were cut to about 20-s duration; their temporal resolution was 29.97 frames per second and their spatial resolution was 1280 by 720 pixels (NTSC HDTV progressive scan). All videos were stored to disk in the MPEG-2 video format with a bit rate of 18.3 Mbit/s. The camera was fixed on a tripod and most movies contained no camera or zooming movements; only four sequences (three of which depicted animals) contained minor pan and tilt camera motion. A representative sample of still shots is given in Figure 1.

Trailers

The official trailers for the Hollywood movies "Star Wars—Episode III" and "War of the Worlds" were used
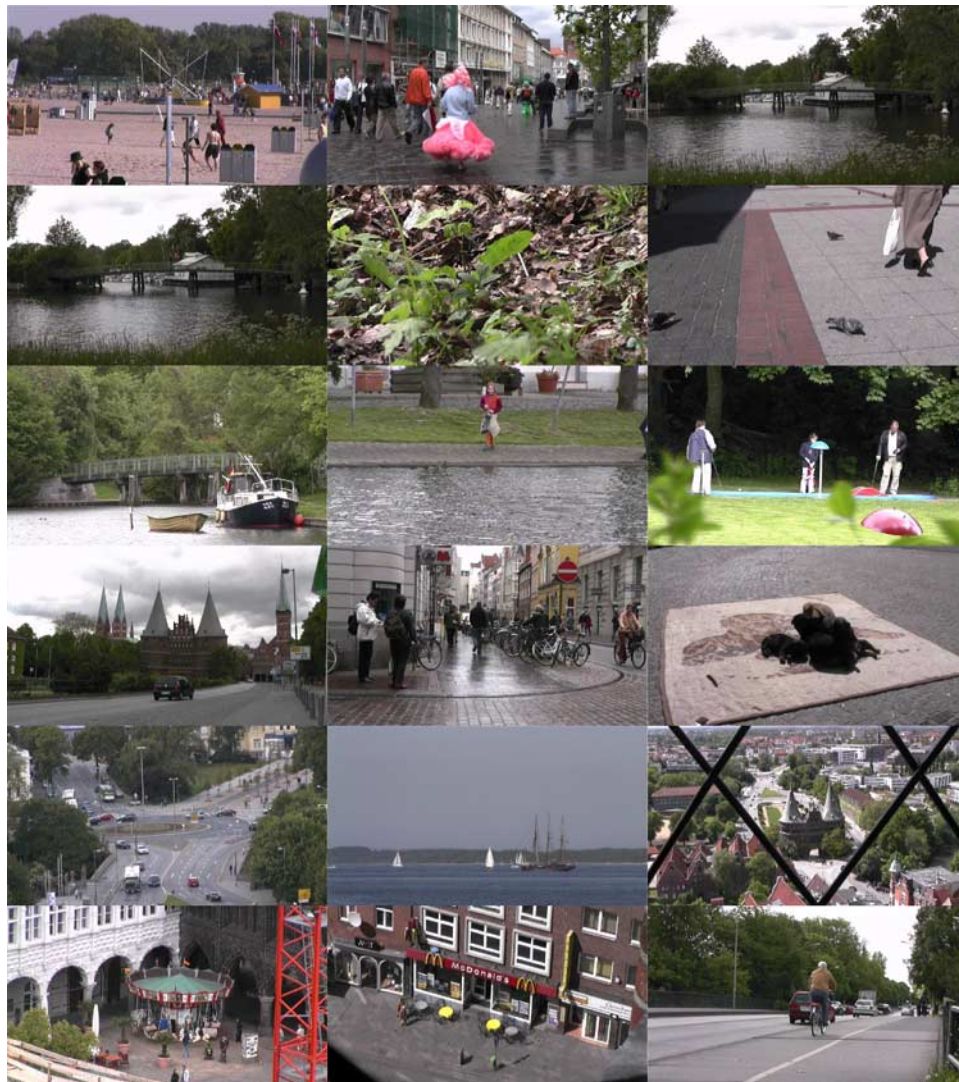

Figure 1. Still shots from all movies used in the natural condition.

for this condition. Both had a duration of about 32 s each and a spatiotemporal resolution of 480 by 360 pixels, 15 fps and 480 by 272 pixels, 24 fps, respectively. Some text on plain background is shown during the first and last few seconds, but in between, these trailers are characterized by a large amount of object motion, explosions, etc., and many scene cuts (21 and 24, respectively). Camera work is deliberately aimed at guiding the viewer's attention, e.g., by zooming in on the face of a scared child. The accompanying sound track was not played during stimulus presentation.

Stop motion

Nine out of the 18 natural movies were also shown in a "stop-motion" condition. Instead of displaying all (around) 600 frames at 30 frames per second, only every 90th frame was displayed for a full 3 s. Thus, the sequence and timing of depicted events was the same as in the original movie but was revealed only in steps similar to scene cuts (note that, typically, the whole scene layout changes with a cut; here, only the position and appearance of moving objects change, whereas the background stays the same).

Static images

Finally, still shots from the nine movies not used in the "stop-motion" condition were used to record eye movements on static images. Similar to the "stop-motion" condition, every 90th frame of a movie was used, but the order was randomized over movies and the temporal sequence of still shots so that subjects could not predict a stimulus from the previous one.

Data recording

All eye-movement recordings were made with an SR Research EyeLink II eye tracker, using information from pupil and corneal reflection to estimate gaze at 250 Hz. This tracker compensates for small head movements, but subjects' heads were still fixated in a chin rest. After an initial binocular calibration, only monocular data from the eye with the smaller validation error were used throughout the experiments (mean validation error of 0.62 deg). Subjects were seated 45 cm away from an Iiyama MA203DT screen that had a width of 40 cm and a height of 30 cm; all stimuli were scaled to make use of the full screen (that was run at a resolution of 1280 by 960 pixels). Since the videos (except for the Hollywood trailers) had an aspect ratio of 16:9 and would not natively fit on the monitor with an aspect ratio of 4:3, they were displayed in the "letterbox" format with black borders below and above such that pixels had the same physical width as

height. Videos covered about 48 by 27 degrees of visual field, and about 26.7 pixels on the screen corresponded to 1 degree of visual angle for the high-resolution movies (1280 by 720 pixels).

For a smooth playback of videos, two computers were used. The first computer ran the eye tracking software; the second was used for stimulus decoding and display. Therefore, gaze recordings and video timing had to be synchronized, for which two strategies were employed. In Experiment 1, the display computer sent a trigger signal to the tracking host via a dedicated ethernet link whenever a new frame was displayed (every 33 ms); these trigger signals and the gaze data were stored to disk using common time stamps by the manufacturer's software. In all other experiments, a three-computer setup was used. Gaze measurements were sent from the tracker across an ethernet link to a relay computer and from there on to the display computer, where independent threads wrote both gaze and video frame time stamps to disk using the same hardware clock. This seemingly complicated setup was necessary because the tracker manufacturer's API requires the network to be constantly monitored (polled) for new gaze samples to arrive, wasting CPU cycles and potentially disturbing the smooth playback of (high-resolution) video. The task of the relay computer thus was to constantly check whether a new gaze sample had arrived from the tracker, using the proprietary software; each sample was then converted to a custom clear-text format and sent on to the display computer, where the receiving thread (performing a "blocking wait" on its network socket) would only run very briefly every 4 ms (at a sampling rate of 250 Hz). Because of the low system load and the low conversion rate, this relay step did not incur a significant delay; the latency of both synchronization approaches is in the single digit millisecond range, and the latter approach has also been used successfully for latency-critical gaze-contingent paradigms (Dorr, 2010).

Subjects were recruited among students (overall age ranging from 18 to 34 years) at the Psychology Department of Giessen University who were paid for their participation. In Experiment 1, data from fifty-four subjects (46 females, eight males) were collected. After an initial nine-point calibration and the selection of the preferred eye, all 18 movies were shown in one block. After every movie presentation, a drift correction was performed; this scheme was also adhered to in the following experiments.

For the repetitive presentation of movies in Experiment 2, 11 subjects came to the laboratory for 2 days in a row. Each day, the trailers and six movies out of the 18 natural movies from Experiment 1 (beach, breite_strasse, ducks_children, koenigstrasse, roundabout, street) were shown five times each in randomized order.

A further 11 subjects participated in Experiment 3 and watched nine "stop-motion" movies, which were created from a subset of the 18 natural movies from Experiment 1

(beach, breite_strasse, bridge_1, bumblebee, ducks_children, golf, koenigstrasse, st_petri_gate, st_petri_mcdonalds). Then, subjects were shown, after another calibration, still shots from the remaining nine movies (bridge_2, doves, ducks_boat, holsten_gate, puppies, roundabout, st_petri_market, street, sea) in randomized order. Still shots were shown for 2 s each.

