Real and implied motion at the center of gaze

Alper Açık University of Osnabrück, Institute of Cognitive Science, Osnabrück, Germany

Andreas Bartel University of Osnabrück, Institute of Cognitive Science, Osnabrück, Germany

Peter König University of Osnabrück, Institute of Cognitive Science, Osnabrück, Germany

Even though the dynamicity of our environment is a given, much of what we know on fixation selection comes from studies of static scene viewing. We performed a direct comparison of fixation selection on static and dynamic visual stimuli and investigated how far identical mechanisms drive these. We recorded eye movements while participants viewed movie clips of natural scenery and static frames taken from the same movies. Both were presented in the same high spatial resolution (1080 × 1920 pixels). The static condition allowed us to check whether local movement features computed from movies are salient even when presented as single frames. We observed that during the first second of viewing, movement and static features are equally salient in both conditions. Furthermore, predictability of fixations based on movement features decreased faster when viewing static frames as compared with viewing movie clips. Yet even during the later portion of static-frame viewing, the predictive value of movement features was still high above chance. Moreover, we demonstrated that, whereas the sets of movement and static features were statistically dependent within these sets, respectively, no dependence was observed between the two sets. Based on these results, we argue that implied motion is predictive of fixation similarly to real movement and that the onset of motion in natural stimuli is more salient than ongoing movement is. The present results allow us to address to what extent and when static image viewing is similar to the perception of a dynamic environment.

Introduction

In order to sample the visual world, humans change their fixation points multiple times per second. Fixations allow a high-resolution (Tootell, Silverman, Switkes, & De Valois, 1982) analysis of the selected scene region until the next saccade moves the eyes to another location of interest. In line with the trend of studying perception under natural conditions with stimuli taken from the real world (Felsen & Dan, 2005), the current standard methodology in the field consists of recording eye movements of subjects who are viewing natural or equally complex images and then looking for properties that characterize fixated regions (for recent reviews, see Schütz, Braun, & Gegenfurtner, 2011; Tatler, 2009; Tatler, Hayhoe, Land, & Ballard, 2011; Wilming, Betz, Kietzmann, & König, 2011). Due to its obvious importance for survival in our dynamic world, understanding how a system coupled with a natural environment (Thompson & Varela, 2001) achieves visual sampling is of great interest.

Unfortunately, even though the dynamicity of the environment is an integral feature of "natural" perception (Gibson, 1979), and studies using artificial stimuli reveal that motion strongly attracts attention (Yantis & Egeth, 1999; Yantis & Jonides, 1990), the list of eye-movement studies that employ dynamic stimuli is relatively short. Some of these studies rely on presenting natural videos to participants in the laboratory (Böhme, Dorr, Krause, Martinetz, & Bartz, 2006; Carmi & Itti, 2006; Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Itti & Baldi, 2009; Le Meur, Le Callet, & Barba, 2007; Marat et al., 2009; Mital et al., 2011; Vig, Dorr, Martinetz, & Barth, 2012); others record people's eye movements while they perform common tasks in a real or virtual environment (Hayhoe & Ballard, 2005; Land & Hayhoe, 2001; Land & McLeod, 2000; Rothkopf, Ballard, & Hayhoe, 2007; Schumann et al., 2008), and recently, an upsurge has occurred in hybrid studies that introduce a mixture of these two approaches (Cristino & Baddeley, 2009; Einhäuser et al., 2009; Foulsham, Walker, & Kingston, 2011; 't Hart et al., 2009). Only a few of these studies

Citation: Açık, A., Bartel, A., & König, P. (2014). Real and implied motion at the center of gaze. *Journal of Vision*, *14*(1):2, 1–19, http://www.journalofvision.org/content/14/1/2, doi:10.1167/14.1.2.

doi: 10.1167/14.1.2 Received January 29, 2013; published January 6, 2014 ISSN 1534-7362 © 2014 ARVO

Downloaded from jov.arvojournals.org on 03/15/2021

(Dorr et al., 2010; Machner et al., 2012; 't Hart et al., 2009) directly compare eye movements recorded under comparable static and dynamic stimulus conditions. In order to check whether findings obtained using static stimuli can be generalized to dynamic scenes and vice versa studies that carefully control for stimulus motion are needed.

One major goal of eye-movement research is to find local statistical properties, usually labeled image features, that distinguish fixated natural scene regions from ones that are not fixated as well as possible (Einhäuser & König, 2003; Kienzle, Franz, Schölkopf, & Wichmann, 2009; Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; D. Parkhurst, Law, & Niebur, 2002; Reinagel & Zador, 1999). Such features can be pooled in order to generate a saliency map (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998): A computational model receives an image as input and returns a two-dimensional map, on which each location is assigned a fixation probability proportional to the weighted sum of the feature values at that location. In the case of dynamic stimuli, the saliency map is computed for each movie frame separately, and the features of interest include temporal variations (Itti & Baldi, 2009; Le Meur et al., 2007; Marat et al., 2009; Vig, Dorr, & Barth, 2009; Vig et al., 2012). Even though saliency (D. Parkhurst et al., 2002; Peters, Iyer, Itti, & Koch, 2005) and, especially, temporal saliency (Itti & Baldi, 2009; Vig et al., 2012) are successful in predicting fixation locations on natural stimuli, whether the human attention system employs a similar mechanism is open to debate (Einhäuser, Rutishauser, & Koch, 2008; Rothkopf et al., 2007).

Due to the complexity of natural stimuli, practically an infinite number of spatial, temporal, and spatiotemporal features can be computed in order to estimate a region's saliency. For a multifeature saliency model, efficiency and success depend on two properties of the feature bank: First, each feature has to correlate well with fixation probability so that fixated regions correspond with locations at which the feature of interest assumes a relatively high value. For instance, features derived from intrinsic dimensionality analysis are better predictors of fixated regions than luminance contrast (LC) is (Açık, Sarwary, Schultze-Kraft, Onat, & König, 2010; Saal, Nortmann, Krüger, & König, 2006; Vig et al., 2012). Second, the features are to display low statistical dependency among them; otherwise, their coexistence in the model would be redundant because they would address the same aspects of saliency. Baddeley and Tatler (2006) have shown that if the correlation between high-frequency edges and contrast is controlled for, the former appears as a good predictor of fixations and the latter not. In information theoretical classification terminology (Peng, Long, & Ding, 2005), the first property corresponds to maximal relevance and can be realized as high mutual information between the target class (fixated vs. not fixated regions) and the values of a given feature. The second criterion can be phrased as a minimum redundancy constraint (Peng et al., 2005), in which features that display low mutual information with one another are preferred. That is, a good set of saliency features has to classify fixations as well as possible while entertaining low statistical dependency among features in order to address different aspects of saliency.

As an example of the phrase "correlation does not imply causation," higher feature values at fixated regions do not necessarily mean that these features cause saccades to land at those locations. Previous research (Açık, Onat, Schumann, Einhäuser, & König, 2009; Einhäuser & König, 2003) has shown that both increases and decreases in local LC increase fixation probability. This finding contradicts the basic assumption of typical saliency map models, which model the influence of contrast on fixation probability as a monotonically increasing function (e.g., Itti et al., 1998, but cf. D. J. Parkhurst & Niebur, 2004). Cristino and Baddeley (2009) presented either intact or spatiotemporally band-pass filtered videos that were recorded with a head-mounted camera during a casual walk on a shopping street. The band-pass filtering for space and time was such that high-pass temporally filtered videos contained only low-frequency spatial information and vice versa. Even though the filtered videos differed greatly in terms of the spatial distribution of low-level features, such as rich information at the edge of the movie if the stimuli are high-temporal filtered but more content at the center if low-frequency temporal information is present, the fixation distributions remained very similar across stimulus conditions (Cristino & Baddeley, 2009). Frey and colleagues (2011) went a step further and investigated the relationship between a feature and fixation probability after removing all of the information in that feature channel. They presented images either intact in their color content or the same images with the red-green or blue-yellow channel removed. Despite this selective removal of color information, the fixated locations were still characterized by higher feature values of, especially, red-green contrast, which was computed on the color-intact images. Thus, paradigms that manipulate or completely remove a certain feature channel suggest that the relationship between features and fixation probability is not always causal.

Motion perception does not have to rely on explicit motion information—that is, real movement—but can be deduced from other cues even if the stimulus is static (Freyd, 1987; Hubbard, 1995). Freyd and Finke (1984) showed a static rectangular target with gradual changes in its orientation in order to induce implied motion, and they asked the participants about the final

Downloaded from jov.arvojournals.org on 03/15/2021

appearance of the target. The participants remembered the stimulus to be more tilted in the direction of implied motion than in its actual orientation. Thus, it was as if the participants had extrapolated the implied motion until they gave their responses (for reviews, see Freyd, 1987; Hubbard, 1995). Freyd (1983) presented photographs taken during irreversible movements, such as someone jumping off of a wall, and probed the participants' recognition memory with either the same stimulus or another snapshot that took place earlier or later in the same action sequence. Participants took longer to answer and made slightly more mistakes when the probe was a later snapshot from the same sequence, suggesting that the memory representation of the action included its natural continuation (Freyd, 1983, 1987). Both real and implied motion led to increases in perceptual estimates of temporal duration (Kanai, Paffen, Hogendoorn, & Verstraten, 2006; Yamamoto & Miura, 2012). Kourtzi and Kanwisher (2000) recorded fMRI data while participants viewed static images of people performing natural movements or the end-states of those movements. They observed that motion areas middle temporal and medial superior temporal were more active in the former condition, suggesting a neural substrate for implied motion effects. Later studies (Krekelberg, Dannenberg, Hoffmann, Bremmer, & Ross, 2003; Proverbio, Riva, & Zani, 2009) replicated this result with different types of implied motion. These findings clearly show that even if the visual system is presented a completely static scene, motion cues can still be extracted and influence behavioral and neural responses.

