Cogn Comput
DOI 10.1007/s12559-010-9061-4

Eye Movements Show Optimal Average Anticipation with Natural Dynamic Scenes

Eleonora Vig · Michael Dorr · Thomas Martinetz · Erhardt Barth

Received: 28 April 2010 / Accepted: 28 July 2010

Abstract A less studied component of gaze allocation in dynamic real-world scenes is the time lag of eye movements in responding to dynamic attention-capturing events. Despite the vast amount of research on anticipatory gaze behaviour in natural situations, such as action execution and observation, little is known about the predictive nature of eye movements when viewing different types of natural or realistic scene sequences. In the present study, we quantify the degree of anticipation during the free viewing of dynamic natural scenes. The cross-correlation analysis of image-based saliency maps with an empirical saliency measure derived from eye movement data reveals the existence of predictive mechanisms responsible for a near-zero average lag between dynamic changes of the environment and the responding eye movements. We also show that the degree of anticipation is reduced when moving away from natural scenes by introducing camera motion, jump cuts, and film-editing.

Keywords Eye movements · Anticipatory gaze behaviour · Natural dynamic scenes · Saliency maps

E. Vig (✉) · M. Dorr · T. Martinetz · E. Barth
Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany
e-mail: vig@inb.uni-luebeck.de

M. Dorr
Schepens Eye Research Institute, Department of Ophthalmology, Harvard Medical School, 20 Staniford Street, Boston, MA 02114, USA

Introduction

Over the last decades, much research has explored the factors that drive eye movements during the viewing of natural, real-world scenes. While most work on gaze allocation in naturalistic scenes has dealt with static stimuli, the study of Itti [15] was among the first to confirm on real-world complex videos that humans look at video regions of higher bottom-up salience than would be expected by chance. Authors found that motion and image transients are more predictive for eye movements than static features, such as colour, intensity, or orientation. Moreover, Carmi and Itti [4] have shown on MTV-style video clips that dynamic visual cues can play an important causal role in drawing attention. 't Hart et al. [29] went further and used recordings of a mobile eyetracking setup to replay the visual input (during in- and outdoor exploration) in the laboratory, under head-fixed viewing conditions. The study showed that gaze recorded in the lab can predict reasonably well eye positions in the real-world, but the temporal continuity of the scene is of importance. Eye movements have been collected and examined on a wide variety of dynamic realistic stimulus types (i.e. video categories). Gaze allocation has been studied while people watched Hollywood movies [11, 28], video games, or even driving scenes [7, 31].

The present work is aimed to focus on a less studied component of gaze allocation in dynamic real-world scenes: it investigates the average time lag of eye movements during the free viewing of natural or realistic videos.

Due to the anatomical structure of the eye, a sophisticated oculomotor system is needed to direct the fovea, the small high-resolution area of the retina, to regions of interest within the periphery. This is achieved by saccades—rapid eye movements by which we shift our line of

Published online: 14 August 2010

sight. However, the required neural processing introduces a certain delay until the oculomotor system reacts to a visual stimulus. In a typical laboratory setup, it takes about 200 ms until a saccade is made towards a spatially and temporally unpredictable target [2, 5]. This delay can, in principle, obstruct immediate reaction to potentially critical events in everyday life. Yet, we are not hindered in our daily activities by this inherent lag in the visual feedback, most likely due to anticipation of the course of future events.

Early studies have shown the existence of predictive mechanisms if the target's spatial and temporal characteristics, such as amplitude, direction, and onset, are known a priori. For example, anticipatory saccades, with near-zero or even negative latencies, occur when the target systematically moves back and forth between two fixed locations [8, 27]. Unlike in passive, free-viewing scenarios, a number of experiments investigating eye movements during natural interaction with the environment, have found that the human visual system can benefit from expectations and prior knowledge about the surrounding world: eye movement patterns were examined during the performance of well-learned everyday tasks, such as tea- and sandwich-making [19, 22], handwashing [24], and driving [20]. These studies show that in everyday life eye movements are "proactive, anticipating actions rather than just responding to stimuli" [18]. That is, saccades are often made to predicted locations of expected events even in advance of the event. However, these authors stress that eye movement patterns are highly task-specific: they seem to be influenced by some learned internal model of the performed actions [13, 18]. More recent experiments examined gaze patterns in more dynamic environments, during the execution of actions requiring specific physical skills. These studies confirm the proactive nature of eye movement control. For example, in the ball game cricket, experienced batsmen make high-precision anticipatory saccades to predict the ball's trajectory [21]. Similar results were reported when gaze patterns of elite-shooters [26] and experienced squash players [6] were compared to that of novices. The main conclusion of these studies is that these predictive mechanisms may have evolved by learning the dynamic properties of the surrounding world (here, of the ball). These studies present evidence for predictive mechanisms during the execution of different natural tasks.

Furthermore, anticipation is found also during action observation. Experiments have shown that predictions are made also during the viewing of block stacking and model building tasks [9, 10, 23]. When subjects watch a block stacking task, their gaze anticipates the hand movements of the actor, as if they performed the task themselves.

Based on these findings, the present study addresses the question: to what extent does the human visual system benefit from predictive mechanisms during the free viewing of dynamic natural scenes? Furthermore, how does the visual system adjust the degree of predictability? Our interest in these questions arose in connection with our work on eye movement prediction in dynamic real-world environments. Our simple model of visual saliency uses low-level image properties to predict where humans look in dynamic scenes [32]. A critical issue, often neglected in the design of computational saliency models, is when exactly a salient location is fixated. Depending on the degree of predictability of a salient event, saccades may lag, coincide with, or even anticipate the event. In the present study, we quantify the average time lag between salient events in the natural scene and the eye movements responding to the events. Insights into these questions may have important implications for the design of computational models of visual saliency.