In all of the above experiments, subjects were not given any specific task other than to "watch the sequences attentively."

Data analysis

Gaze data preprocessing

The eye tracker marks invalid samples and blinks, during which gaze position cannot be reliably estimated. Furthermore, blinks are often flanked by short periods of seemingly high gaze velocity because the pupil gets partially occluded by the eye lid during lid closure, which in turn leads to an erroneous gaze estimation by the tracker. These artifacts were removed and recordings that contained more than 5% of such low confidence samples were discarded. In Experiment 1, this left between 37 and 52 recordings per video sequence and 844 (out of 972) recordings overall (Experiment 2: 707 out of 840; Experiment 3: 627 out of 792).

Saccades are typically extracted from raw gaze recordings based on the high velocity of saccadic samples. However, the choice of an optimal threshold for saccade velocity is difficult: a low threshold might lead to a high false positive rate, i.e., the detection of too many saccades due to microsaccades and impulse noise in the eye tracker measurements; a high threshold, on the other hand, might forfeit information from the beginning and end of saccades, where velocity is still accelerating or decelerating, respectively. Therefore, we labeled saccadic samples in a two-step procedure. To initialize search for a saccade onset, velocity had to exceed a relatively high threshold (138 deg/s) first. Then, going back in time, the first sample was searched where velocity exceeded a lower threshold θ_{off} (17 deg/s) that is biologically more plausible but less robust to noise (both parameters were determined by comparing detection results with a hand-labeled subset of our data). In a similar fashion, saccade offset was the first sample at which velocity fell below the lower threshold again. Finally, several tests of biological plausibility were carried out to ensure that impulse noise was not identified as a saccade: minimal and maximal saccade durations (15 and 160 ms, respectively) and average and maximum velocities (17 and 1030 deg/s, respectively).

Determining fixation periods is particularly difficult for recordings made on dynamic stimuli (Munn et al., 2008). Smooth pursuit eye movements cannot occur on static images and are hard to distinguish from fixations because of their relatively low velocity of up to tens of degrees per second; but even a small, noise-induced displacement in the gaze measurement of just 1 pixel from one sample to the next already corresponds to about 9 degrees per second. However, manual labeling of fixations is not feasible on such large data sets as that of Experiment 1 (about 40,000 fixations); we therefore used a hybrid velocity- and dispersion-based approach (Salvucci & Goldberg, 2000) and validated its parameters on a smaller data set of hand-labeled fixations. After saccade detection, the intra-saccadic samples were extracted. Here, a sliding window of at least 100 ms was moved across the samples until a fixation was detected. This minimum duration of 100 ms ensured that very brief stationary phases in the gaze data were not labeled as fixations. Then, this fixation window was extended until either one of two conditions was met: the maximum distance of any sample in the window to the center of the fixation window exceeded 0.35 deg (this threshold was gradually increased to 0.55 deg with longer fixation duration); or the average velocity from beginning to end of the window exceeded 5 degrees per second. The latter condition served to distinguish pursuit-like motion from noise where sample-to-sample velocities might be high, but velocities integrated over longer time intervals are low because the direction of gaze displacements is random. We also varied the minimum duration to 50 and 150 ms, respectively, and found qualitatively similar results.

Eye movement similarity

A variety of methods has been proposed in the literature to assess the consistency of eye movements across different observers. The fundamental problem is that there is no obvious metric for eye movement similarity since there is no direct (known) mapping from eye position to its perceptual consequences. In practice, there is only a small probability that two observers will fixate exactly the same location at exactly the same time; small spatiotemporal distances between eye positions, however, might have been introduced in the measurement only by fixational instability and the limited eye tracker accuracy and are thus of little practical relevance. For larger distances of more than about 1 degree and a few tens of milliseconds, on the other hand, it is not clear how a similarity metric should scale: is a fixation twice as far also twice as different? How about two fixations to the same location, but of different duration? In the case of our (moving) stimuli, a further problem arises that looking at the same image region at different points in time, e.g., in the background of the scene, might carry a different notion depending on what is (or is not) occurring elsewhere, e.g., in the foreground. As pointed out by Tatler, Baddeley, and Gilchrist (2005), a good similarity metric should be robust to extreme outliers and sensitive not only to location differences but also to differences in the probability of such locations; if all but one of the subjects looked at the same location A and the remaining subject looked at location B, this should be reflected as more coherent than an

even distribution of fixations over A and B. Additionally, hard thresholds should be avoided in order to deal with the inherent spatiotemporal uncertainty in the eye tracker measurements. Finally, an ideal metric would yield an intuitively interpretable result and allow for fine-grained distinctions.

We will now discuss similarity metrics proposed in the literature according to the above criteria and then describe our modification of the Normalized Scanpath Saliency method that will be used in the remainder of this paper.

Several authors have used clustering algorithms to group fixations and then determined what percentage of fixations fell into the main cluster, or how large an image region must be to contain the gaze traces of a certain number of observers (Goldstein et al., 2007; Osberger & Rohaly, 2001; Stelmach et al., 1991). Obviously, these measures yield very intuitive values and are also robust to outliers. However, they might be sensitive to cluster initialization, and even if they were extended to regard the fixations in several clusters, they cannot capture differences in the distribution of fixations across several locations. Furthermore, a fixation can either be counted as inside the cluster or not, which means that a small spatial displacement can have a significant impact on the result. Some clustering algorithms introduce a certain smoothness to overcome this problem, e.g., mean-shift clustering (Santella & DeCarlo, 2004), but the scale of the resulting cluster becomes unpredictable, so that for densely distributed data, even two fixations that are very far apart might be classified as similar.

Another popular approach is to assign a set of letters to image regions and to create a string where the *i*th letter corresponds to the location of fixation *i*. The resulting strings can then be compared by string editing algorithms, which sum penalties for every letter mismatch or other string dissimilarity such as letter insertions or transpositions. Drawbacks of this method are the need for an a priori definition of regions of interest for the string alphabet and of a penalty table; inherently, it cannot distinguish between fixations of different duration. Nevertheless, the string-editing approach has been used successfully on line drawings (Noton & Stark, 1971) and on semi-realistic dynamic natural scenes (Blackmon et al., 1999) and has been extended to handle the case where the order of fixated regions matters (Clauss, Bayerl, & Neumann, 2004).

Mannan, Ruddock, and Wooding (1996) developed a measure to compare two sets of fixations by summing up the distances between the closest pairs of fixations from both sets. This is problematic because the result is dominated by outliers and probability distribution differences are not accounted for.

Hasson, Yang, Vallines, Heeger, and Rubin (2008) cross-correlated horizontal and vertical eye trace components of observers across two presentations of the same movie. The intuitive range of the measure is from −1 for highly dissimilar scanpaths to 1 for exactly the same

scanpaths, with zero indicating no correlation between the traces. However, similarity here is defined relative to the mean position of the eye (which usually also is roughly the center of the screen, see below); this means that two scanpaths oscillating between two fixations in counter-phase, i.e., *ABAB*… and *BABA*… will always be classified as very dissimilar, regardless of the actual distance between *A* and *B*.

Another class of methods operates on fixation maps or probability distributions created by the additive super-position of Gaussians, each centered at one fixation location $\vec{x} = (x, y)$ (to obtain a probability distribution function, a subsequent normalization step is required so that the sum of probabilities over the fixation map equals one). The inherent smoothness of the Gaussians offers the advantage that two fixations at exactly the same location will sum up to a higher value than two closely spaced fixations, whereas very distant fixations will contribute only very little to their respective probabilities. This means that noise both in the visual system and the measurement has only a small impact on the final result; by definition, these methods also are sensitive to location distribution differences. There now are various possibilities to assess the similarity of two fixation maps, which includes both the comparison of two different groups of observers and the comparison of just one observer to another. Since, in practice, fixation maps can only be created for a finite set of locations anyway, the most straightforward difference metric is the sum over a squared pointwise subtraction of two maps (Wooding, 2002). Pomplun, Ritter, and Velichkovsky (1996) have computed the angle between the vectors formed by a linearization of the two-dimensional fixation maps. In the latter study, fixations were also weighted with their duration, a modification that in principle could also be applied to the other fixation map-based measures as well.

An approach based on information theory, the Kullback–Leibler Divergence, was chosen by Rajashekar, Cormack, and Bovik (2004) and Tatler et al. (2005). This measure, which strictly speaking is not a distance metric and needs minor modifications to fulfill metric requirements (Rajashekar et al., 2004), specifies the information one distribution provides given knowledge of the second distribution. The KLD matches all of the above criteria for a good similarity measure with the possible exception of intuitiveness: identical distributions have a KLD of zero, but the interpretation of the (theoretically unbounded) result for non-identical distributions is not straightforward.