Eye-movement data gathered during viewing of a movie offer two types of analysis. First, one can apply the steps taken in the analysis of picture-viewing data. For instance, it is known that during scene viewing, feature saliencies are higher at the targets of shorter saccades compared to longer saccades (Açık et al., 2010; Tatler, Baddeley, & Vincent, 2006). It would be important to know whether this saccade amplitude and feature saliency relationship remains the same if the stimulus is dynamic. Similarly, it is still debated whether feature saliencies decrease with viewing time (Açık et al., 2009; D. Parkhurst et al., 2002; Tatler, Baddeley, & Gilchrist, 2005). The data gathered by Tatler and colleagues (2005) revealed no such temporal saliency decrease. In a study comparing different image categories (Açık et al., 2009), we observed the decrease only in the case of landscape images devoid of any man-made objects. Accordingly, for movie studies revealing a similar decrease in dynamic feature saliency over time (Carmi & Itti, 2006; Marat et al., 2009), one needs to ask whether this is because of the dynamic nature of the stimulus or due to the semantic aspects of selected videos. This question can be answered by using pictures and videos depicting the very same scene. Thus, saccade metric–related and temporal variations in feature saliencies can be addressed in static and dynamic stimuli alike. The second type of analysis, however, relies on the temporally varying nature of the stimulus and hence can be performed only with movies. Given a fixation at a certain location, one can analyze local stimulus properties before and after the onset of the fixation in order to address whether saccades are predictive or reactive (Land & Hayhoe, 2001; Vig et al., 2009, 2011). Thus, for a comprehensive understanding of eye-movement data obtained with movies, analysis techniques used with static images are to be used in conjunction with methods that rely on the temporal variation in dynamic stimuli.

Imagine you are reading the newspaper or visiting a web page on which you encounter a picture taken during a motor sports race, such as Formula 1 or NASCAR. It is quite plausible that one of the first things you inspect is a car on the racetrack. Obviously, even though no motion is present in the photograph, the object of your interest, the racing car, was moving at the time the photo was taken. This thought experiment reveals the main research question addressed in the current study: While viewing static images, do humans preferentially fixate on regions that were moving or have the potential to move? In order to seek an answer, we showed human observers short movies as well as frames taken from these movies. Selecting the static stimuli from the movie frames was crucial because this enabled us to quantify to what degree motion features, which can be computed only with the movies, correlated with fixation probability in the absence of stimulus motion. Moreover, we quantified the time course of such saliency effects on fixation selection. Finally, we calculated statistical dependencies within static features, within motion features, and between static and motion features in order to reveal the least redundant feature pairs. Our results allow us to characterize the extent to which static-image viewing is comparable to the perception of a dynamic stimuli.

Methods

Participants

A total of 23 university students (11 females, age range 21–30, median age 25 years) participated in the study. All had normal or corrected-to-normal visual acuity and were naïve to the purpose of the experiment. For their participation, they either received monetary compensation (5€) or were granted extra university course credits. The study was conducted in compliance with the Declaration of Helsinki as well as with

Downloaded from jov.arvojournals.org on 03/15/2021

national and institutional guidelines for experiments with human subjects.

Stimuli

The stimuli (Figure 1) consisted of short movie clips ("movie" condition) and frames taken from these clips ("frame" condition). The stimulus set was assembled from two commercial DVDs, featuring the documentaries *Highway One* and *Belize* from the *Colourful Planet* collection (Telepool Media GmbH, Leipzig, Germany, courtesy of www.mdr.de). Both provide the content at a frame rate of 25 frames per seconds (fps) and in the WMV HD-DVD (Microsoft Media Video High Definition) format at the HD resolution of 1080 × 1920 pixels. We consider this high resolution to be an important property as static and dynamic visual stimuli are presented at identical quality. The DVDs include scenes with various content, such as natural landscapes, wildlife, close-ups of vegetation and animals, man-made objects, cars and traffic scenes, humans and close-ups of talking faces, open waters, and even fire—that is, a large variety of spatiotemporal events. The selection of the movie clips consisting of a single continuous shot was based on subjective criteria. These criteria corresponded with the presence of at least some object motion in the given scene, an absence of camera movement in order to minimize egomotion-like perception, avoidance of compression noise, and semantic unrelatedness of the clips in the final set. Accordingly, 216 movie clips with durations ranging from 0.8 to 15.4 s (mean duration 4.0 s; median duration 3.8 s) were selected for the "movie" condition. The lossless HuffYUV compression (Huffman, 1952) was applied in order to optimize the file sizes. The middle frames of the movie clips served for the "frame" condition, in which a static frame was shown for the same duration as the underlying video. In summary, 216 short movie clips featuring different kinds of object motion and one frame taken from each of these clips established the stimuli used in the study.

Apparatus

The experiment was programmed in Python and run on an Apple Mac Pro (Apple, Cupertino, CA) operated with Linux. For the extraction of the frames from the DVDs, the lossless compression of movie clips, all further video editing, and the presentation of the stimuli, MPlayer and MEncoder (www.mplayerhq.hu) were used. The stimuli were shown on a 30-in. Apple Cinema HD Display (Apple) at a resolution of 1600 × 2560 pixels and with a refresh rate of 60 Hz. The latter setting, coupled with the movie stimuli at 25 fps, resulted in the repetition of every fifth movie frame for one additional refresh of the monitor, which MPlayer controlled. This correction ensured that no progressive increase in the temporal synchronization error of the eye tracker and presentation tools existed.

Eye-position data were recorded using the Eyelink II system (SR Research, Mississauge, Ontario, Canada) with a sampling rate of 500 Hz. Events, such as saccades, fixations, and blinks, were automatically detected and parameterized via the eye-tracker system. Because the fixation-detection algorithm of the eye tracker doesn't define smooth-pursuit movements explicitly, there is the risk of detecting spurious saccades while eyes follow slowly moving objects in our movies. Accordingly, we have chosen the higher velocity threshold of the algorithm (30°/s) for saccades, decreasing the probability of saccade onset detections during smooth pursuit. This might mean that we have treated some smooth-pursuit movements as fixations, but in the Discussion, we will return to this issue and argue that this is not important for the analysis performed here.

Design and procedure

During the experiment, each participant viewed 432 stimuli, consisting of 216 movie clips and 216 static images that were the middle frames of the movie clips. The presentation order of the stimuli was pseudorandomized for each subject. For that reason, half of the movie clips were randomly selected and assigned to the first half of the experiment, and their frame counterparts were assigned to the second half. The remaining frames were then added to the first half and the remaining movie clips to the second half. The presentation order of the stimuli in each half was completely random: This ensured that, in each half of the experiment, equal numbers of frames and clips were shown and that a movie clip and a frame taken from it were never shown in the same half.

Each participant, after receiving the instructions about the study, was brought to the darkened experiment room. The task was free viewing, and the participant was told solely to "study the movie clips and images carefully." The distance between the participant's eyes and the screen was 80 cm, and at this distance, the stimuli covered 27° × 43° of the participant's visual field. Given this relatively high stimulus width and because the eye tracker is able to compensate for head movements with its third camera, no chin rest was used. After its installation on the participant's head, the eye tracker was calibrated. When the calibration error of a single eye was 0.33° of visual angle or less, that eye was selected for tracking, and the experiment began with the appearance of a