Most models of visual attention are based on the concept of a saliency map, which topographically encodes stimulus conspicuity [17]. At every location in the visual scene, simple features such as edges, contrast, or colour are stored in feature-activation maps and then combined together to form a global saliency map. In the following, we will refer to these maps as "analytical saliency maps", as they are computed analytically by means of local low-level image properties.

Outline of the Approach

To measure the delay between events in a video and saccades towards these events, we temporally aligned analytical saliency maps with an "empirical" saliency measure based on real gaze data (see Fig. 1 for a sketch of the analysis). According to our hypothesis, a dynamic event, such as the appearance of an object (e.g. the car on the left), would yield a local spatio-temporal maximum in our dynamic analytical saliency measures (middle row in Fig. 1). After a certain time, any saccade made towards this dynamic event would, in turn, yield a local spatio-temporal maximum in the empirical saliency map (bottom row). To determine this time lag, analytical and empirical saliency can be cross-correlated, that is, multiplied when shifted against each other in time by varying amounts. Therefore, the time lag at which the cross-correlation function reaches its maximum denotes the average response delay.

Here, we use the analytical saliency measure introduced in [32] to predict gaze-capturing events. It is a simple and fast alternative to state-of-the-art saliency algorithms (e.g. [16]), requiring less parameter tuning and providing comparable prediction results [33]. A brief overview of the model is given in Sect. 2.2.

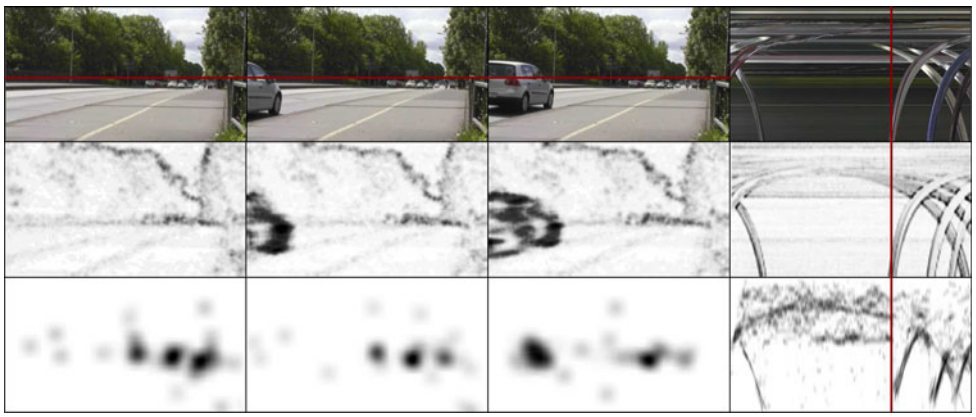

Fig. 1 Different (x, y) and (x, t) slices of the spatio-temporal volume of a video and corresponding analytical and empirical saliency maps. *Top row* Three neighbouring (but not consecutive) frames (i.e. (x, y) slices) of a video and a horizontal (x, t) slice of the movie cube at fixed $y = 400$ pixels (*red horizontal line* in the spatial screenshots). For the (x, t) slice, time axis is along the horizontal direction. Here, the time of the sudden appearance of the car is marked by the *red vertical line*. *Middle row* Corresponding frames from the analytical saliency map (invariant K of the structure tensor). The sudden appearance of a car from left yields a strong response in the analytical saliency map. *Bottom row* Empirical saliency map based on raw gaze samples of all subjects. Attention is drawn to the salient event (appearance of the car in the scene), but the eyes arrive at the target only after a certain time lag. Saccadic responses yield a spatio-temporal maximum in the empirical saliency map. The two saliency maps can be cross-correlated, when shifted against each other in time, to determine the average time lag between the two maps (colour figure online)

Methods

Stimuli and Data Collection

Natural Dynamic Scenes with Static Camera

In a free-viewing task, 54 participants watched 18 high-resolution (HDTV standard, 1,280 × 720 pixels, 29.97 Hz) natural outdoor video sequences with a duration of about 20 s each. The clips depicted real-world outdoor scenes: people in a pedestrian area (on the beach, playing in a park), populated streets and roundabouts, animals. The videos were displayed at 45 cm viewing distance and at a visual angle of 48 × 27 degrees, so that the maximum spatial frequency of the display was 13.3 cycles per degree. The commercially available videographic eye tracker EyeLink II was used to record gaze data at 250 Hz. The experiment was conducted using two computers, the first of which was used to display the videos, while the second ran the eye tracking software. Recordings were performed in Karl Gegenfurtner's lab at the Dept. of Psychology of Giessen University. To synchronize gaze recording and video timing, the display of a new movie frame was signalled to the tracking computer with a UDP packet sent over a dedicated Gigabit Ethernet link; there, these packets were stored together with the gaze data using common timestamps by the manufacturer's software. From these recordings, about 40,000 saccades were extracted using a dual-threshold velocity-based procedure [3].