For this reason, we use the Normalized Scanpath Saliency (NSS) measure as proposed by Peters et al. (2005). Originally, this measure has been developed to evaluate how closely artificial saliency models match human gaze data, but NSS can be directly applied to assess inter-subject variability as well. The underlying idea is to construct a fixation map by superposition of Gaussians as above, but with a different normalization scheme: mean intensity is subtracted and the resulting

distribution is scaled to unit standard deviation. This has the effect that a random sampling of locations in the NSS map has an expected value of zero, with positive values resulting from fixated locations and negative values from non-fixated regions. To evaluate the similarity of eye movements of multiple observers, it is possible to use a standard method from machine learning, "leave one out." For each observer A, the scanpaths of all other observers are used to create the NSS map; the values of this NSS map are then summed up over all fixations made by A. If A tends to look at regions that were fixated by the other observers, the sum will be positive; for essentially uncorrelated gaze patterns, this value will be zero and it will be negative for very dissimilar eye movements. NSS has been used on videos before (Marat et al., 2009), but only on a frame-by-frame basis, similar to the analysis of static images by Peters et al. (2005). To achieve temporal smoothing, so that slightly shifted fixation onsets are not considered to be dissimilar by a hard cut-off, we extended NSS to the three-dimensional case.

Formally, for each movie and observer $i = 1, …, N, M_i$ gaze positions $\vec{x}_i^j = (x, y, t)$ were obtained, $j = 1, …, M_i$. Then, for each $\vec{x}_i^j$ of the training set of observers $S = \{1, …, k − 1, k + 1, …, N\}$, a spatiotemporal Gaussian centered around $\vec{x}_i^j$ was placed in a spatiotemporal fixation map F:

$$F(\vec{x}) = \sum_{i \in S} \sum_{j=1}^{M_i} G_i^j(\vec{x}), \tag{1}$$

with

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

This fixation map F was subsequently normalized to zero mean and unit standard deviation to compute an NSS map N:

$$N\left(\vec{x}\right) = \frac{F(\vec{x}) - \overline{F(\vec{x})}}{\text{Std}(F)}. \tag{3}$$

Finally, the NSS score was evaluated as the mean of the NSS map values at the gaze samples of test observer k:

$$\text{NSS} = \sum_{j=1}^{M_k} N(\vec{x}_k^j)/M_k, \tag{4}$$

and this was repeated for all possible training sets (i.e., N times with N different test subjects).

The spatiotemporal Gaussian G had parameters $\sigma_x = \sigma_y = 1.2$ deg, $\sigma_t = 26.25$ ms. To evaluate gaze variability over the 20-s time course of the videos, NSS was not computed on the whole movie at once, but on temporal windows of 225-ms length that were moved forward by 25 ms every step. These parameters were chosen to roughly match the size of the fovea and a short fixation and to have a temporal resolution better than one video frame; they were also varied systematically with qualitatively similar results (σ from 0.6 to 2.4 deg, temporal windows from 75 to 325 ms). Because NSS is sensitive to the size of the Gaussian G, all results that are presented in the following were normalized with the inverse of the NSS of a single Gaussian.

Gaze position $\vec{x}$ here refers to the raw gaze samples provided by the eye tracker except for those samples that were labeled as part of a saccade. Because visual processing is greatly reduced during saccades, these saccadic samples are of no practical relevance for the present analysis. In principle, the fixation spots could have been used instead of the raw samples as well, which would have significantly reduced the computational cost of this analysis; however, this might have biased results during episodes of pursuit, where automatic fixation detection algorithms still have problems and potentially ascribe fixations to random positions on the pursuit trajectory. Indeed, it was those movie parts in which many subjects made pursuit eye movements where we informally found eye movements to be particularly coherent. Furthermore, using the raw data allows for a distinction of different fixation durations; two fixations to the same location, but with varying duration will be classified as less similar than two fixations of identical length (given they take place at similar points in time).

In theory, this measure is independent of the number of training samples because it normalizes the training distribution to unit standard deviation. In practice, however, small training set sizes may lead to quantization artifacts; where applicable, we therefore matched the number of training samples when comparing two conditions. This was particularly important for the comparison of "local" and "repetitive," because in the latter condition each scanpath had to be evaluated in terms of a maximum of only four other scanpaths (the stimuli were repeated five times per day). A further consequence is that, in the following, different absolute NSS values are occasionally reported for the same condition (but in the context of different comparisons).

Finally, we ran a comparison of the NSS measure with the Kullback–Leibler Divergence to exclude the possibility that our results might underlie some methodological bias. Even though the NSS analysis yields a more intuitive absolute score, NSS and KLD differ only slightly in their relative results. We computed both NSS and KLD scores over time for all movies in the "local" condition and found

that they are highly correlated ($r = 0.87$, $SD = 0.05$), i.e., both methods approximately mark eye movements on the same video parts as coherent or incoherent, respectively.

Results

Saccadic amplitudes and fixation durations

The distribution of saccadic amplitudes for natural movies and for the other stimulus types is shown in Figures 2 and 3A, respectively.

On natural movies, saccadic amplitudes follow a skewed distribution with a mean of 7.2 deg and a median of 5.5 deg. Looking at the shape of the empirical cumulative distribution function (ECDF) in comparison to that of the other stimulus types, the ECDF for natural movies rises quickly but saturates late. This means that observers tend to make both more small and more large saccades (with amplitudes of less than 5 and more than 10 degrees, respectively) on natural movies, whereas saccades of intermediate amplitudes are less frequent than in the other conditions. In contrast to this, the saccades on Hollywood trailers show the smallest fraction of large amplitudes, e.g., only 7.8% have an amplitude of 12 deg or more (natural movies: 16.7%). All the conditions differ from each other highly significantly (Kolmogorov–Smirnov test, $p < 0.001$), with the only exception that the
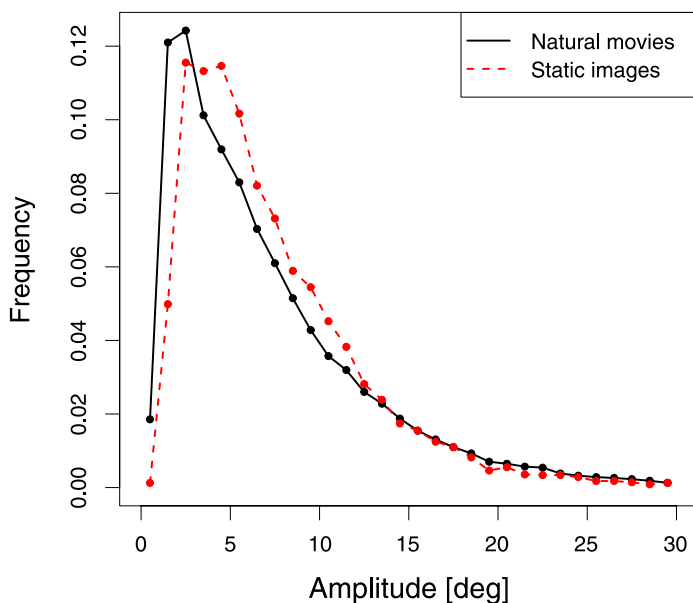