Downloaded from jov.arvojournals.org on 03/15/2021

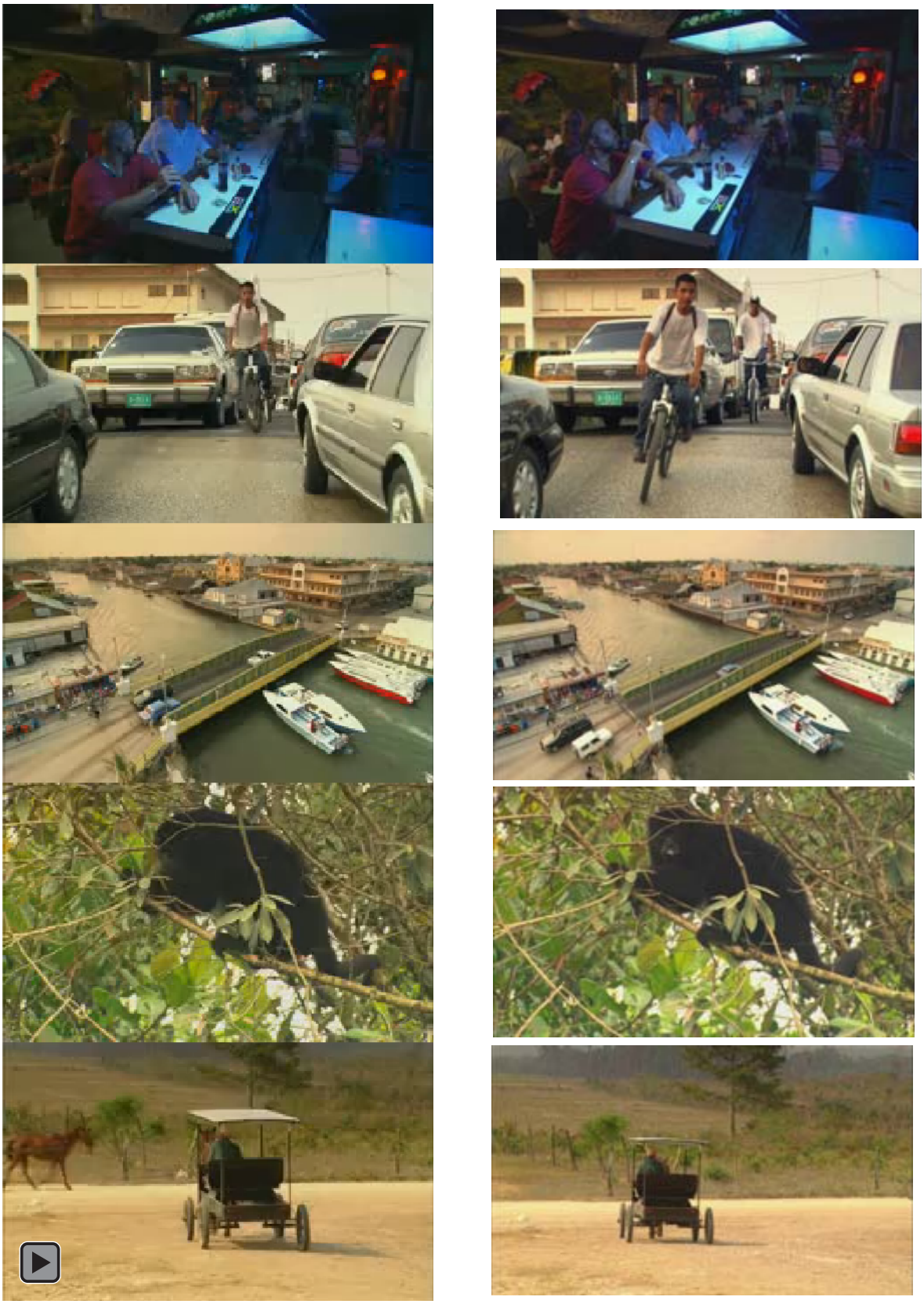

Figure 1. Stimuli. Representative examples of the stimuli used in the movie (left) and frame conditions (right) of the experiment. The frame stimuli correspond to the middle frames of the movie stimuli. Clips are taken from *Highway One* and *Belize* of the *Colourful Planet* DVD collection (Telepool Media GmbH, Leipzig, Germany, courtesy of www.mdr.de).

Downloaded from jov.arvojournals.org on 03/15/2021

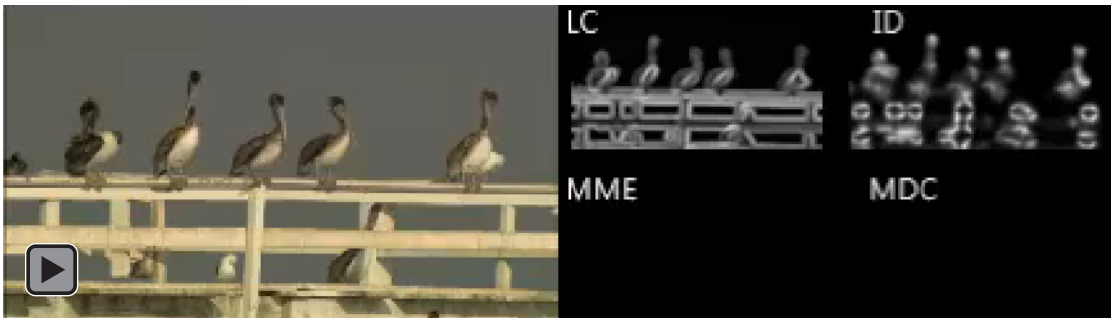

Figure 2. Fixations and features. Large movie: A movie stimulus with actual (green) and control (red) fixations overlaid. The fixations remain on screen for their recorded duration. The control fixations come from the presentation of the rest of the movies with the same temporal constraints. Smaller movies: The feature maps computed from the stimulus shown. The very first frame in the movement feature videos is blank because they are computed from the difference between two successive frames. In frame stimuli, the control fixations for a given stimulus are taken from the presentation of the rest of the frame stimuli, and the fixation onset and offset did not play a role because the stimulus does not vary temporally. The clip is taken from *Highway One* of the *Colourful Planet* DVD collection (Telepool Media GmbH, Leipzig, Germany, courtesy of www.mdr.de).

central fixation cross, followed by the presentation of stimuli. After every third stimulus, the fixation cross appeared again for the purpose of drift correction. Upon the completion of 108 trials, a 5-min break was given each time, and the subject was free to remove the eye tracker. After the break, the calibration was performed anew, and the experiment continued. Together with calibrations and breaks, the entire experiment lasted for about an hour.

Data analysis

The data analysis addressed the relationship between local features and fixation probability, the alteration of this relationship with certain viewing parameters, and the statistical dependencies among different local features.

Features

We used two "static" features that were computed from local patches of single frames (Figure 2, upper part). The first static feature is LC, which is the standard deviation of local luminance, and it was computed in circular regions with a diameter of 1°. The so-called "i2D" feature (abbreviated as ID here) of the intrinsic dimensionality analysis (Saal et al., 2006; Zetzsche, Barth, & Wegmann, 1993) quantifies the amount of junction and corner-like structures that are present in a local region. ID is known to be a better predictor of fixation locations when compared with LC (Açık et al., 2010). Nevertheless, LC is a physiologically plausible feature and an invariable constituent of saliency map models (Itti & Koch, 2000). Thus, a standard feature, LC, and a better fixation predictor, ID, were selected as the static features in the present study.

We have introduced two "movement" features, which are calculated from local patches of movement-related differences between two successive frames (Figure 2, lower part). Feature extraction from movement assumes that the movement of objects depicted in a scene can be reliably detected. Detection and quantification of motion from a two-dimensional frame series is a major challenge in computer vision that is commonly labeled "optical flow estimation" (Horn & Schunck, 1981) even though that term is occasionally used exclusively for egomotion (Lappe, 2000). Especially in the case of natural movies employed here, optic-flow computations are difficult because temporal changes in luminance that are used to extract movement vectors are not always due to motion but can stem from other factors, such as shading, occlusion, and visual noise (Anandan, 1989; Black & Anandan, 1996). Several algorithms are developed for optic flow estimations, and it is beyond the scope of the present study to evaluate and compare their performance (see Baker et al., 2011 and the corresponding webpage http://vision.middlebury.edu/flow/ for a comprehensive review and comparison of state-of-the-art optical flow algorithms). We have chosen the algorithm that Black and colleagues (Black & Anandan, 1996; Black & Jepson, 1996) introduced because it is freely available (http://www.cs.brown.edu/~black/), is now a classic approach in the field, and suits the present purpose. The advantages of the algorithm include the ability to detect multiple motions in a small image region, robustness in the case of brightness changes without movement, and very reliable detection of spatially abrupt motion boundaries due to object movement (Black & Anandan, 1996). The default parameters of the algorithm were kept constant, but the maximum number of pyramids was set to seven in

Downloaded from jov.arvojournals.org on 03/15/2021

order to address the relatively large size of the stimuli as Michael J. Black (personal communication, May, 2007) suggested. For a given successive frame pair, the algorithm computes, for each pixel, the horizontal and vertical components of motion vectors in pixel units and returns these in two frame-size matrices. Using these two matrices, we have calculated the two movement features that we call mean motion energy (MME) and motion-directed contrast (MDC). For a given location, the MME quantifies the amount of movement in a circular region with a diameter of 1° and is defined as

$$\frac{1}{N}\sum_{k=1}^{N}\sqrt{h_k^2+v_k^2}, \quad (1)$$

where k runs over the pixels in the circular patch, and h and v are the lengths of the horizontal and vertical motion components, respectively. As can be seen, this is simply the arithmetic mean of the motion vector amplitudes inside the region and ignores the direction of the vectors. This is computed for each pixel to obtain a frame-sized MME map. MDC, on the other hand, corresponds to the variation of motion in a region around a given pixel and is defined as

$$\sqrt{Var(H)+Var(V)}, \quad (2)$$

that is, the square root of the summation of the individual variances of the horizontal (H) and vertical (V) motion components inside the region. We call the feature "directed" because the horizontal and vertical motion component variances are separately computed. For the frame stimuli, the motion feature maps are obtained using the presented frame and the one preceding it. Accordingly, whereas the MME feature is high for all moving regions in the image, the MDC feature is high only at motion transitions: the regions in which static or slowly moving parts are found together with faster-moving parts or in which the movement direction of neighboring regions differs.

Features and fixation probability

In order to check how well a given feature discriminates fixated points (actual fixations) from points that are not fixated (control fixations), we have employed the now-standard area under the curve (AUC) measure (Tatler et al., 2005; Vig et al., 2012; Wilming et al., 2011, but cf. Carmi & Itti, 2006), which corresponds to the integral of the receiver–operating characteristic curve. If the feature is always higher at actual fixations than at control fixations, then AUC becomes 1, and if the actual and control fixation distributions of the feature are identical, the AUC is 0.5. Certain factors, such as the central bias in viewing, render the selection of control fixations nontrivial (Tatler, 2007; Tatler & Vincent, 2009). The most common choice in scene-viewing literature (Açık et al., 2010; Einhäuser & König, 2003; Marat et al., 2009; Tatler et al., 2005) is to compare the feature values at fixations during the viewing of a given stimulus with the feature values again taken from this image but at locations that were fixated during the viewing of other stimuli. Because such fixations are themselves a result of natural viewing behavior, they carry the same image content–independent viewing biases. Moreover, if a certain type of analysis requires a specific set of actual fixations, such as fixations following saccades with certain parameters (Tatler et al., 2006), the control fixations can be selected according to the same criteria, taking care of biases that are peculiar to these criteria. In the present study, the feature values of a given frame stimulus at its actual fixations were compared with the feature values of the same stimulus at the locations fixated during the presentation of all other frame stimuli. In the case of movie clips, however, one has to consider time as well (Marat et al., 2009). Fixations have, together with their horizontal and vertical locations, an onset, which corresponds with a specific movie-clip frame. As such, we assigned each fixation to the frame that was on screen at the time of its onset and used the feature value at the fixated location in this frame; this was done both for actual and control fixations. The control fixations for a given movie were all fixations performed during the presentation of all other movie clips with the constraint that the onset of the fixation is not later than the duration of the current clip.