Moving Camera and Edited Videos

As a control data set, we use the CRCNS eye-1 database[1] [14], a benchmark data set for the analysis of eye movement data on complex video stimuli. The database consists of 100 video clips (640 × 480 pixels, 30 Hz) and the gaze data of eight human subjects freely viewing these videos. In the current study, we used a subset of 50 clips and their corresponding eye traces called "original" experiment [14, 15]. The sequences include indoor and outdoor scenes, television broadcasts (commercials, sports, news, talkshows, etc.), and video games. In case of all videos, transitions between shots are achieved by camera movements, such as panning, tilting, and zooming. Besides these, transitions are realized in television clips (23 out of 50) through jump cuts and special video editing effects, such as fading, dissolving, and wiping. Text overlays are also common. The total number of saccades extracted from the raw gaze data with the aforementioned saccade detection procedure was about 11,000.

Analytical Saliency Measures

In search of saccade triggering stimuli, we use a simple measure to detect salient events in the video. We have previously shown that the intrinsic dimension of the visual signal can be used to predict saccade targets in natural dynamic scenes [32]. The intrinsic dimension describes

[1] http://www.crcns.org/data-sets/eye/eye-1.

how many parameters (or degrees of freedom) are necessary to locally represent the observed data. Thus, locally a movie can be *i0D* (static and homogeneous regions), *i1D* (stationary edges, uniform regions that change in time), *i2D* (corners, edges that change in time), and *i3D* (space-time corners, non-constant motion).

In estimating the intrinsic dimension, we use the structure tensor $\mathbf{J}$, defined in terms of the spatio-temporal gradient of the video intensity function $f(x, y, t)$ ($f : \mathbb{R}^3 \rightarrow \mathbb{R}$):

$$\mathbf{J} = \int_{\Omega} \begin{bmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{bmatrix} \mathrm{d}\Omega, \quad (1)$$

where subscripts denote partial derivatives and Ω is a spatio-temporal smoothing kernel. The intrinsic dimension of f corresponds to the rank of the matrix $\mathbf{J}$ and may be derived from the eigenvalue analysis of $\mathbf{J}$ or from its symmetric invariants defined as:

$$\begin{aligned} H &= 1/3\,\mathrm{trace}(\mathbf{J}) = \lambda_1 + \lambda_2 + \lambda_3 \\ S &= |M_{11}| + |M_{22}| + |M_{33}| = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3\,, \\ K &= \mathrm{determinant}(\mathbf{J}) = \lambda_1\lambda_2\lambda_3 \end{aligned} \quad (2)$$

where λ_i are eigenvalues and $|M_{ij}|$ are minors (i.e. determinants of submatrices) of $\mathbf{J}$. If $K \neq 0$, the intrinsic dimension is 3 (*i3D*); if $S \neq 0$, it is at least *i2D*; and if $H \neq 0$, it is at least *i1D*.

Previously, we have shown that eye movement predictability increases with the intrinsic dimension: the higher the intrinsic dimension the higher the predictive power [32].

To improve noise resilience, we performed our analysis on a lowpass-filtered video (6.6 cycles/degree) that was created by filtering the video with a 5-tap spatial binomial filter and downsampling it (in space) by a factor of two.

To obtain the structure tensor $\mathbf{J}$, partial derivatives were calculated by first smoothing the input with spatio-temporal 3-tap binomial kernels, and then applying $[-1, 0, 1]$ kernels to compute the differences of neighbouring pixel values. The smoothing of the products of derivatives (with Ω) was done with another spatio-temporal 3-tap Gaussian. In principle, pooling these derivatives over a larger spatio-temporal neighbourhood is desirable for a robust computation of $\mathbf{J}$, but for the present analysis, localized responses were more important than robustness against noise.

In addition to being symmetric, the above filter kernels are centred at the detected events, i.e. are non-causal. Note that with a non-causal filter, the output can anticipate the next event. For our purpose, however, a non-causal filter is more appropriate as its output is maximal at the time of the event, whereas the maximum response of a causal filter would be lagging behind the event and, therefore, would decrease the separation between the analytical and empirical saliency measures.

One might argue that for registering the temporal events with eye movements, it would suffice to consider simple temporal differences, but we prefer to register those spatio-temporal events that yield the highest degree of predictability.

Empirical Saliency Measures

Average Scanpaths (Fixation Density Distribution)

We defined our empirical saliency measure as the density of the gaze points averaged over all subjects. These probability maps were computed for each video, by placing two-dimensional spatial Gaussians at each fixation location of all subjects, similarly to the well-known fixation density distribution [34]. The Gaussian kernels had a spatial support of about 4.8 degrees of visual angle and a standard deviation σ of 0.25. The superposition of these Gaussians resulted in the empirical saliency map.

Average Saccades (Saccade Density Distribution)

In the standard approach, all raw gaze samples are used for creation of the empirical saliency map, which includes samples throughout or even to the end of fixations, although ultimately we are interested only in fixation onsets. Therefore, we also created much sparser empirical saliency maps with the above parameters but using only the saccade landing points of all subjects.

Single Saccades

As the traditional empirical saliency map contains gaze data of several viewers, saccadic responses to a certain salient event might arrive at slightly different times within a short time interval. How does this influence our analysis? To gain a deeper understanding of the underlying causes, we also examined the average time lag of individual saccades in responding to changes in the visual scene. For each saccade, we created a sparse response map (similar to the empirical saliency), by placing a single two-dimensional Gaussian at the endpoint of the saccade. Individually, saccade landing points are more prone to noise than the full empirical saliency map. However, they are more localized in space–time and in such a large number of samples (about 40,000 saccades in the first data set of natural outdoor scenes), noise should cancel out.