Figure 2. Distribution of saccadic amplitudes on natural movies and static images, which are still shots from the natural movies that were shown in randomized order. Saccades of medium amplitude (4–12 deg) are more frequent in the static images condition, whereas saccades on natural movies have small amplitude (up to 4 degrees) more often.
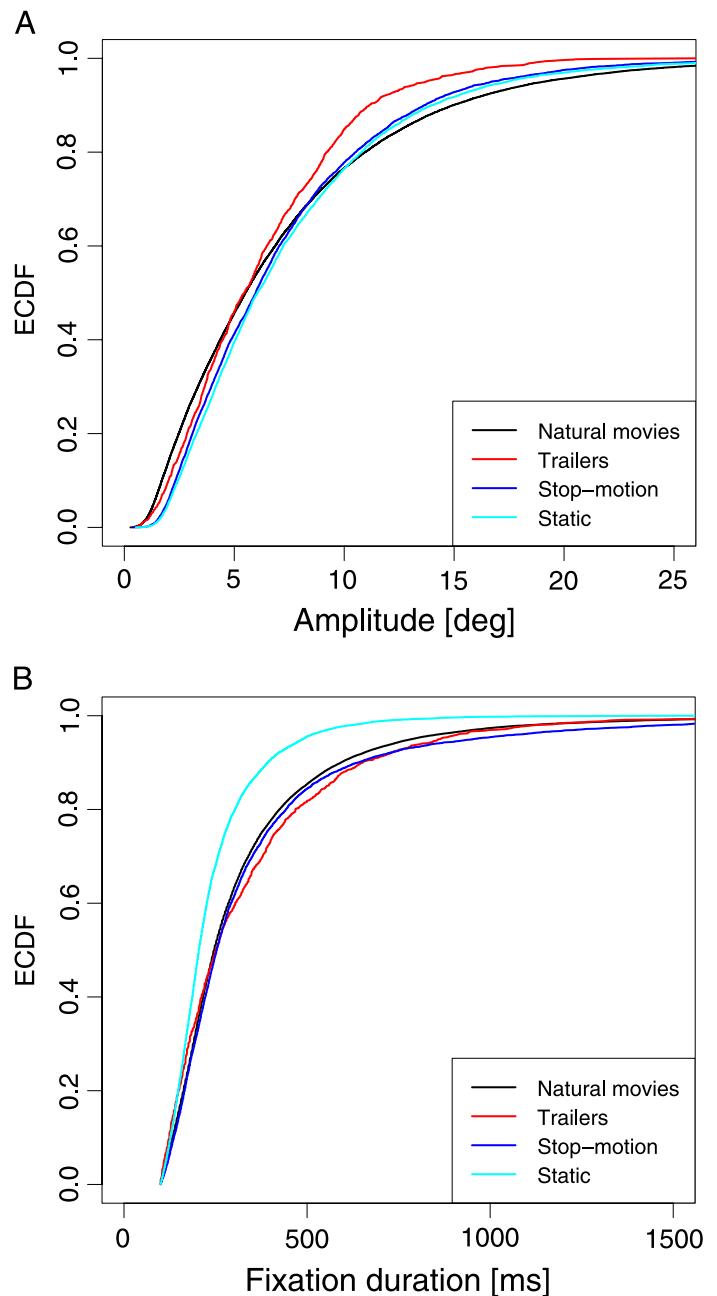

Figure 3. Empirical cumulative distribution functions (ECDFs) of (A) saccadic amplitudes and (B) fixation durations for the different movie types. Natural movies elicit a higher number of either small or large (but not intermediate) saccades and relatively short fixations; trailers and stop-motion movies elicit longer fixations.

difference of saccadic amplitude distributions between stop-motion movies and static images is only weakly significant ($p < 0.02$).

Fixation durations are depicted in Figure 3B. We here also find stimulus type-specific effects. Similar to saccadic amplitudes, the distributions are heavily skewed, which is reflected in a pronounced difference of mean and median values (for natural movies, 326 and 247 ms, respectively).

Fixations on stop-motion movies (354 and 253 ms) and on Hollywood trailers are longer than on natural movies (mean 340, median 251 ms), and the shortest fixations occur on static images (mean 240, median 204 ms).

All these differences are statistically significant (Kolmogorov–Smirnov test, $p < 0.001$).

Center bias of gaze and stimuli

A well-documented property of human viewing behavior is that observers preferentially look at the center of the stimulus, the "center bias" (Buswell, 1935; Parkhurst, Law, & Niebur, 2002; Tatler, 2007; Tseng et al., 2009). This stands to reason since the center of the screen is the most informative location: because of the decline in peripheral acuity of the retina, a fixation to one side of the screen will lead to an even lower resolution on the opposite side of the display. Because at least a coarse "snapshot" of the scene is particularly important during the first few, exploratory fixations, the central bias is strongest directly after stimulus onset (Tatler, 2007). In Figure 4, density estimates are shown for the different stimulus categories. Eye movements on the Hollywood

trailers are the most centered; here, the densest 10% of screen area (15.2 by 8.5 deg) contain about 74% of all fixations, whereas for natural movies this number is only 30% (and 62% of fixations in the densest 30% of the screen). For stop-motion movies, the center bias again is slightly stronger than for natural movies (39% and 75% in the densest 30%) and similar to the center bias of static images (35% and 72%). Here, fixations are redrawn toward the center at every new frame onset (data not shown).

A further common explanation for the center bias of fixations is that there usually is a bias already in the stimuli because photographers (consciously or subconsciously) place objects of interest in the image center. When recording the natural movies, no particular care was taken to avoid such central bias; on the contrary, the goal was to record image sequences "from a human standpoint" to fulfill a common definition of natural scenes (Henderson & Ferreira, 2004), which ruled out any truly random sampling. To assess the magnitude of this potential bias, the spatial distribution of image features was computed (see Figure 5). The feature used here is a generic image descriptor, namely the geometrical invariant K. This invariant encodes those image regions that
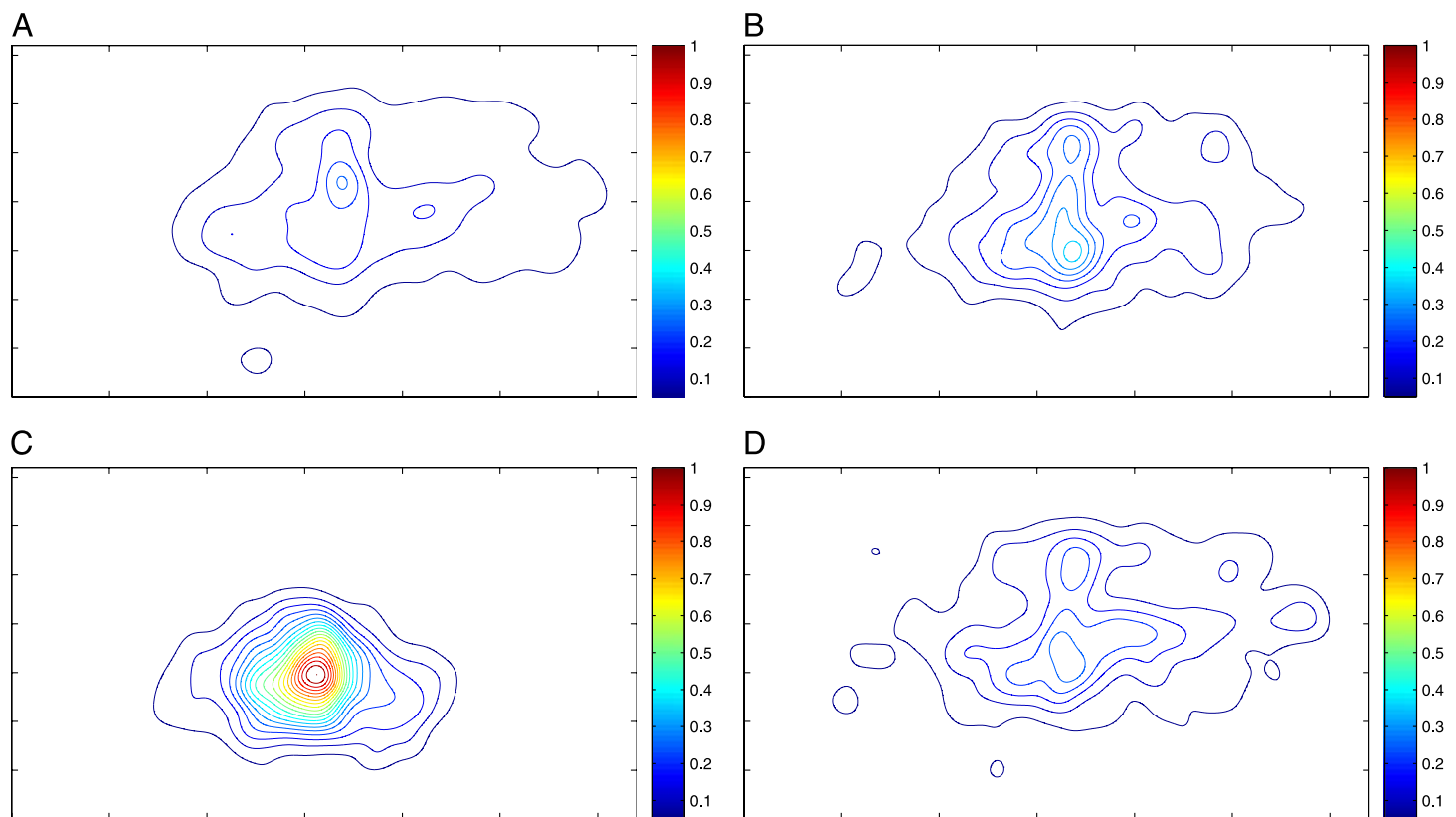