For additional analysis, we have compared AUCs obtained with fixations following shorter saccades and those following longer saccades after dividing the actual and control fixations into two groups with a median split on the sizes of the preceding saccades (Açık et al., 2010). Finally, by choosing only those actual and control fixations that had an onset in a specific temporal interval, we have investigated whether the AUC values changed over time. This was done with a sliding window analysis with a 500-ms window length and with a window overlap of 250 ms between shifts. Moreover, in order to address whether saccades are reactive or predictive, we have repeated the AUC analysis with feature values found in frames that slightly preceded or followed the fixation onset, respectively. All AUCs are computed individually for each stimulus, and median values are reported together with their 95% confidence intervals (CIs), calculated by randomly resampling—from the stimulus-specific AUCs—as many samples as were in the set with replacement and taking the median of that resampled distribution. The question of whether the AUCs for a given feature are different in the two experimental conditions was answered by performing bootstrap-based statistical testing (Efron & Tibshirani, 1993;

Downloaded from jov.arvojournals.org on 03/15/2021

Tatler et al., 2005). Thus, the AUC measure served as the basis of the analysis addressing the relationship between features and fixation probability, that is, the saliency of features.

Statistical dependence between features

Do the features that are employed here display statistical dependencies, and if yes, what is the magnitude of their dependency? In order to address this question, we have first computed the Pearson product-moment correlation coefficient. For each stimulus and condition, separately, the values of two given features either along actual or control fixations were gathered, and the correlation between them was measured. As such, for each stimulus, we have computed four correlation coefficients: movie clip actual, movie clip control, frame actual, and frame control. However, the correlation coefficient addresses solely linear dependencies between two variables and can fail to uncover more complex relationships between static and movement features. Accordingly, we have employed mutual information (MI), which measures the mutual dependence of two random variables, independent of the nature of the dependence. In the discrete case, it is defined as

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \quad (3)$$

where $p(x)$ and $p(y)$ are the marginal probability density functions of the random variables X and Y, and $p(x,y)$ is the joint probability distribution. Choosing two as the log base returns the MI in bits. For the construction of discrete probability density functions, one has to choose a certain bin size, which is significant for reliably estimating the probability. Having too few bins leads to a poor estimation of the probability function, and having too many bins would require a large number of samples to fill those bins so as to allow for a reliable estimate. Because we have only 227 actual fixations per stimulus, we have performed the MI analysis only on the control fixation feature values. In summary, the correlation coefficient and mutual information measures are used to reveal the statistical dependencies between the features.

Results

Fixation data size

We have collected 53,411 fixations while observers were presented the static frame stimuli, and we collected 44,555 fixations during the presentation of movies. The medians (plus or minus standard deviations) over stimuli for the frame and movie conditions are 234 (±112.1) and 189.5 (±99.7) fixations, respectively. Unpaired bootstrap tests (Efron & Tibshirani, 1993) revealed that these medians were significantly different ($p < 10^{-5}$). However, the median of mean fixation durations was 270.8 (±29.4) ms in the frame condition, compared with 334.7 (±60.6) ms in the movie condition, and these differences were, again, highly significant ($p < 10^{-5}$). Combining these two types of information, we have computed the total fixation time for each stimulus, that is, the time that is left after removing the periods in which saccades are present. The median of total fixation time was 61.9 (±35.1) s in the frame condition and 63.0 (±38.6) s in the movie condition, and the difference was not significant ($p = 0.38$). Furthermore, for both stimulus types, fixations with durations shorter than 100 ms amounted to less than 2% of the data, suggesting that few, if any, smooth-pursuit regions were incorrectly labeled as saccades. Thus, even though more fixations were observed in the frame condition, the longer fixation durations while viewing movie stimuli resulted in comparable total fixation times for the two conditions.

Features and fixations

The main aim of the present study was to quantify how well local features differentiate actual and control fixations (see Methods) using the AUC measure. As can be seen in Figure 3, all AUC medians over stimuli and the lower bounds of 95% CIs were clearly above the chance level of 0.50. This shows that all four features were partially successful in distinguishing fixated locations in both stimulus conditions. MME AUC dropped from 0.59 (lower CI boundary/upper CI boundary, 0.58/0.62) in the movie condition to 0.57 (0.55/0.59) in the frame condition (bootstrap test for equal median, $p = 0.015$). A similar decrease from 0.61 (0.59/0.62) to 0.57 (0.55/0.59) was observed for MDC ($p = 0.0003$). This demonstrates a systematic difference between the movement feature AUCs of the two conditions. LC AUC was 0.59 (0.58/0.61) in the movie condition and was not significantly different in the frame condition ($p = 0.45$) with 0.59 (0.57/0.61). The corner detector of intrinsic dimensionality measure remained at 0.61 (0.59/0.63 for both conditions, $p = 0.45$). This result shows that the static features have comparable predictive value in static and movie conditions. In summary, whereas static features displayed a high and constant predictive value of fixation locations, the predictive value of movement features dropped for static stimuli in comparison with movie clips while still remaining well above chance level.

Downloaded from jov.arvojournals.org on 03/15/2021

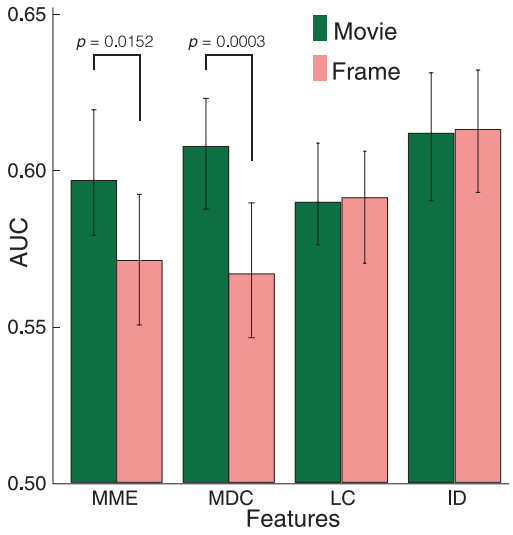

Figure 3. Fixation predictability (saliency) of features. AUC values for each feature and stimulus condition are shown together with significant across-stimulus condition comparisons. Whereas the discrimination performances of movement features are higher in the movie condition, no such difference exists in the case of static features.

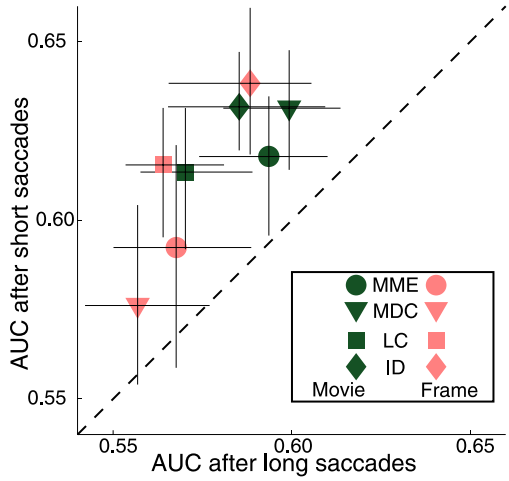

Figure 4. Saccade size influence on the fixation predictability of features. It can be clearly seen that for both stimulus conditions and all features used, the AUC of fixations following shorter saccades is higher.

Does it matter that a given stimulus is viewed before or after its frame or movie counterpart? During the first half of the experiment, all stimuli were shown for the first time, and during the second part of the experiment, the frame and movie counterparts of the first half stimuli were presented. Accordingly, we could address whether having seen one version of a stimulus—movie or frame—during the first half of the experiment influenced viewing the stimuli that appeared during the second half. For each feature and stimulus condition, we have compared the stimulus-specific AUCs obtained for the different parts of the experiment with permutation tests. None of the eight comparisons yielded statistically significant differences (all ps > 0.35 without multiple comparison correction). Thus, prior exposure to a stimulus did not change the fixation and feature relationship when it was later viewed with different motion content.

Do the features display greater discriminability after shorter saccades? In order to answer this question, we have median-split both the actual and control fixations of each stimulus according to the amplitude of the saccade that preceded it and then computed the AUCs. For all features and stimulus conditions, the AUCs obtained from the shorter saccade group are greater than the larger saccade group's AUCs (Figure 4).

Temporal aspects of feature-related viewing

In order to check whether fixation discriminability of features decreases over time, we have performed a sliding window analysis (Figure 5). Linear regression slopes (AUC change per second) for these time series were computed both from the whole data and from the bootstrap samples of the data that were also used for CI estimations of the time-specific medians (shaded regions in Figure 5). In the case of the ID feature, the 95% CIs for the slope included zero, suggesting that the linear decrease in AUCs was statistically unreliable. For LC, the slopes were –0.014 (–0.027/–0.002) and –0.018 (–0.033/–0.003) for the frame and movie conditions, respectively. The movement features, on the other hand, have stronger AUC decreases over time. In the frame condition, MME displayed a slope of –0.041 (–0.060/–0.023) and MDC a slope of –0.027 (–0.043/–0.013). The slopes in the movie condition were –0.026 (MME, –0.050/–0.006) and –0.023 (MDC, –0.048/–0.005). Two-sample Kolmogorov-Smirnov tests on bootstrapped slope distributions revealed that the AUC decrease in the frame condition was faster only in the case of movement features ($p < 0.05$). Thus, the relationship between fixation probability and feature values is higher at the start of the stimulus presentation in the case of movement features with small but significant decreases over time for LC, too.