Normalized Cross-Correlation

Our analysis is based on the cross-correlation of the above described saliency maps shifted, relative to each other, in the time domain [12]. The normalized cross-correlation function (ncc) between two spatio-temporal signals f and g is defined as

$$\mathrm{ncc}(f,g,\tau) = \frac{\sum_{x,y,t}(f(x,y,t)-\bar{f})\cdot(g(x,y,t+\tau)-\bar{g})}{\sqrt{\sum_{x,y,t}(f(x,y,t)-\bar{f})^2\cdot\sum_{x,y,t}(g(x,y,t+\tau)-\bar{g})^2}}$$

where τ is the temporal offset and $\bar{f}$ and $\bar{g}$ stand for the DC components (means) of the two signals. ncc was computed for each analytical and empirical saliency map pair. To determine the correlation expected by chance, as a control condition, we randomly paired analytical and empirical saliency maps of different movies and proceeded as above. This shuffling of scanpaths and videos among each other is a standard procedure in relating low-level image features to gaze data [25, 30].

Results

In our analysis, we shifted the empirical saliency relative to the analytical saliency one frame per temporal unit (approximately 33.367 ms in the first movie set of natural outdoor scenes with static camera and 33.333 ms in the CRCNS eye-1 movie set), within a range of 61 (±30) frames. Here, we are less concerned with the absolute values of the correlation coefficient obtained for the different time lags, but with the value of the time shift at which the maximum correlation occurs. We will mainly restrict ourselves to invariant K, as it proved to be superior to the other invariants in terms of eye movement predictability. However, we shall show that the observations also hold for the other invariants.

Natural Dynamic Scenes with Static Camera

Figure 2 summarizes results obtained for cross-correlating the analytical with the average empirical saliency map of all subjects. Mean correlation coefficients for the 18 movies are plotted against the frame shift in Fig. 2a (red cross curve). A positive lag of t ms indicates that the empirical saliency map follows the invariant movie by t ms. The maximum of the averaged coefficients, for correlating invariant K with the empirical saliency map ("average scanpaths" case), is detected at a lag of 66.73 ms (i.e. two frames).

As can be expected, the maximum is slightly shifted in time to the left when the fixation data are discarded ("average saccades" case—dotted curve in Fig. 2a). In this case, quite surprisingly, highest average correlation is found at −33.36 ms, i.e., on average, the empirical saliency map was ahead of the analytical map by one frame. In both conditions (empirical saliency based on raw gaze samples and on saccade endpoints only), mean correlation curves have a Gaussian-like shape and a pronounced peak,

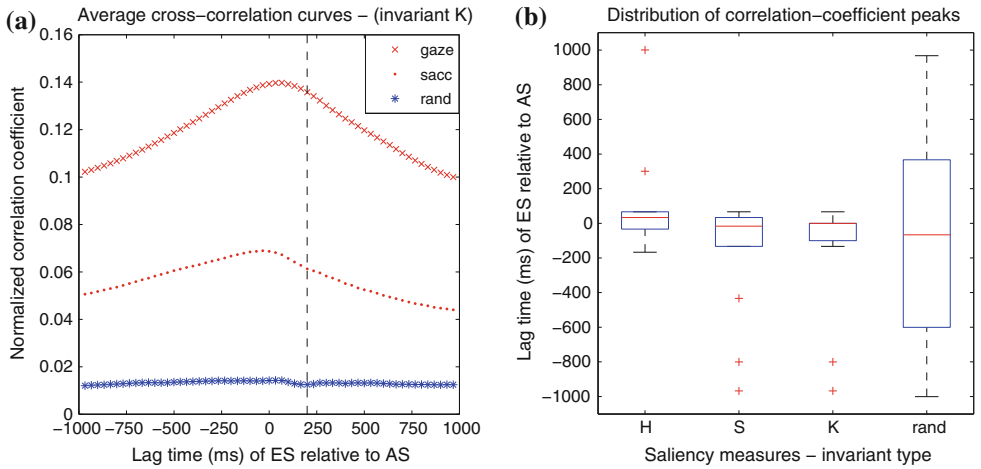

Fig. 2 The empirical saliency map (ES) is offset (with respect to the analytical saliency map—AS) along the time dimension by one frame (33.367 ms) per temporal unit within a predetermined range (±30 frames). A correlation coefficient is calculated for each individual frame shift. **a** Average correlation coefficients over all movies are plotted against the frame shift (*red cross* ES based on average scanpaths, *red dot* ES based on average saccades, *blue asterisk* random AS–ES pairing). The *dashed vertical line* represents the normal mean value of the saccadic reaction time (in the order of 200 ms) to unpredictable targets [2]. **b** Box plot comparing distributions of correlation peaks over the movie set for the AS measures *H*, *S*, *K*, and random AS–ES pairing (*middle line* median, *box* upper and lower quartile, *whiskers* data extent, *plus sign* outliers) (colour figure online)

whereas randomly pairing and then correlating analytical and empirical maps of different videos yields a flat curve (asterisk curve in Fig. 2a).

In the following, we will restrict our considerations to an empirical map based on saccade endpoints only, because, as results suggest, raw gaze data introduce further undesired shifts in the eye movement response.