Figure 4. Distribution of gaze in the different conditions, averaged over all movies and subjects. (A) Natural movies. (B) Stop-motion movies. (C) Hollywood trailers. (D) Static images. Probability maps were computed for each condition by the superposition of Gaussians (σ = 0.96 deg) at each gaze sample and subsequent normalization; shown here are contour lines. The distribution of gaze on Hollywood trailers is clearly more centered than in the other conditions. Gaze on natural movies has the widest distribution; in the other conditions, frequent reorienting saccades to the center are elicited by scene cuts (trailers) or frame onsets (stop-motion).
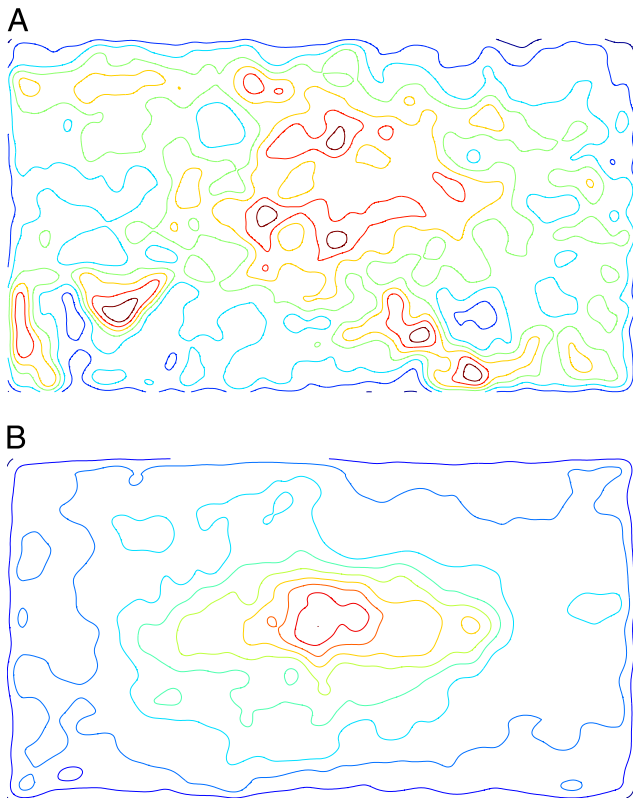
Figure 5. Distribution of spatiotemporal structure for (A) natural movies and (B) Hollywood trailers. Shown here is the average spatial distribution of intrinsically three-dimensional regions as measured by the structure tensor, i.e., transient or non-rigidly moving corners, which have been shown to be highly predictive of eye movements (Vig et al., 2009). The trailers show a stronger bias for placing structure in the center.

change in three spatiotemporal directions, i.e., transient corners, and it has been shown to be correlated with eye movements (Vig, Dorr, & Barth, 2009). Even for the natural movies, there is a certain predominance of central features, but this effect is particularly strong for the Hollywood trailers (in fact, Figure 5B still underestimates the central bias because the frequent scene cuts introduce globally homogeneous temporal transients). It is worth pointing out that the fixation distribution for Hollywood trailers also reflects this central feature distribution; nevertheless, this does not necessarily imply a causal connection. Indeed, Tatler (2007) found that the center bias of fixations on natural static images was independent of spatial shifts in the underlying feature distributions.

Variability of eye movements on natural videos

After these general observations, we will now present results on the variability of eye movements. We start out with the variability across different observers watching the

same natural movie for a single presentation of the stimulus (which Stark coined the "local" condition), see Figure 6 for some prototypical cases in more detail and Figure 7 for an overview. Shown here are one example where variability is very high, one example where most observers look at the same region at least temporarily, and data for one Hollywood movie trailer. Common to all movies is that variability is relatively low (coherence, as shown in the figures, is high) during the first 1 to 2 s due to the central bias of the first few saccades. After this initial phase, gaze patterns for the movie "roundabout" diverge and remain relatively incoherent until the end of the movie; this is not surprising since the scene is composed of a crowded roundabout seen from an elevated viewpoint, i.e., moving objects (cars, pedestrians, cyclists) are distributed almost uniformly across the screen. Nevertheless, gaze patterns are still more similar than the random baseline of different observers looking at different movies. The latter was coined by Stark as the "global" condition, which models stimulus- and subject-independent effects such as the central bias and, therefore, is still higher than pure chance, which would result in an NSS of 0 (mean NSS for "roundabout" is 0.27; for "global," 0.13, $p < 0.001$). NSS for the movie "ducks boat" is shown by the peaked curve in Figure 6. The overall scene is fairly static with two boats moored on a canal but no humans or moving objects (see Figure 1). At about the 5-s mark, a bird flies by, followed by another bird at 10 s; both these events make most observers look at the same location (max NSS 2.61, mean 0.84). Because of pursuit with different gains, ongoing saccades, and the ultimately limited calibration accuracy, it is difficult to strictly determine how many subjects looked at the birds simultaneously; an informal count, however, reveals that an NSS of about 2.5 corresponds to about 80% of subjects looking at the same location, with the remaining fifth of fixations quasi-randomly spread over the remaining scene.

For a comparison, NSS for the trailer "War of the Worlds" is also plotted and exhibits several such highly coherent peaks; on average, gaze on trailers is significantly more coherent than on natural movies (1.37 vs. 0.72, $p < 0.001$).

A further prediction by the scanpath theory is that "idiosyncratic" viewing behavior should be less variable than the "global" condition, i.e., the eye movements of one person watching different movies should be more coherent than those of different persons watching different movies. However, our data do not support this hypothesis; indeed, NSS for the idiosyncratic condition is even lower than for global (0.11 vs. 0.16).

Variability of eye movements on stop-motion movies

Figure 8 shows the average NSS for the stop-motion movies and for the matched set of natural movies (only
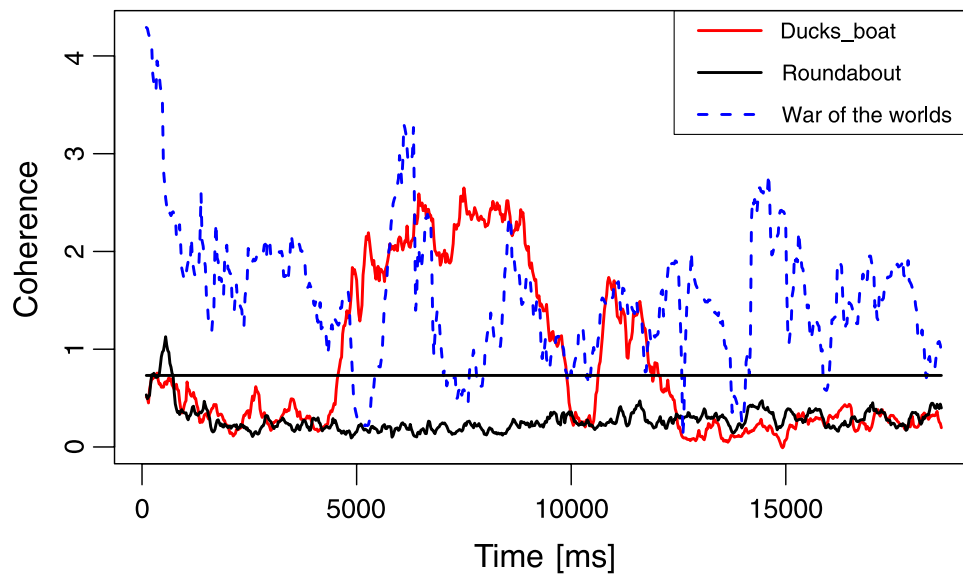
Figure 6. Normalized Scanpath Saliency on natural movies: when a bird flies by (from 5 to 10 s, another bird follows 11–13 s), almost all observers orient their attention to the same spot (red line); in the "roundabout" video with small, moving objects evenly distributed across the scene, eye movements are highly variable and thus have a low coherence (black line). For comparison, the horizontal line denotes the average across all natural movies; the much higher coherence for one Hollywood trailer is also shown (dashed line).
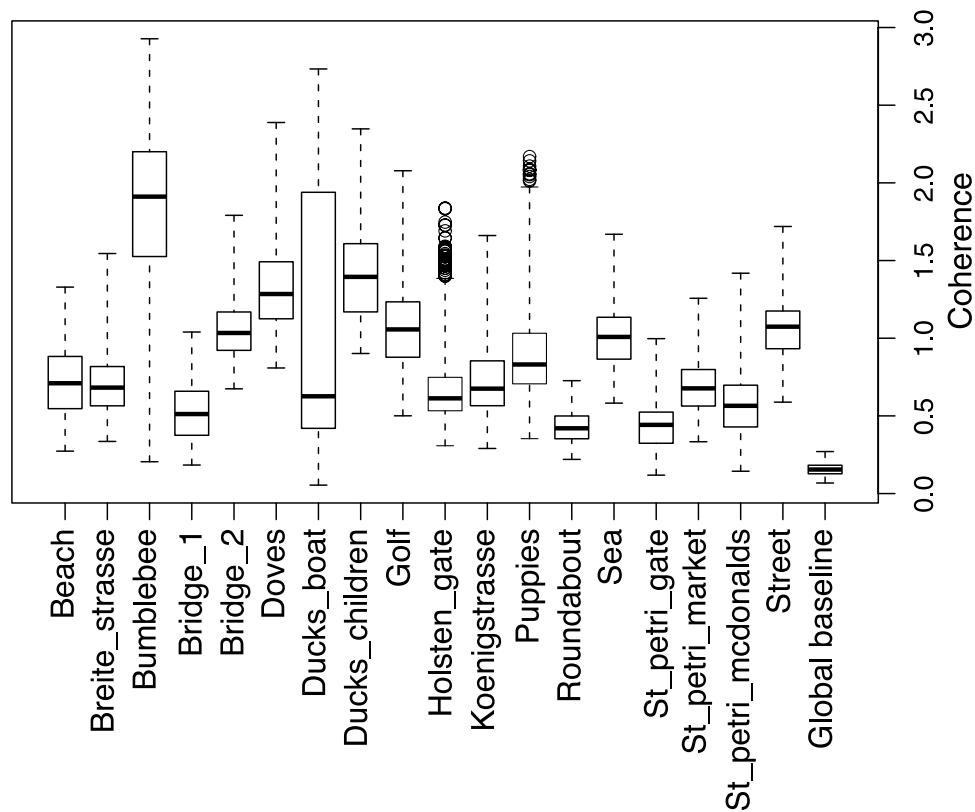