The above results bring to mind the question of whether the MDC and MME AUC differences between the two experimental conditions arise over time. Resampling tests reveal that a significant difference between the conditions appears only after 1 s for MME ($p = 0.008$) and after 0.75 s for MDC ($p = 0.005$), and the difference remains significant ($p < 0.05$) for at least another second. That is, movement features are equally effective in detecting fixations for movie and frame conditions during the early phase of the stimulus

Downloaded from jov.arvojournals.org on 03/15/2021

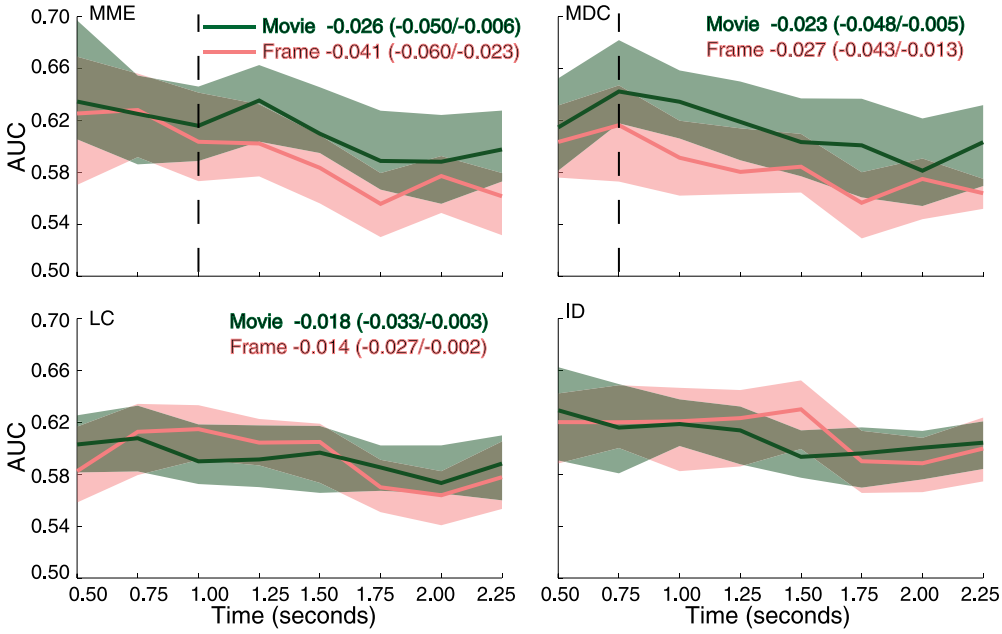

Figure 5. Fixation predictability of features as a function of time. AUCs are computed in 0.5-s-long temporal windows with the first window centered at 0.5 s after stimulus onset. Earlier fixations are not considered because many of them are at the center of the screen due to the preceding drift-correction cross. The vertical dashed lines denote the first-time point from the time on the movie and frame conditions at which AUCs are significantly different ($p < 0.05$) for at least 1 s. The shaded regions cover the 95% CIs, bootstrapped. The slopes of the linear fits and their 95% CIs are given in insets. In the case of ID, the CIs included 0, and hence, the statistics are not shown. For clarity, the fits themselves are not drawn.

presentation, and a difference in favor of the former condition appears thereafter.

A second question concerning the role of time that is unrelated to the above analysis deals with how well feature values at a fixated location discriminate fixations before or after the fixation onset. This analysis extends the AUC calculations computed on the frame that was visible at the onset of a fixation to frames that precede and follow the fixation (Figure 6). Please note that in the case of static stimuli this analysis employs frames that are not presented. As can be seen, the CIs for different time points on a given curve are largely overlapping. Even though visual inspection suggests that movement features in the static condition induce a prediction, resample tests reveal that not a single median difference between two points on a given curve is significant despite the lack of correction for multiple comparisons (36 comparisons for each curve, all ps $>$ 0.05). Thus, the fixation discrimination ability of a given feature is roughly equal if AUCs are computed with the feature values taken from frames that appear slightly before or after the fixation onset.

Dependencies between features

In order to see whether the four features employed in the study have linear dependencies among them, Pearson's correlation coefficient is computed for each feature pair along the control fixations. This is done separately for each stimulus and experimental condition. The median (along stimuli) correlations within static features (ID and LC) were positive and rather high: frame condition $R = 0.54$ (± 0.15) and movie condition $R = 0.53$ (± 0.17). Within movement features (MME and MDC), correlation coefficients were even higher: frame condition 0.62 (± 0.26) and movement condition 0.60 (± 0.20). The between-movement and static-feature correlations, on the other hand, were much closer to zero: LC MME 0.01 (± 0.24) and 0.00 (± 0.21); LC MDC 0.17 (0.19) and 0.15 (0.16); ID MME -0.02 (± 0.25) and -0.04 (0.23); and ID MDC 0.13 (0.20) and 0.10 (0.18) for the frame and movie conditions, respectively. Thus, even though the correlations within static features and within movement features were positive and high, the linear dependencies between static and movement feature pairs are relatively weaker.

Because the statistical dependencies between movement and static features do not have to be solely linear, mutual information is a more appropriate measure. As can be seen in Figure 7, the MI results agree with the correlation results. The information gained about a static feature by observing another static feature is at least double the information obtained regarding a movement feature by observing a static feature. The difference of MME-MDC MI across two stimulus

Downloaded from jov.arvojournals.org on 03/15/2021

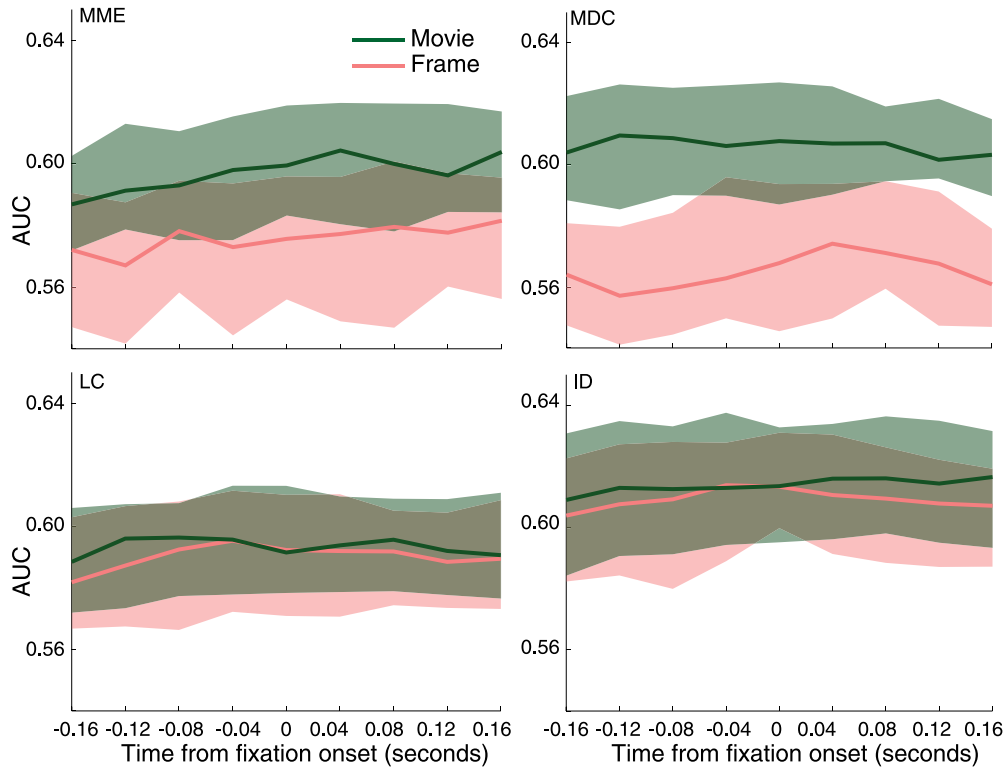

Figure 6. Fixation predictability of features before and after fixation onset. In the preceding analysis, AUCs are computed with feature values taken from the frames that were visible exactly at the onset of fixations. Here, the same analysis (time point zero) is repeated with frames that appeared just before (negative time points) and just after (positive time points) the fixation onset. Negative time points correspond with frames that precede the fixation onset and the positive ones to frames that follow the fixation onset. The shaded regions cover the 95% CIs (bootstrapped). In order to keep the data analyzed for each frame constant, fixations on the first and last four frames were discarded. Even though the movement features in the frame condition display higher AUC values on the right and suggest predictive fixations, the comparisons with other time points do not reach significance (all $ps > 0.05$, no correction for multiple comparisons). Note that despite the visual resemblance to Figure 5, the analysis performed and the data included are different.

conditions can be explained away considering that the feature values of the frame stimuli are only a small subset of the feature values found in the movie stimuli. The mutual information analysis conclusively shows that feature pairs consisting of one movement and one static feature display much less statistical dependency, compared with what one observes between LC and ID or between MME and MDC.

Discussion

We have recorded the eye movements of human observers who viewed high-resolution natural movies and static frames taken from those movies in the absence of an explicit task. Akin to looking at a photograph of objects in motion, the latter condition allowed us to perform analysis aimed at uncovering the degree to which natural-image viewing differs from what happens when confronted with dynamic stimuli. We have considered the role of two types of spatially local features in the guidance of attention: static features that are computed on single frames and movement features selective for local motion for their computations are based on differences between successive movie frames. Comparing fixated locations and those that are not fixated on frames in terms of movement features that are extracted from movies enabled us to address two different but related questions: To what degree are these features causal for or correlated with higher probability of fixation, and does implied motion—motion deduced from static cues in the absence of real motion—display a role in gaze allocation? The present study fills a gap in the study of eye guidance by characterizing the role of movement features while the visual system is coupled with static and dynamic scenes.