The box plot in Fig. 2b shows the distribution (over the 18 movies) of time shifts at which maximum correlation was measured. Here, we compare the distributions of correlation peaks obtained for the three invariants, *H*, *S*, and *K*, and for the random analytical–empirical pairing case. As already expected, for the invariants, the peaks are all centred around 0 ms with only few exceptions (red crosses in Fig. 2b). For example, for invariant K, the correlation peaks of two movies are identified at very large negative offsets, meaning that the response in the empirical saliency preceded the signal by an unrealistically large amount of time. An inspection of the shape of the individual correlation curves indicates that these two curves are flatter than those of the other movies, with no pronounced peaks. Indeed, a closer look at the content of these movies reveals that they are of almost still-life character (e.g. unpopulated bridge) and so, as invariant *K* is only sensitive to dynamic content, it is not surprising that the correlation curves have no distinctive peaks. In the following, unrealistically large positive and negative shifts are considered outliers.

Overall, we found that the three distributions of the invariants are very similar with a median of one frame (33.367 ms) for invariant *H*, −16.68 ms for invariant *S*, and 0 ms for *K*. Unlike the concentrated distributions of the invariants, the lags at which maximum correlation occurs in the random pairing case are scattered throughout the correlation window.

Results for correlating invariant *K* with individual saccades are shown in Fig. 3. The maximum of the average correlation curves of all saccades is here, too, identified at −33.367 ms (Fig. 3a). The average correlation curve has again a pronounced peak when compared to the curve of the control condition. However, the peak around 0 ms in the distribution of the time shifts with maximum correlation (in Fig. 3b) is not very distinctive. This is again due to low values of *K* resulting in a flat correlation curve with no pronounced peaks. To measure the curves' "flatness", we used the following simple measure: in Fig. 3c, we sorted

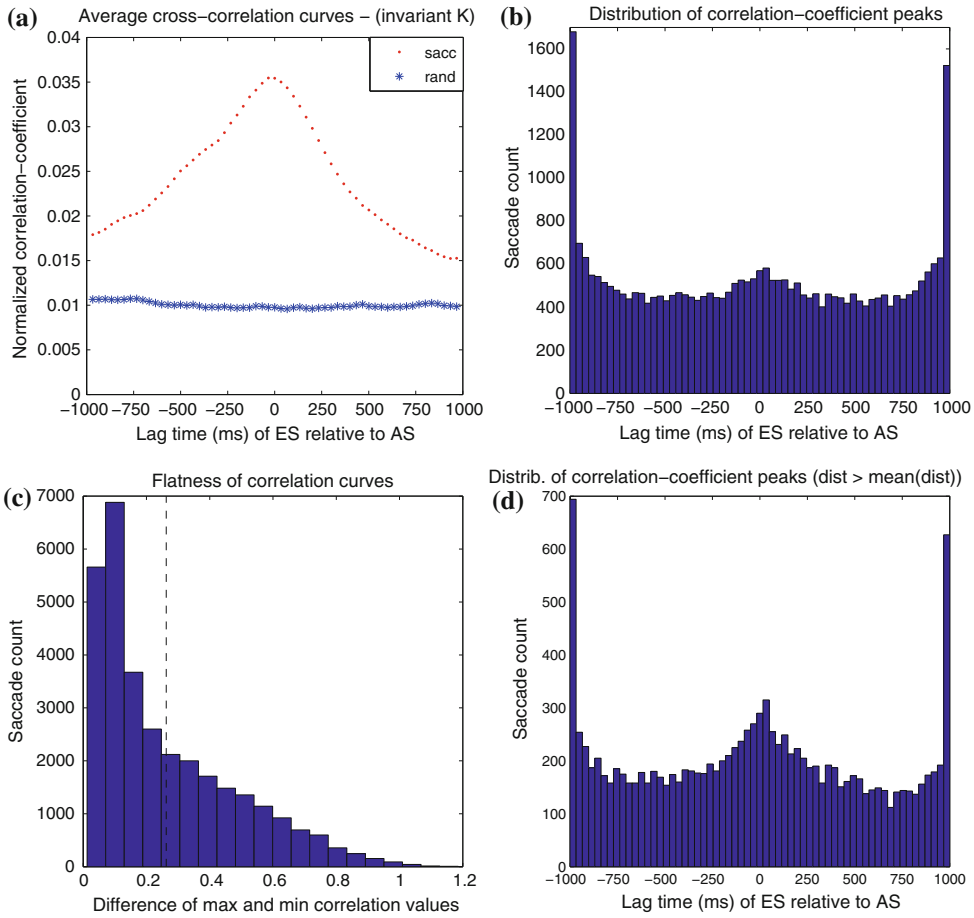

Fig. 3 Individual saccade landing points cross-correlated with the analytical saliency map *K*. **a** Average correlation coefficients over all saccade endpoints (*red dot* landing points, *blue asterisk* shuffled locations). Peak identified at −33.36 ms (−1 frame). **b** Distribution of time shifts (over all saccades) with maximum correlation. **c** Histogram of the distribution of saccades sorted according to the difference between the correlation curves' extreme points. A threshold is set at the mean of the differences removing around 60% of saccades with a flatness measure smaller than the mean measured "flatness". **d** Distribution of correlation peaks of curves after thresholding (colour figure online)