Figure 7. Distribution of Normalized Scanpath Saliency scores for all natural movies. To remove the onset effect where central bias is strongest, data from the first 2.5 s were discarded. The boxes enclose data between the first and third quartiles; whiskers extend to the most extreme point that is at most 1.5 times the inter-quartile distance from the box. For comparison, the rightmost bar shows data for the "global" baseline. (Image size 48 by 27 degrees.)
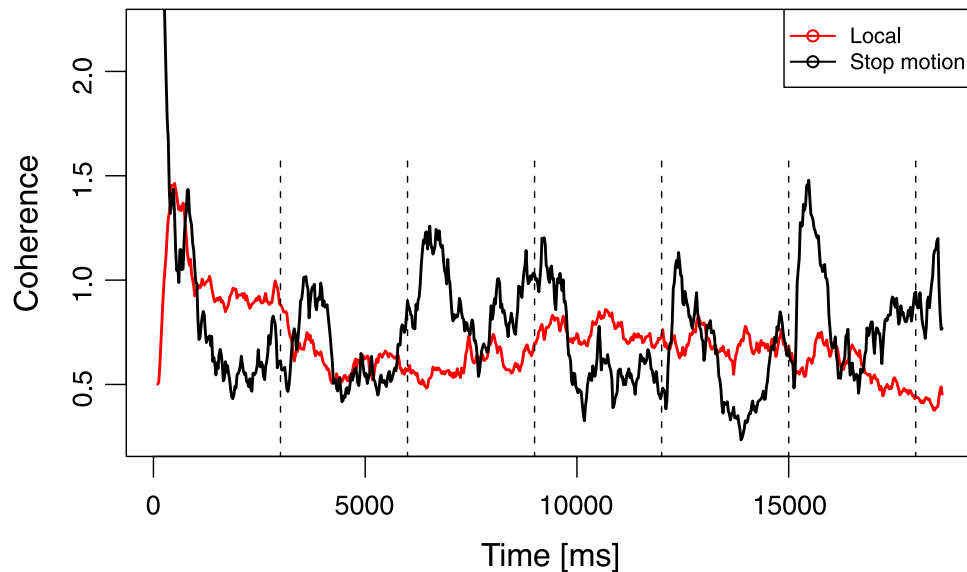
Figure 8. Eye movement coherence on the same set of movies for continuous display (local condition) and for the stop-motion condition, where one frame is shown every 3 s. In the stop-motion condition, coherence spikes after each frame transition and then drops again steeply until the next frame onset. This demonstrates a systematic difference in gaze behavior on static and dynamic stimuli.

nine out of the 18 natural movies were shown in a stop-motion version), with dashed vertical lines denoting the onset of new stop-motion frames. Inter-subject coherence spikes after every frame onset to above the NSS score on the continuous movies; after about 1 to 2 s, however, variability increases and the NSS score drops below that of the continuous case. This observation is statistically significant when pooling the first and last seconds of the 3-s frame intervals: initially, mean stop-motion NSS is higher than local NSS (paired Wilcoxon signed rank test, $p < 0.032$); in the last second, this relationship is reversed ($p < 0.032$).

Variability increases with repetitive viewing of the same stimulus

Several studies have found that repetitive presentation of the same stimulus leads to similar scanpaths (on static images, Foulsham & Underwood, 2008; Hasson, Yang et al., 2008; for simple artificial dynamic scenes, Blackmon et al., 1999). Results from Experiment 2 confirm these earlier findings; indeed intra-subject variability is lower than inter-subject variability (mean NSS for repetitive 0.67, local 0.45 on natural movies; on trailers, repetitive 1.4, local 0.88. For both stimulus types, Kolmogorov–Smirnov test, $p < 0.001$; the local score here is smaller than above because of the matched sample size, see Methods section). One possible confound is that when recording eye movements from one subject in one session, calibration inaccuracies might not be independent across trials, i.e., eye movement coherence might be overesti-

mated; we therefore compared one subject's scanpaths only with scanpaths from the other day of data collection (and indeed found that failure to do so resulted in an even higher increase in eye movement coherence than above). However, pooling together up to five repetitions of a movie may also underestimate how similar gaze patterns evoked by the same stimulus are: the variability of the individual presentations, i.e., for the first, second, … presentation is shown in Figure 9. With increasing number of repetitions, the variability of eye movements across subjects increased ($p < 0.001$, paired Wilcoxon's test). Because the bottom-up stimulus properties were kept constant by definition, this means that individual viewing strategies had an increasing influence. Interestingly, though, this effect was reversed when the stimuli were presented again the following day. The first presentation on the second day (presentation 6 in Figure 9) led to a coherence across subjects comparable to that of the very first presentation (on day one); for subsequent presentations, coherence declined again.

Correlation of basic eye movement parameters with variability/hotspots

Finally, we investigated whether the fixations at locations with high observer similarity, or hotspots, are different from random fixations. Figure 10 shows fixation duration and amplitude of the saccade preceding that fixation as a function of NSS at fixation (relative to the maximum NSS over all movies; because of the small sample size for larger values, the range of NSS is clipped
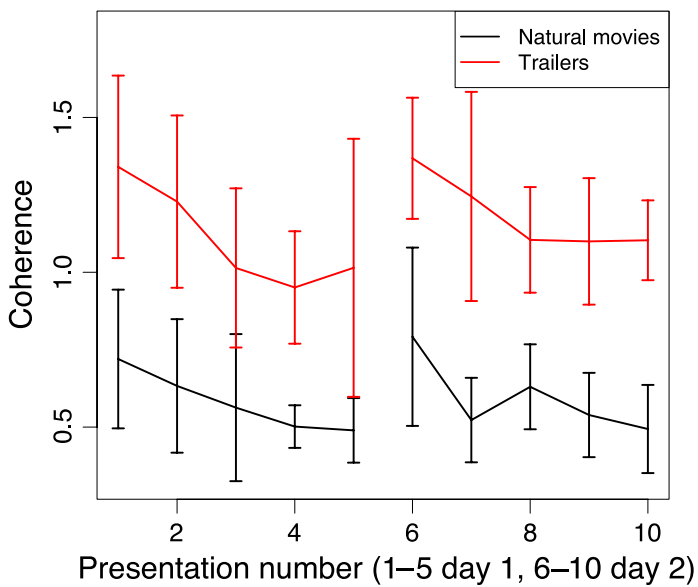
Figure 9. Evolution of coherence during repeated presentation of the same stimulus on natural movies and Hollywood trailers. Each movie was presented five times in one session (trials 1–5) and another five times the following day (trials 6–10). Later presentations at the same day are significantly more variable (paired Wilcoxon's test, $p < 0.001$), but coherence is comparable between both days.

at 70% of the maximum). Locations with high coherence, i.e., locations that were looked at by many observers simultaneously, were examined with fixations of longer duration compared to random locations; also, observers tended to make small saccades toward such highly coherent locations (in both cases, $p < 0.001$). In other words, the image regions that attract attention by a number of people also attract the attention of individual observers for longer and more small, object-investigating saccades.