Before we discuss our findings' implications, we want to comment on one possible source of criticism for the eye-movement analysis performed here. For both dynamic and static stimulus conditions, we have used

Downloaded from jov.arvojournals.org on 03/15/2021

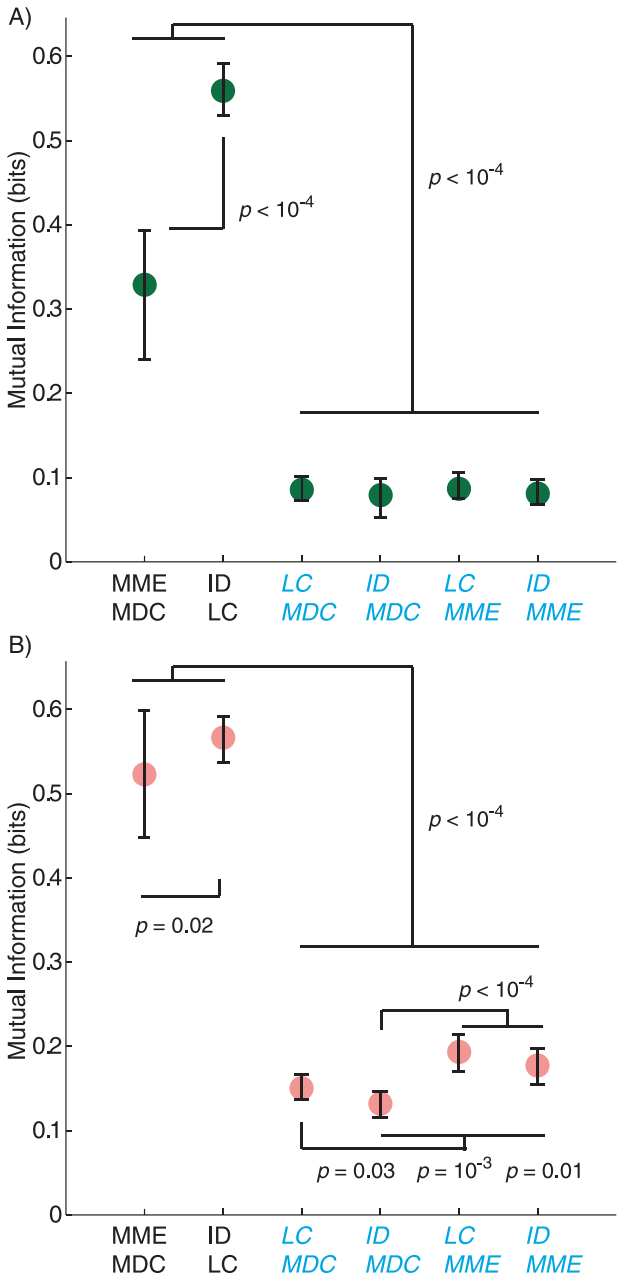

Figure 7. Statistical-dependence analysis. For each stimulus separately, the values of two features were measured along the control fixations, and the MI in bits between these two feature distributions was computed. Shown are the medians and 95% CIs of the MI. Note that whereas the MI of within-movement (MME-MDC) and within-static (LC-ID) feature pairs is relatively high, the MI for movement-static feature pairs is low. Bootstrap test results are shown only for $p < 0.05$.

the same parameters for detecting saccades and performed the analysis on fixation periods that are left after the removal of saccades. Accordingly, smooth pursuit—movement of the eyes while an object in motion is tracked (Robinson, 1965)—was not addressed explicitly. Although some studies (e.g., Dorr et al., 2010) of movie viewing carefully exclude smooth-pursuit periods from their data, others (e.g., Marat et al., 2009) considered all eye-position samples regardless of whether they were fixations, saccades, or smooth pursuit. Our approach falls into a third category of studies in which the data are analyzed at the endpoints of fast eye movements (e.g., Böhme et al., 2006; Itti & Baldi, 2009; Vig, Dorr, Martinetz, & Barth, 2011). Because eye position at the fixation onset is expected to fall on a region that drew attention while still at the periphery, for the question of why do we look where we do (Schütz et al., 2011), it does not matter whether the eyes will later move following the object featured in that region. Importantly, the fixation detection algorithm employed here detected more fixations in the frame condition and fixations lasting longer in the movie condition. If the algorithm was incorrectly labeling smooth-pursuit movements as saccades, the opposite would be expected. Moreover, regardless of the stimulus type, there were few fixations with durations shorter than 100 ms. We have shown that the total amount of time spent during fixations is very similar across conditions because the fixation durations in the frame condition are longer. 't Hart et al.'s (2009) results obtained from a comparison of movie and stop-motion conditions, in which single frames from a movie were shown for 3 s, are identical to our observations. Thus, even though we did not address smooth-pursuit movements directly, given that we had fewer fixations during movie viewing and that the vast majority of fixation durations were longer than 100 ms, we are confident that our choice of analyzing image properties at saccadic target regions is legitimate.

Our results revealed that both static and movement features predict fixations better than chance regardless of the movement content of the stimulus. However, only static features, LC, and ID, maintain identical saliency for both frames and movies. Furthermore, the former feature is relatively less salient, confirming earlier observations (Açık et al., 2010; Saal et al., 2006). On the other hand, movement features introduced here, which quantify the average (MME) and deviation (MDC) of local motion, display high fixation prediction performance similar to ID only in the case of movies. Their performance is significantly worse yet well above chance for the static stimuli. Moreover, even though all features had higher saliency for fixations in both stimulus conditions following shorter saccades, thus replicating previous findings (Açık et al., 2010; Tatler et al., 2006), the condition differences were identical. These results might suggest that movement features have separable correlational and causal contributions to fixation selection, which has previously been claimed for luminance and color features (Açık et al., 2009; Einhäuser & König, 2003; Frey et al., 2011). This would entail that movement features

Downloaded from jov.arvojournals.org on 03/15/2021

predict fixations on static scenes because they are correlated with other features or properties, such as object locations (Einhäuser, Spain, & Perona, 2008), which are the actual causes of overt attentional allocation. However, if the stimulus contains movement, the causal role of movement features is added on top of the indirect contribution, and this is measured as even higher fixation predictability. This scenario is reminiscent of what Frey and colleagues (2011) have described with color features. They took images that were rich in color content and gathered eye-movement data from people who viewed these images either with intact color content or after one of the red-green or blue-yellow channels was removed. In color-intact images, the saliency of blue-yellow contrast was very low, but red-green contrast predicted fixations reasonably well. Crucially, red-green contrast remained predictive of fixation locations even after the removal of that channel albeit less than before (Frey et al., 2011). Thus, fixation predictability difference of movement across static and dynamic viewing conditions suggests two attentional mechanisms: one relying on motion and corresponding to a causal mechanism, and a second motion-independent mechanism that nevertheless reveals a correlation between motion and fixation probability.

A close inspection of the temporal course of movement feature saliencies reveals that there is no need to postulate two separate mechanisms. We have shown that static feature saliencies are characterized with either nonexistent (ID) or very weak (LC) decreases over time. An earlier study (D. Parkhurst et al., 2002), revealing a prominent temporal saliency decrease for static features, has been criticized (Tatler et al., 2005) due to a biased selection of control fixation points. Studies controlling for that bias have either failed to uncover a temporal variation of static saliency (Tatler, 2007; Tatler et al., 2005), or it was observed for a small subset of the presented stimuli (Açık et al., 2009). Crucially, however, movement feature saliencies, measured here with the same unbiased control fixations, decrease with time for both dynamic and static scenes. That is, the fixation predictability of these features is highest during the first second of the stimulus presentation, and it decreases gradually thereafter. To our knowledge, the only two studies that have analyzed the time course of movement saliency in movies have observed the same attenuation (Carmi & Itti, 2006; Marat et al., 2009). Our results generalize those findings to static stimuli. Most importantly, during the first second of viewing, the predictability is comparable for movie and frame fixations. That is, in the early phase of viewing, the saliencies of movement features are the same regardless of whether the scene is dynamic or not. Only after that, the movie fixations are predicted better because the saliency in the case of static scenes shows a faster decrease over time. The fact that movement features predict fixations equally well for static and dynamic scenes in the early phase of viewing argues against an explanation based on separate contributions from causal and correlational roles of saliency.

How come movement features are equally salient during the viewing of dynamic and static scenes in the early phase of viewing? Many theoretical accounts (Freyd, 1987; Gibson, 1979) suggest that the default mode of visual perception assumes the dynamicity of the world. That is, even if the perceptual system is presented a "snapshot" of the environment (Freyd, 1987), the corresponding visual processing proceeds as if the stimulus is dynamic. Strong experimental support for these claims comes from studies on implied motion (Hubbard, 1995; Kourtzi & Kanwisher, 2000), motion that is deduced from static cues in the absence of real movement. For several types of motion (Freyd & Finke, 1984; Freyd & Jones, 1994; Hubbard, 1995), it has been shown that participants who view successive snapshots of a continuous movement misremember the final stimulus as depicting a later stage of that movement. Freyd (1983) demonstrated the same distortion in memory with a single snapshot of a real-world movement. This suggests that the static stimuli used in these studies are represented dynamically, and the implied dynamicity affects the participants' memory (Freyd, 1987; Gilden, Blake, & Hurst, 1995; Hubbard, 1995). Our results reveal that implied motion relates to cognition already during interaction with the stimulus before it is represented in the memory. The fact that the perception of static natural stimuli that imply bodily movement activate cortical visual motion areas provides additional support for this argument (Kourtzi & Kanwisher, 2000; Proverbio et al., 2009). When first confronted with the static scenes, our subjects looked at regions that would be moving in the corresponding movies, that is, at sources of implied motion. Either due to mechanisms of sensory adaptation, as observed with imaginary visual motion (Gilden et al., 1995), or simply because most motion-containing regions are already looked at, the movement feature saliency decreases after the first fixations. Because dynamic stimuli reveal new sources of motion, the fixation predictability of movement remains high. In sum, motion, regardless of whether it is real or deduced from static cues as in the case of implied motion, predicts fixated locations reasonably well.