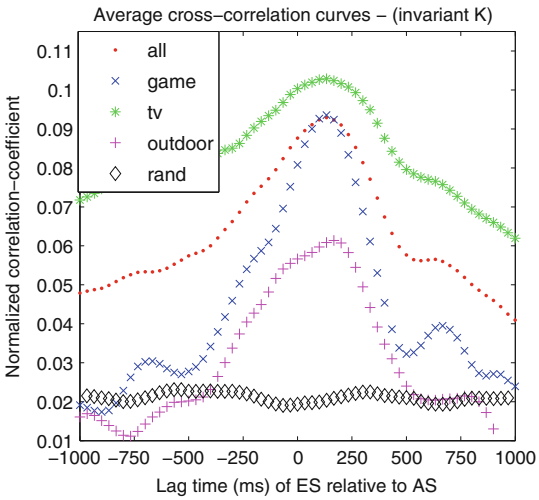

Fig. 4 Mean correlation curves when cross-correlating individual saccades with the invariant K (CRCNS eye-1 data set). For the "original" experiment: averaging over correlation curves of all movies (*red dot*), computer game videos (*blue cross*), TV-clips (*green asterisk*), outdoor scenes (*magenta plus sign*), and randomly shuffled locations (*black diamond*) (colour figure online)

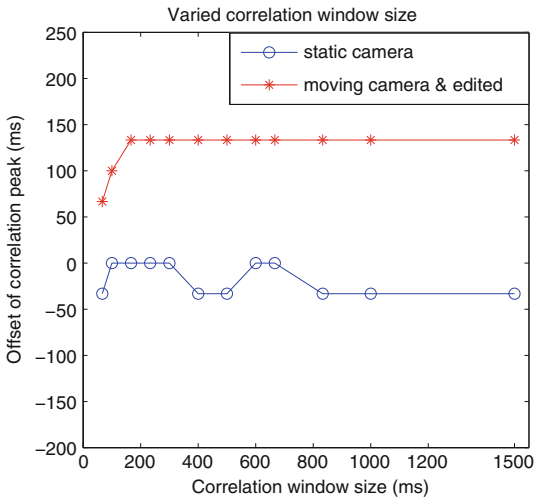

Fig. 5 Offsets of correlation curve peaks when the correlation window size is varied. Individual saccades were cross-correlated with the invariant K. *Blue circle* first movie set of outdoor scenes with static camera, *red asterisk* moving camera and edited videos (colour figure online)

the curves according to the difference of maximum and minimum correlation values over the frame shifts. The more curved the correlation line, the larger this difference. Indeed, when plotting in Fig. 3d only the distribution of saccades for which this difference exceeded the mean of all differences (i.e. 0.26), the peak becomes more prominent. Nevertheless, this simple measure cannot eliminate outliers, such as peaks at implausibly large time offsets.

Moving Camera and Edited Videos

Next, we compare these findings with those obtained on the CRCNS eye-1 data set. When cross-correlating invariant K with individual saccades, a noticeable shift is observed in the location of the peak of the mean correlation coefficients (red dotted curve in Fig. 4). The correlation maximum is here identified at about 133.33 ms (four frames). This larger average time shift could, however, be explained by the fact that a significant number of the clips (television broadcasts and quasi-realistic computer game scenes) are physically quite different from real-world natural scenes. Jump cuts, camera movements, and movie-editing techniques introduce unnatural temporal discontinuities which could entail delayed oculomotor responses. For instance, movie cuts elicit reorienting saccades towards the centre of the screen. To further investigate whether the presence of camera motion and movie-editing techniques affects average response delays, we categorized the fifty movie sequences into three groups based on stimulus type: TV-broadcasts (23 clips), computer games (9 videos), and outdoor scenes (17 sequences; parks, crowds, rooftop bar). Note that the outdoor scenes, too, were captured using basic camera movement techniques (i.e. tilt, pan, and zoom). We excluded from our analysis a synthetic clip of a disc drifting on a textured background. Average cross-correlation curves of the three stimulus groups are plotted in Fig. 4. Although the three curves reach their maximum at very similar time shifts (at about 133.33 ms), notice the difference in how peaked the curves are. The correlation curve of the quasi-realistic computer game stimuli is the most sharply peaked, whereas the curve of the more natural outdoor scenes reaches a plateau at around −66.66 ms after which only limited increase occurs. Considering that we are looking at averages of several individual correlation curves, a pronounced peak and high correlation values (e.g. as in the case of computer games) suggest that the majority of the underlying individual curves reach their maximum at roughly the same time lag. In case of the outdoor scenes, however, the distribution of time shifts at which a peak occurs is more scattered; therefore, averaged coefficient values are lower and the maximum is not very pronounced.

Finally, we show that the size of the sliding window has no impact on the outcome of the correlation analysis. In Fig. 5, we plotted, for various correlation window sizes, the offsets of the peaks (of mean correlation curves) for the two movie sets (blue circle—dynamic scenes with static camera, red asterisk—moving camera and edited videos). When the window is smaller than the actual optimal offset, the peak is detected at the border of the correlation window, otherwise the curves are almost flat, i.e. offsets are consistent, with only small fluctuations of one frame in case of the dynamic scenes with static camera. Here, peaks were detected at an offset of either −1 or 0 frames.