Discussion

We have collected a large set of eye movements on natural movies and on several other stimulus types. To investigate the role of temporal change in dynamic stimuli, we used stop-motion stimuli that have very similar semantic content as the natural movies but lack continuous motion. To probe the upper limit of eye movement coherence, we used trailers for Hollywood action movies where both low-level features and semantically meaningful objects were deliberately arranged in order to guide the viewer's attention. We found systematic differences throughout these different stimulus types, which emphasize the need to study vision under as natural conditions as possible (Felsen & Dan, 2005). In the following, we will discuss some of these findings in more detail.

General eye movement parameters

Saccadic amplitude and fixation duration are two well-studied, basic eye movement parameters. In line with earlier findings, saccadic amplitudes on natural stimuli follow a heavily skewed distribution biased toward short amplitudes, with a long tail of relatively rare saccades of larger amplitude. In a review of several studies, von
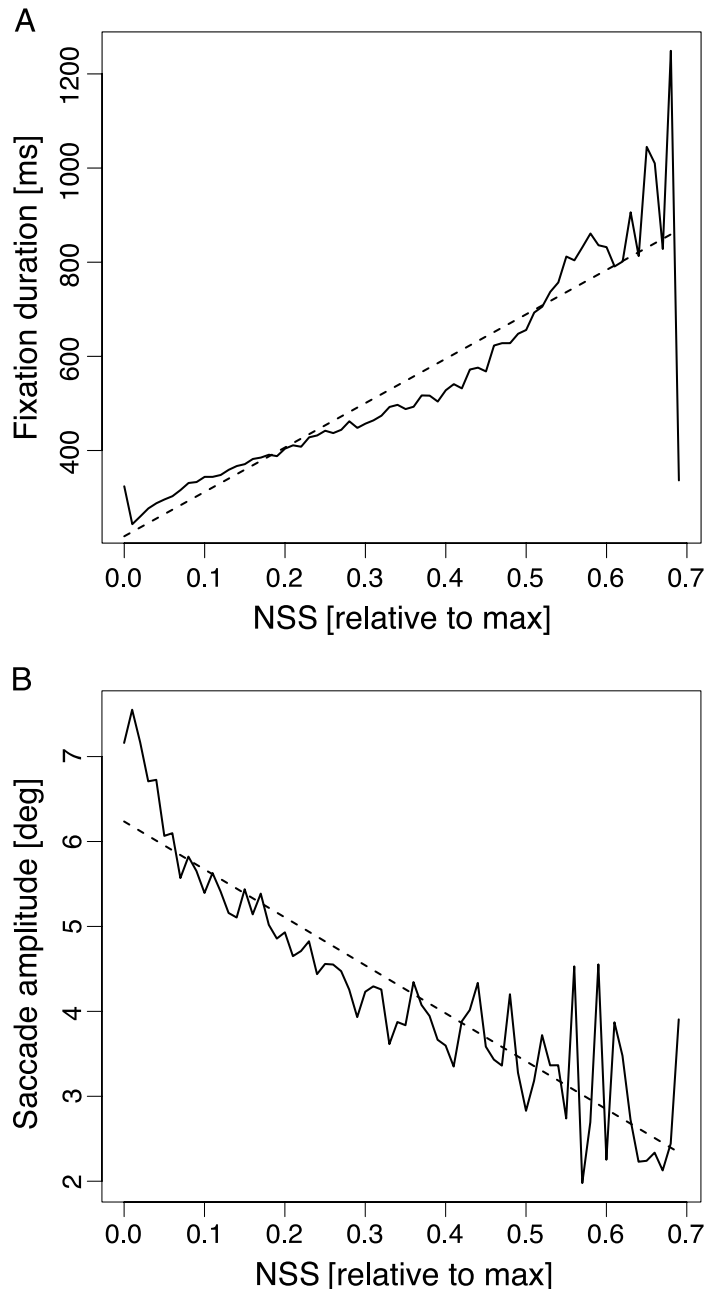

Figure 10. (A) Correlation of NSS values on natural movies and fixation duration. "Hotspots", where many subjects look simultaneously and NSS is high, are fixated for longer periods of time (9.4 ms/%, $R^2 = 0.79$). (B) Correlation of NSS values and saccadic amplitudes. Saccades toward hotspots are typically of small amplitude (−0.05 deg/%, $R^2 = 0.81$).

Wartburg et al. ([2007](#)) found that mean saccadic amplitude scales linearly with stimulus size; the largest natural stimuli reported had an extent of 34 by 26 deg and resulted in a mean saccadic amplitude of 6.3 deg (median 5.2 deg). In contrast to this, we measured only slightly larger saccades (mean 7.2, median 5.5 deg) on more than 30% larger stimuli (image extent of our videos 48 by 27 deg). However, this probably can be explained by the fact that there are obvious mechanical limits to the range of eye movements: under natural viewing conditions, saccades typically are accompanied by a head movement (Einhäuser et al., [2007](#); Guitton & Volle, [1987](#); Morasso, Bizzi, & Dichgans, [1973](#)), but in the present experiments, these were suppressed by a chin rest. When comparing the distributions of saccadic amplitudes across the different stimulus types, eye movements on natural movies comprised more either small or large saccades, with fewer saccades of intermediate amplitude than in the other conditions. Apparently, viewing behavior on natural movies can be characterized by occasional larger jumps between clusters of interesting objects, which are then examined in detail by small, intra-object saccades; this effect has been observed before (for static images, see Tatler & Vincent, [2008](#); for vision during natural tasks, see Land, Mennie, & Rusted, [1999](#)), but we here find it to be more pronounced for movies than for static images. Such a phenomenon could be related to the "two modes of visual processing" described by Pannasch and Velichkovsky ([2009](#)) and Velichkovsky, Dornhoefer, Pannasch, and Unema ([2000](#)), e.g., that longer saccades are often followed by fixations of shorter duration (ambient processing) and shorter saccades are followed by longer fixations (focal processing). On Hollywood trailers, the smallest fraction of large amplitudes was observed; here, the producers deliberately capture the viewer's attention in the center of the screen, using special effects such as explosions, tracking shots, etc., so that there is little incentive for large saccades toward the periphery. This was also reflected in the fact that saccades on this type of movie showed the highest center bias.

Multiple studies have found that fixation duration varies with task (Canosa, [2009](#); Loschky, McConkie, Yang, & Miller, [2005](#); Tatler, Baddeley, & Vincent, [2006](#)). We found the shortest fixation durations on static images, which possibly can be explained by an artifact of the experimental setup: the short presentation time of static images puts pressure on the subjects to quickly scan the image before it disappears again. A small but significant difference was found between natural and stop-motion movies. Despite the very similar semantic content of these stimuli types, fixations on stop-motion movies were slightly longer; intuitively, however, one might expect shorter fixations because of the cuts that elicit saccades. Looking at the time course of this difference in detail, we indeed find that average fixation duration is slightly lower on stop-motion movies in the first second after every cut but then increases during the following 2 s of static frame presentation. The predictability of the individual frames in the stop-motion condition thus led to less exploratory viewing behavior than on the static images that were presented in random order.

Variability

Not surprisingly, we found that eye movements of several observers on the same natural movie are less variable than eye movements on different movies; in other words, that eye movements are at least partially determined by the visual input. This effect was even stronger for professionally cut Hollywood trailers.

It is a well-established fact that the consistency in fixation locations between observers decreases with prolonged viewing (Tatler et al., [2005](#)). Some authors have argued that the direct contribution of low-level saliency to the choice of fixation targets decreases with viewing time (Itti, [2006](#); Parkhurst et al., [2002](#)), while others, e.g., Tatler et al. ([2005](#)), argue that only the top-down strategy changes (that picks targets from a low-level defined set of candidate locations). Cristino and Baddeley ([2009](#)), however, found that eye movements on slightly filtered movie versions were as predictive of those on the original movie as were eye movements on heavily (and differently) filtered versions and reasoned that it is not image saliency but the behavioral relevance of an image region that draws attention (such as the curb for navigation, or looking at other pedestrians for obstacle avoidance). Nevertheless, we here found a systematic difference in viewing behavior on dynamic, i.e., more natural stimuli compared to static images predominantly used in eye movement studies. Whereas the first few fixations on the static images used in our stop-motion movies are heavily influenced by stimulus onset and drawn toward the center of the stimulus, viewing after as little as 1.5 s becomes unnatural and idiosyncratic in the absence of continuous temporal change. 't Hart et al. ([2009](#)) found a similar result in a recent study. They presented their subjects with a random sequence of still shots from a set of videos for 1 s each; these still shots lacked continuous motion and elicited more centered eye movements than the original videos. The time course of such gaze behavior led the authors to interpret this finding as a dominance of stimulus onset effects. We here confirm this finding for longer presentation times (3 s instead of 1 s) and extend it to a case where scenes are highly predictable; even though the still shots are taken from the same movie and presented in their correct chronological sequence, frame transitions still elicit reorienting responses toward the center.