Even though implied motion can explain the temporal decrease of movement saliency in static scenes, why we, and others (Carmi & Itti, 2006; Marat et al., 2009), have observed a similar temporal decrease with movies—albeit a slower one—remains to be answered. A large body of studies that employ artificial stimuli (Abrams & Christ, 2003; Hillstrom & Yantis,

Downloaded from jov.arvojournals.org on 03/15/2021

1994; Yantis & Egeth, 1999; Yantis & Jonides, 1984, 1990) demonstrates that it is the onset of visual motion and not movement per se that captures attention in search tasks. In two simple but elegant experiments, Abrams and Christ (2003) demonstrated both attentional capture and the following inhibition of return (Klein, 2000) for items that have just started moving. They presented static and moving placeholders, which have either kept or reversed this property after a few seconds. When a target appeared at the placeholder that was initially stationary and started moving later, the response, relative to target appearance at other placeholders, was very fast if the movement transition coincided with the target appearance. Moreover, it was slowed down if a delay occurred between the two events, demonstrating inhibition of return (Abrams & Christ, 2003). Our observation of higher movement saliency in the first second of viewing can thus be explained with the appearance of multiple motion onsets. When the movie first starts, all movement therein is novel and, as such, attracts attention. With time, however, the already existing motion is not salient anymore, and it is only those movement sources that enter the frame after the start of the presentation that attract attention. That is, the onset of motion, real or implied, is more salient than movement as such, and accordingly, movement feature saliencies are highest when the stimulus first appears.

Do the movement and static features capture the same aspects of saliency? The present correlation coefficient and mutual information analyses have demonstrated that, whereas the statistical dependencies within movement features and within static features are relatively high, movement and static feature pairs display statistical independence. Saliency map models (Itti & Koch, 2000; Itti et al., 1998) combine the information that different features provide in order to optimize fixation predictability. Which features to include in the model and how to choose the weights for different feature channels are open questions in the field (Zhao & Koch, 2011). Because the predictability of features that are statistically dependent on one another would not add up optimally (Peng et al., 2005), the statistical independence of features is a useful criterion for saliency modeling. Here, we show that movement and static features establish such independent pairs, which explains how previous saliency map studies have reached significantly better results following the integration of static and movement features, compared with their individual performance (Carmi & Itti, 2006; Marat et al., 2009). Together with promoting the inclusion of both movement and static features in saliency maps, the statistical independence between these two feature types reveals that they capture different aspects of saliency.

So far, we have argued that motion is salient even in the absence of real motion. Motion can be either a causal determinant of overt visual attention or correlate with other low- or high-level scene properties, which are the actual causes of fixation selection. The absence of statistical dependencies between movement and static features rules out a correlation between motion and other low-level features. At first sight, the causal determinant stance appears as a parsimonious conclusion. Alternatively, the gaze might be directed to different locations according to the knowledge and expectations of the observer in a task-dependent manner (Cristino & Baddeley, 2009). Low-level features, such as color and motion, are statistically dependent on object identities, and humans learn such regularities in everyday life. Hansen, Gegenfurtner, and colleagues (Hansen, Olkkonen, Walter, & Gegenfurtner, 2006; Olkkonen, Hansen, & Gegenfurtner, 2008; Witzel, Valkova, Hansen, & Gegenfurtner, 2011) showed their participants objects with known colors, such as a red strawberry or a yellow German mailbox. If the participants were told to adjust the color of the objects in order to make them appear gray, they would overshoot the target color along the opponent color axis, such as making the mailbox slightly bluish and the strawberry greenish. The effect was not observed for meaningless shapes or natural objects that appear in several different colors (Hansen et al., 2006; Olkkonen et al., 2008). Thus, subjective color perception depends on previous knowledge about object identity and cannot be explained solely with bottom-up processing. Similarly, we possess knowledge of object postures that are diagnostic of motion, which defines the very notion of implied motion (Freyd, 1983; Kourtzi & Kanwisher, 2000). Both eye-tracking (Einhäuser, Spain, et al., 2008) and modeling studies (Wischnewski, Belardinelli, Schneider, & Steil, 2010) suggest that attentional selection prioritizes object-level analysis of scenes. In the comprehensive visual attention model introduced by Wischnewski and colleagues (2010), static and dynamic features are used not in order to guide attention to certain salient pixels, but to extract proto-objects in the scene. These proto-objects carry location, color, and rough shape information, and selective attention operates at this medium level of visual representation. Importantly, the extraction of these proto-objects enables a straightforward addition of task-dependent control to the attention model (Wischnewski et al., 2010). Accordingly, if attention is guided by knowledge about the environment and aimed at uncovering informative scene regions, such as objects, low-level features, including movement, might still correlate with fixation probability.

The goal of the present study was not optimal fixation prediction but rather a comparison of movement saliencies in dynamic and static viewing condi-

Downloaded from jov.arvojournals.org on 03/15/2021

tions. Still, we would like to comment on our effect size in relation to previous studies (Carmi & Itti, 2006; Itti & Baldi, 2009; Le Meur et al., 2007; Marat et al., 2009; Vig et al., 2009; Vig et al., 2012). This is not straightforward because the metrics quantifying saliency differ across studies (Wilming et al., 2011). Here, we have employed the AUC measure that has become a standard since its introduction into the field (Tatler et al., 2005). A recent review (Wilming et al., 2011) combining theoretical and empirical evaluations of several saliency metrics concludes that for the type of the data recorded here AUC is the best choice. Furthermore, methodological preferences, such as the choice of which regions that are not fixated will be compared with fixated ones (Carmi & Itti, 2006; D. Parkhurst et al., 2002; Tatler, 2007), have a large influence on the value of the saliency metric. Most importantly, the central bias (D. Parkhurst et al., 2002; Tatler 2007)—the tendency of observers to select relatively more saccade endpoints around the center of the screen—appears as a strong predictor of fixations in the absence of any information about image content (Tatler 2007). Accordingly, Wilming and colleagues (2011) suggest that the fixation predictability of central bias must be a lower boundary that any saliency measure has to surpass. For instance, the dynamic saliency model that Le Meur et al. (2007) proposed, albeit better than other saliency models they have tested, performs worse than the central bias. One way to control for the central bias, as is done in the current study, is to select actual and control fixations with identical spatial (Açık et al., 2009; Einhäuser & König, 2003; Tatler et al., 2005) and temporal (Vig et al., 2009; Vig et al., 2012) distributions. The present bias-controlled dynamic feature AUC of 0.60 is comparable with the saliency of individual static features previously reported (Açık et al., 2010; Mital et al., 2011; Tatler et al., 2005). It is of note here that dynamic feature AUCs reach higher values around 0.70 for those portions of video viewing in which the fixated locations of different participants tend to form tight clusters (Mital et al., 2011). Carmi and Itti (2006) use low-resolution movie clips and a percentile metric that, as AUC, is bound between 0.50 and 1 and report values around 0.70 for their dynamic saliency map modeling. Nevertheless, they take as the actual fixation point the maximally salient location in a circle with a radius of 3.15° around the measured saccade endpoint (Carmi & Itti, 2006); accordingly, the real saliency is expected to be lower. The "surprise" measure (Itti & Baldi, 2009) quantifies the unexpectedness of spatiotemporal events in movies using Bayes' theorem. Even though no direct comparisons are made, the surprise measure appears to outperform the previous models of Itti and colleagues (Carmi & Itti, 2006; Peters et al., 2005, cf. Itti & Baldi, 2009). The most recent spatiotemporal intrinsic dimensionality analysis of Vig and colleagues (2012) reaches a bias-free AUC of 0.70, which is, to the best of our knowledge, the highest value reported in studies that employ free viewing and similar stimuli. What is common to these dynamic saliency studies is their integration of dynamic and static information. While the Itti lab (Carmi & Itti, 2006; Itti & Baldi, 2009) computes static and dynamic features separately and then pools them, Vig et al. (2012) measure the saliency of spatiotemporal features. Moreover, in these studies, the features of interest are computed at multiple temporal and spatial scales, and fixation predictability is measured after the information from all scales is combined. Here, we have computed each feature at a single spatial and temporal scale, and the dynamic features were extracted from differences between two frames only. In summary, for studies aimed at optimal fixation prediction, combining dynamic and static information at several temporal and spatial scales appears to be the most fruitful methodology, an idea that our statistical dependence analysis supports.

Are saccades reactive—that is, do they bring the eyes to locations where salient events just happened—or are they predictive in the sense that the eyes land on a location just before the salient event moves there? Some previous studies either assume (Itti & Baldi, 2009) or measure (Vig et al., 2009) that the saliency of fixated regions is at maximum just before the fixation onset. Interestingly, the later work of Vig and colleagues (2010) shows that this reactive characteristic is stimulus-dependent. If the stimuli contain camera movement and scene cuts, the saccades are indeed reactive, but if the dynamic stimuli are more natural, the saccades have a small anticipatory role. The latter observation is in line with studies performed in real environments with everyday tasks (Land & McLeod, 2000; Patla & Vickers, 2003). Land and McLeod (2000) showed that cricket batsmen make predictive saccades to locations at which they expect the thrown ball to bounce. In a similar vein, while walking, humans fixate locations on the floor approximately 1 s before the foot lands at that location (Patla & Vickers, 2003). Our results reveal small and statistically unreliable differences for such reactive and predictive roles that vary with stimulus and feature type. The absence of an explicit natural task in the current study can explain why we have missed this ecological property of everyday gaze behavior. Furthermore, because natural stimuli are characterized by large-scale temporal correlations (Einhäuser, Kayser, König, & Körding, 2002; Kayser, Einhäuser, & König, 2003), analyzing the saliency of features in the temporal periphery of all fixation onsets might have been inadequate. Future studies that restrict themselves to fixations on regions that are task-related or contain large spatiotemporal changes might be more

Downloaded from jov.arvojournals.org on 03/15/2021

appropriate for addressing whether saccades are reactive or predictive.