Discussion

An often neglected question in the design of computational models of saliency is what the typical response lag is to changes in the visual scene. The choice of a specific value is typically motivated by laboratory investigations of saccadic response latencies to synthetic stimuli. In [4] for instance, authors manually choose a particular latency that agrees with the timing of human saccades in the context of a synthetic test clip. However, depending on the stimulus type, the average lag can vary quite substantially: in [18], authors distinguished between "reactive saccades of the laboratory" (having positive lags) and "proactive saccades of normal life" (with near-zero or even negative lags). In this study, we aimed to infer the mean response delay, in laboratory settings under head-fixed viewing conditions, when free-viewing dynamic natural scenes. Using cross-correlation analysis of analytical saliency maps (encoding saccade-triggering changes in the video) and spatio-temporal fixation maps (encoding eye movement responses to the salient events), we identified the time shift at which the two maps have the maximum correlation. We then averaged results over several movies or individual saccades to determine the mean lag in the stimulus class of natural videos. In addition, we examined whether this average response delay differs from that obtained on similar natural and quasi-natural (video game) stimuli, which were captured using basic camera movement techniques and, depending on the movie type, postprocessed with video editing software.

In the first data set of dynamic natural scenes, we found a near-zero mean lag, meaning that, on average, reactions to salient events coincided with or even slightly preceded the events themselves. This result was consistent for all analytical saliency maps (invariants H, S, and K) and both when scanpaths of all subjects and individual saccades were cross-correlated with analytical maps. This somewhat surprising finding may be attributable to an adaptation of the human visual system to the environmental dynamics of the surrounding world. Most dynamic events in natural scenes are, at least to some extent, predictable. Such anticipatory mechanisms (e.g. looking ahead of the movement) imply some sort of scene knowledge of the dynamic characteristics of the environment that is due, for instance, to experience with the physical laws of motion.

In line with the studies on task-specific gaze control [9, 13], one could also speculate that, during the viewing of a particular scene, observers might identify certain higher-level (hidden) tasks and actions, such as playing beach ball, walking on a bridge, driving in a roundabout. If we think of the free viewing of natural videos as action observation of what is happening in the video, possessed knowledge about these actions could possibly generate anticipatory gaze behaviour.

Note that we are here not aiming at explaining gaze behaviour with a simple bottom-up model but merely at measuring the time lag between events in the video and the responding eye movements. We use a plausible model of bottom-up saliency simply to improve the measurement of this time lag. In other words, our bottom-up saliency model based on the invariants merely serves as an "event detector". The fact that this time lag is small can indeed be attributed to top–down mechanisms but our result does not depend on such interpretations. The anticipation that we find can be due to many different predictive mechanisms starting from very simple (low-level) models, such as a Kalman filter, to more complex (high-level) ones, such as action planning. Given this possible continuum of mechanisms of increasing complexity, it seems unnecessary to draw a "bottom-up top-down borderline".

The analysis of the second set of complex stimuli (CRCNS eye-1 data set) reveals a longer average delay of about 133 ms between a dynamic event in the scene and saccades responding to it. We argue that, due to the presence of jump cuts, camera motion, and other movie editing techniques, the amount of bottom-up influence in these stimuli is, on average, higher than in truly natural scenes. The introduced temporal discontinuities and the sudden appearance of text overlays in television broadcasts trigger a high number of reactive saccades. Similarly, to passive observers, the moves of the video game character are less predictable than to the game player himself. Looking at the average correlation curves of the three video subsets (TV-clips, games, and outdoor scenes), the curve of the outdoor scenes pops out. Its global maximum is identified shortly after the overall average of 133 ms but mean correlation values are comparably high already beginning with -66.66 ms. This could suggest that, in comparison with the first set of natural outdoor movies in which the great majority of saccades were rather predictive (therefore, the peak shortly before zero), here the ratio of visually guided and anticipatory saccades is more balanced.

Computational models of attention either assume no time shift between their analytical saliency maps and the responding eye movements, or they do not try to optimize this value but use subjective observations [4]. We argue that by introducing an artificial time lag adjusted to the stimulus type (i.e. eliciting maximum response in the analytical saliency at the time of the expected gaze response, not at the time of the event), saliency models could significantly increase their performance in predicting eye movements. As an alternative, temporal uncertainty could be introduced in the model in order to account for the different stimulus-specific time lags [32].

Our primary motivation to investigate this problem stems from our work on integrating gaze into future visual and communication systems by measuring and guiding eye movements[2] [1]. To perform gaze guidance, we derive transformations that alter the saliency distribution of the scene in real-time. In such a scenario, the right timing of the so-called gaze-capturing events is critical for achieving the desired effect, i.e. for attention to be drawn to a specific movie region at a specific time, the temporal placement of the gaze-capturing event must take into consideration the stimulus-specific average response lag.

In summary, in this study, we have characterized a special class of visual stimuli, namely, that of real-world natural scenes, in terms of the typical time lags between salient changes in the scene and the responding eye movements. To measure this typical time lag, we temporally aligned analytical spatio-temporal saliency maps with response maps encoding saccadic reaction to the salient events. We argue that the near-zero average lag could be attributable to an adaptation of the human visual system to the (often predictable) dynamics of the environment. We show that the degree of anticipation is reduced when moving away from natural scenes by introducing cuts, camera motion, and film-editing. Finally, we suggest that the stimulus dependent mean response lag should be an important consideration in the design of computational models of visual saliency and provide a method for computing the average time shift between movie events and eye movements.

Acknowledgements We would like to thank Karl Gegenfurtner: data were collected in his lab at the Dept. of Psychology of Giessen University. Our research has received funding from the European Commission within the project GazeCom (contract no. IST-C-033816, http://www.gazecom.eu) of the 6th Framework Programme. All views expressed herein are those of the authors alone; the European Community is not liable for any use made of the information.