Conclusion

We have extended the study of variability of eye movements to the temporal domain and natural videos

and measured basic eye movement parameters on a range of different stimulus categories. We investigated the variability introduced by the temporal dynamics of a stimulus using novel "stop-motion" stimuli and found that briefly presented static images, as used in common psychophysical paradigms, are a special case and not very representative of natural human viewing behavior. Gaze patterns on professionally cut Hollywood trailers were also different from those on natural movies. In particular, the trailers evoked very similar eye movements among observers and showed the strongest bias for the center of the screen. We also put to test one aspect of the "scanpath theory" on natural videos and found that repetitive viewing of the same stimulus by the same observer elicited more coherent eye movements than single stimulus presentations from different observers. However, we did not find evidence for idiosyncratic viewing patterns of the same subject across different movies.

In summary, we would like to highlight the importance of studying vision under naturalistic conditions. Eye movements are presumably optimized to deal with natural scenes, and dynamic features in a scene have a major effect on viewing behavior. To encourage further research on eye movements made on dynamic natural scenes, we provide all videos and gaze data online at http://www.inb.uni-luebeck.de/tools-demos/gaze.

Acknowledgments

Commercial relationships: none.
Corresponding author: Michael Dorr.
Email: dorr@inb.uni-luebeck.de.
Address: Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany.

References

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: Proceedings of SPIE* (vol. 6057). Invited contribution for a special session on Eye Movements, Visual Search, and Attention: A Tribute to Larry Stark. Bellingham, WA: SPIE Press.

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision, 9*(5):19, 1–15, http://www.journalofvision.org/content/9/5/19, doi:10.1167/9.5.19. [PubMed] [Article]

Blackmon, T. T., Ho, Y. F., Chernyak, D. A., Azzariti, M., & Stark, L. W. (1999). Dynamic scanpaths: eye movement analysis methods. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human vision and electronic imaging IV: SPIE proceedings* (vol. 3644, pp. 511–519). Bellingham, WA: SPIE Press.

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology of perception in art*. Chicago: University of Chicago Press.

Canosa, R. L. (2009). Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception, 6,* 1–34.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision, 6*(9):4, 898–914, http://www.journalofvision.org/content/6/9/4, doi:10.1167/6.9.4. [PubMed] [Article]

Clauss, M., Bayerl, P., & Neumann, H. (2004). A statistical measure for evaluating regions-of-interest based attention algorithms. In C. E. Rasmussen, H. H. Bülthoff, B. Schölkopf, & M. A. Giese (Eds.), *Pattern recognition: Lecture notes in computer science* (vol. 3175, pp. 383–390). Hiedelberg, Germany: Springer.

Cristino, F., & Baddeley, R. (2009). The nature of the visual representations involved in eye movements when walking down the street. *Visual Cognition, 17,* 880–903.

Dorr, M. (2010). Computational models and systems for gaze guidance. PhD thesis, University of Lübeck, Germany, http://d-nb.info/1004914482.

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., et al. (2007). Human eye–head co-ordination in natural exploration. *Network: Computation in Neural Systems, 18,* 267–297.

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience, 8,* 1643–1646.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2):6, 1–17, http://www.journalofvision.org/content/8/2/6, doi:10.1167/8.2.6. [PubMed] [Article]

Goldstein, R. B., Woods, R. L., & Peli, E. (2007). Where people look when watching movies: Do all viewers

look at the same place? *Computers in Biology and Medicine, 3,* 957–964.

Guitton, D., & Volle, M. (1987). Gaze control in humans: Eye–head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of Neurophysiology, 58,* 427–459.

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

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience, 28,* 2539–2550.

Henderson, J. M., & Ferreira, F. (Eds.) (2004). *The interface of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press.

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

Itti, L. (2006). Quantitative modeling of perceptual salience at human eye position. *Visual Cognition, 14,* 959–984.

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

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature, 369,* 742–744.

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,* 1311–1328.

Land, M. F., & Tatler, B. W. (2001). Steering with the head: The visual strategy of a racing driver. *Current Biology, 11,* 1215–1220.

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

Loschky, L. C., McConkie, G. W., Yang, J., & Miller, M. E. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition, 12,* 1057–1092.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10,* 165–188.

Marat, S., Phuoc, T. H., Granjon, L., Guyader, N., Pellerin, D., & Gúerin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision, 82,* 231–243.

Morasso, P., Bizzi, E., & Dichgans, J. (1973). Adjustment of saccade characteristics during head movements. *Experimental Brain Research, 16,* 492–500.

Munn, S. M., Stefano, L., & Pelz, J. B. (2008). Fixation-identification in dynamic scenes: Comparing an automated algorithm to manual coding. In S. Creem-Regehr & K. Myszkowski (Eds.), *APGV'08: Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization* (pp. 33–42). New York: ACM.

Noton, D., & Stark, L. (1971). Eye movements and visual perception. *Scientific American, 224,* 34–43.

Osberger, W., & Rohaly, A. M. (2001). Automatic detection of regions of interest in complex video sequences. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human vision and electronic imaging VI* (vol. 4299). Bellingham, WA: SPIE Press.

Pannasch, S., & Velichkovsky, B. M. (2009). Distractor effect and saccade amplitudes: Further evidence on different modes of processing in free exploration of visual images. *Visual Cognition, 17,* 1109–1131.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123.

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

Pomplun, M., Ritter, H., & Velichkovsky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception, 25,* 931–948.

Rajashekar, U. Cormack, L. K., & Bovik, A. C. (2004). Point of gaze analysis reveals visual search strategies. In B. E. Rogowitz & T. N. Pappas (Eds.), *Proceedings of SPIE Human Vision and Electronic Imaging IX* (vol. 5292, pp. 296–306). Bellingham, WA: SPIE Press.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In A. T. Duchowski (Ed.), *ETRA'00: Proceedings of the 2000 Symposium on Eye Tracking Research and Applications* (pp. 71–78). New York: ACM.

Santella, A., & DeCarlo, D. (2004). Robust clustering of eye movement recordings for quantification of visual interest. In A. T. Duchowski & R. Vertegaal (Eds.), *ETRA'04: Proceedings of the 2004 Symposium on Eye Tracking Research and Applications* (pp. 27–34). New York: ACM.

Schneider, E., Villgrattner, T., Vockeroth, J., Bartl, K., Kohlbecher, S., Bardins, S., et al. (2009). EyeSeeCam: An eye movement-driven head camera for the examination of natural visual exploration. *Annals of the New York Academy of Sciences, 1164,* 461–467.

Stelmach, L. B., & Tam, W. J. (1994). Processing image sequences based on eye movements. In B. E.

Rogowitz & J. P. Allebach (Eds.), *Human vision, visual processing and digital display: Proceedings of the SPIE* (vol. 2179, pp. 90–98). Washington, DC: IEEE Computer Press.

Stelmach, L. B., Tam, W. J., & Hearty, P. J. (1991). Static and dynamic spatial resolution in image coding: An investigation of eye movements. In B. E. Rogowitz, M. H. Brill, & J. P. Allebach (Eds.), *Human vision, visual processing and digital display II: Proceedings of the SPIE* (vol. 1453, pp. 147–152). Washington, DC: IEEE Computer Press.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

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., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research, 2,* 1–18.

't Hart, B. M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P., et al. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition, 17,* 1132–1158.

Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009) Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7):4, 1–16, http://www.journalofvision.org/content/9/7/4, doi:10.1167/9.7.4. [PubMed] [Article]

Velichkovsky, B. M., Dornhoefer, S. M., Pannasch, S., & Unema, P. J. A. (2000). Visual fixations and level of attentional processing. In A. T. Duchowski (Ed.), *ETRA'00: Proceedings of the 2000 Symposium on Eye Tracking Research and Applications* (pp. 79–85). New York: ACM.

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

von Wartburg, R., Wurtz, P., Pflugshaupt, T., Nyffeler, T., Lüthi, M., & Müri, R. (2007). Size matters: Saccades during scene perception. *Perception, 36,* 355–365.

Wooding, D. S. (2002). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers, 34,* 518–528.