Conclusions

Here, we have demonstrated that during the viewing of natural scenes, local movement features are salient even in the absence of any real motion. The saliency of these features is comparable with static features with good fixation predictability in the early phase of viewing. This is because motion deduced from static cues—that is, implied motion—is enough to reveal the movement in the scene. Moreover, we have argued that motion onset is more salient than movement as such. Movement features display statistical independence with static features, which promotes their inclusion in saliency map models. We conclude that as we sample the environment with eye movements, even static stimuli are treated as dynamic because the visual system is attuned to a changing environment.

Keywords: eye movements, motion perception, natural scene statistics, visual attention

Acknowledgments

We are grateful to Johannes Steger for technical assistance and to Till Becker for suggestions on data analysis. Alper Açık's work was supported by ERC-2010-AdG #269716 - MULTISENSE. Peter König's work was supported by the Cognition and Neuro-ergonomics/Collaborative Technology Alliance grant #W911NF-10-2-0022.

Commercial relationships: none.
Corresponding author: Alper Açık.
Email: alper.acik.81@gmail.com.
Current address: Özyeğin University, Psychology Dept, Çekmeköy Campus Nişantepe District, Çekmeköy — Istanbul.

References

Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science, 14,* 427–432.

Açık, A., Onat, S., Schumann, F., Einhäuser, W., & König, P. (2009). Effects of luminance contrast and its modifications on fixation behavior during free viewing of images from different categories. *Vision Research, 49,* 1541–1553.

Açık, A., Sarwary, A., Schultze-Kraft, R., Onat, S., & König, P. (2010). Developmental changes in natural viewing behavior: Bottom-up and top-down differences between children, young adults and older adults. *Frontiers in Psychology, 1*(1207), 1–14, doi: 10.3389/fpsyg.2010.00207.

Anandan, P. (1989). A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision, 2,* 283–310.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46,* 2824–2833.

Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., & Szeliski, R. (2011). A database and evaluation methodology for optical flow. *International Journal of Computer Vision, 92,* 1–31.

Black, M., & Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding, 63,* 75–104.

Black, M., & Jepson, A. (1996). Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18,* 972–986.

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

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

Cristino, F., & Baddeley, C. (2009). The nature of the visual reperesentations involved in eye movements when walking down the street. *Visual Cognition, 17*(6/7), 880–903.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10):28, 1–17, http://www.journalofvision.org/content/10/10/28, doi:10.1167/10.10.28. [PubMed] [Article]

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall, Ltd.

Einhäuser, W., Kayser, C., König, P., & Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience, 15,* 475–486.

Downloaded from jov.arvojournals.org on 03/15/2021

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17,* 1089–1097.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1–19, http://journalofvision.org/content/8/2/2/, doi:10.1167/8.2.2. [PubMed] [Article]

Einhäuser, W., Schumann, F., Vockeroth, J., Bartl, K., Cerf, M., Harel, J., Schneider, E., & König, P. (2009). Distinct roles for eye and head movements in selecting salient image parts during natural exploration. *Annals of the New York Academy of Sciences, 1164,* 188–193.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14):18, 1–26, http://journalofvision.org/content/8/14/18/, doi:10.1167/8.14.18. [PubMed] [Article]

Felsen, G., & Dan, Y. A natural approach to study vision. *Nature Neuroscience, 8,* 1643–1646.

Foulsham, T., Walker, E., & Kingston, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research, 51,* 1920–1931.

Frey, H.-P., Wirz, K., Willenbockel, V., Betz, T., Schreiber, C., Troscianko, T., & König, P. (2011). Beyond correlation: Do color features influence attention in rainforest? *Frontiers in Human Neuroscience, 5*(36), 1–13, doi:10.3389/fnhum.2011.00036.

Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception and Psychophysics, 33,* 575–581.

Freyd, J. J. (1987). Dynamic representations. *Psychological Review, 94,* 427–438.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 126–132.

Freyd, J. J., & Jones, K. T. (1994). Representational momentum for a spiral path. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 968–976.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gilden, D., Blake, R., & Hurst, G. (1995). Neural adaptation of imaginary visual motion. *Cognitive Psychology, 28,* 1–16.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*(11), 1367–1368.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9,* 188–194.

Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception & Psychophysics, 55,* 399–411.

Horn, B., & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence, 17,* 185–203.

Hubbard, T. L. (1995). Environmental invariants in the representation of motion: Implied dynamics and representational momentum, gravity, friction, and centripetal force. *Psychonomic Bulletin & Review, 2,* 322–338.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the I.R.E, 40*(9), 1098–1102.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 29,* 1295–1306.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506.

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

Kanai, R., Paffen, C. L. E., Hogendoorn, H., & Verstraten, F. A. J. (2006). Time dilation in dynamic visual display. *Journal of Vision, 6*(12):8, 1421–1430, http://journalofvision.org/content/6/12/8/, doi:10.1167/6.12.8. [PubMed] [Article]

Kayser, C., Einhäuser, W., & König, P. (2003). Temporal correlations of orientations in natural scenes. *Neurocomputing, 52,* 117–123.

Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision, 9*(5):7, 1–15, http://journalofvision.org/content/9/5/7, doi:10.1167/9.5.7. [PubMed] [Article]

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. *Journal of Cognitive Neuroscience, 12,* 48–55.

Krekelberg, B., Dannenberg, S., Hoffmann, K.-P., Bremmer, F., & Ross, J. (2003). Neural correlates of implied motion. *Nature, 424,* 674–677.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by

Downloaded from jov.arvojournals.org on 03/15/2021

saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision, 13,* 201–214.

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

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

Lappe, M. (2000). Computational mechanisms for optic flow analysis in primate cortex. *International Review of Neurobiology, 44,* 235–268.

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.

Machner, B., Dorr, M., Sprenger, A., von der Gablentz, J., Heide, W., Barth, E., & Helmchen, C. (2012). Impact of dynamic bottom-up features and top-down control on the visual exploration of moving real-world scenes in hemispatial neglect. *Neuropsychologia, 50,* 2415–2425.

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

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

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision, 8*(5):13, 1–16, http://journalofvision.org/content/8/5/13, doi:10.1167/8.5.13. [PubMed] [Article]

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

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt attention in natural scenes. *European Journal of Neuroscience, 19,* 783–789.

Patla, A. E., & Vickers, J. N. (2003). How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Experimental Brain Research, 148,* 133–138.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27,* 1226–1238.

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

Proverbio, A. M., Riva, F., & Zani, A. (2009). Observation of static pictures of dynamic actions enhances the activity of movement-related brain areas. *PLoS ONE, 4*(5), e5389, doi:10.1371/journal.pone.0005389.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems, 10,* 1–10.

Robinson, D. A. (1965). The mechanics of human smooth pursuit eye movement. *Journal of Physiology, 180,* 569–591.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14):16, 1–20, http://journalofvision.org/content/7/14/16, doi:10.1167/7.14.16.

Saal, H., Nortmann, N., Krüger, N., & König, P. (2006). Salient image regions as a guide for useful visual features. *IEEE AICS 2006*. Sheffield Hallam University.

Schumann, F., Einhäuser-Treyer, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision, 8*(14):12, 1–17, http://journalofvision.org/content/8/14/12/, doi:10.1167/8.14.12. [PubMed] [Article]

Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, *11*(5):9, 1–30, http://www.journalofvision.org/content/11/5/9, doi:10.1167/11.5.9. [PubMed] [Article]

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://journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W. (2009). Current understanding of eye guidance. *Visual Cognition, 17,* 777–789.

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

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research, 46*(12), 1857–1862.

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

Downloaded from jov.arvojournals.org on 03/15/2021

1–23, http://www.journalofvision.org/content/11/5/5, doi:10.1167/11.5.5. [PubMed] [Article]

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

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

Thompson, E., & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences, 5*(10), 418–425.

Tootell, R. B., Silverman, M. S., Switkes, E., & De Valois, R. L. (1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science, 218,* 902–904.

Vig, E., Dorr, M., & Barth, E. (2009). Efficient visual coding and the predictability of eye movements on natural movies. *Spatial Vision, 34,* 1080–1091.

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

Vig, E., Dorr, M., Martinetz, T., & Barth, E. (2012). Intrinsic dimensionality predicts the saliency of natural dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34,* 1080–1091.

Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and limits of models of fixation selection. *PLoS ONE, 6*(9), e24038, doi:10.1371/journal.pone.0024038.

Wischnewski, M., Belardinelli, A., Schneider, W. X., & Steil, J. J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation, 2,* 326–343.

Witzel, C., Valkova, H., Hansen, T., & Gegenfurtner, K. R. (2011). Object knowledge modulates color appearance. *i-Perception, 2,* 13–49.

Yamamoto, K., & Miura, K. (2012). Time dilation caused by static images with implied motion. *Experimental Brain Research*, *223*(2), 311–319.

Yantis, S., & Egeth, H. E. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 661–676.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 601–621.

Yantis, S., & Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 121–134.

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

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3):9, 1–15, http://www.journalofvision.org/content/11/3/9, doi:10.1167/11.3.9. [PubMed] [Article]

Downloaded from jov.arvojournals.org on 03/15/2021