References

1. Barth E, Dorr M, Böhme M, Gegenfurtner KR, Martinetz T. Guiding the mind's eye: improving communication and vision by external control of the scanpath. In: Rogowitz BE, Pappas TN, Daly SJ, editors. Human vision and electronic imaging, vol 6057 of Proceedings of SPIE. Invited contribution for a special session on Eye Movements, Visual Search, and Attention: a Tribute to Larry Stark; 2006.
2. Becker W. Saccades. In: Carpenter RHS, editor. Vision & visual dysfunction, vol 8: Eye movements. London: CRC Press; 1991. p. 95–137.
3. Böhme M, Dorr M, Krause C, Martinetz T, Barth E. Eye movement predictions on natural videos. Neurocomputing. 2006;69(16–18):1996–2004.
4. Carmi R, Itti L. Visual causes versus correlates of attentional selection in dynamic scenes. Vis Res. 2006;46:4333–45.
5. Carpenter RHS. Oculomotor procrastination. In: Fisher DF, Monty RA, Senders JW, editors. Eye movements: cognition and visual perception. Hillsdale, NJ: Lawrence Erlbaum; 1981. p. 237–46.
6. Chajka K, Hayhoe M, Sullivan B, Pelz J, Mennie N, Droll J. Predictive eye movements in squash. J Vis. 2006;6(6):481–6.
7. Crundall D, Chapman P, Phelps N, Underwood G. Eye movements and hazard perception in police pursuit and emergency response driving. J Exp Psychol. 2003;9(3):163–74.
8. Findlay JM. Spatial and temporal factors in the predictive generation of saccadic eye movements. Vis Res. 1981;21(3):347–54.
9. Flanagan JR, Johansson RS. Action plans used in action observation. Nature. 2003;424(6950):769–71.
10. Gesierich B, Bruzzo A, Ottoboni G, Finos L. Human gaze behaviour during action execution and observation. Acta Psychol. 2008;128(2):324–30.
11. Goldstein RB, Woods RL, Peli E. Where people look when watching movies: do all viewers look at the same place?. Comput Biol Med. 2007;3(7):957–64.
12. Gonzalez RC, Woods RE. Digital image processing, 2nd edn. Boston, MA: Addison-Wesley Longman Publishing Co., Inc; 2001.
13. Hayhoe M, Ballard D. Eye movements in natural behavior. Trends Cogn Sci. 2005;9(4):188–94.
14. Itti L. Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Trans Image Process. 2004;13(10):1304–18.
15. Itti L. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cogn. 2005;12(6):1093–123.
16. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell. 1998;20(11):1254–9.
17. Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol. 1985;4(4):219–27.
18. Land MF, Furneaux S. The knowledge base of the oculomotor system. Philos Trans R Soc B Biol Sci. 1997;352:1231–9.
19. Land MF, Hayhoe M. In what ways do eye movements contribute to everyday activities?. Vis Res. 2001;41:3559–65.
20. Land MF, Lee DN. Where we look when we steer. Nature. 1994;369:742–4.
21. Land MF, McLeod P. From eye movements to actions: how batsmen hit the ball. Nat Neurosci. 2000;3:1340–5.
22. Land MF, Mennie N, Rusted J. The roles of vision and eye movements in the control of activities of daily living. Perception. 1999;28:1311–28.
23. Mennie N, Hayhoe M, Sullivan B. Look-ahead fixations: anticipatory eye movements in natural tasks. Exp Brain Res. 2007;179(3):427–42.
24. Pelz JB, Canosa R. Oculomotor behavior and perceptual strategies in complex tasks. Vis Res. 2001;41:3587–96.
25. Reinagel P, Zador AM. Natural scene statistics at the centre of gaze. Network Comput Neural Syst. 1999;10:341–50.
26. Russo FD, Pitzalis S, Spinelli D. Fixation stability and saccadic latency in élite shooters. Vis Res. 2003;43(17):1837–45.
27. Smit AC, Van Gisbergen JAM. A short-latency transition in saccade dynamics during square-wave tracking and its significance for the differentiation of visually-guided and predictive saccades. Exp Brain Res. 1989;76:64–74.
28. Smith TJ, Henderson JM. Edit blindness: the relationship between attention and global change blindness in dynamic scenes. J Eye Movement Res. 2008;2(2):1–17.

[2] For further information visit http://www.gazecom.eu.

29. 't Hart BM, Vockeroth J, Schumann F, Bartl K, Schneider E, König P, Einhäuser W. Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions. Visual Cogn. 2009;17(6/7):1132–58.
30. Tatler BW, Baddeley RJ, Gilchrist ID. Visual correlates of fixation selection: effects of scale and time. Vis Res. 2005;45: 643–59.
31. Underwood G, Phelps N, Wright C, van Loon E, Galpin A. Eye fixation scanpaths of younger and older drivers in a hazard perception task. Ophthal Physiol Opt. 2005;25:346–56.
32. Vig E, Dorr M, Barth E. Efficient visual coding and the predictability of eye movements on natural movies. Spat Vis 2009;22(5):397–408.
33. Vig E, Dorr M, Martinetz T, Barth E. A learned saliency predictor for dynamic natural scenes. In: Diamantaras K, Duch W, Iliadis LS, editors. ICANN 2010, Part III, LNCS 6354, Berlin: Springer; 2010. p. 52–61.
34. Wooding DS. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. Behav Res Methods Instruments Comput. 2002;34(4):518–28.