Are fixations in static natural scenes a useful predictor of attention in the real world?

Tom Foulsham[1]* and Alan Kingstone[2]

[1] Department of Psychology, University of Essex, UK

[2] Department of Psychology, University of British Columbia, Canada

*Corresponding author:

Department of Psychology, Wivenhoe Park, Colchester, Essex, UK, CO4 3SQ.

Tel. (+44) 1206 874159. Email: foulsham@essex.ac.uk

Running head: Comparing attention in real world and static scenes

Keywords: attention, eye movements, scene perception

Word count (main body): 7275

Abstract

Research investigating scene perception normally proceeds in laboratory experiments using static images. Much has been learned about how observers look at pictures of the real world and the attentional mechanisms underlying this behaviour. However, the use of static, isolated pictures as a proxy for studying everyday attention in real environments has led to the criticism that such experiments are artificial. We report a new study that tests the extent to which the real world can be reduced to simpler laboratory stimuli. We recorded the gaze of participants walking on a university campus with a mobile eye tracker, and then showed static frames from this walk to new participants, in either a random or sequential order. The aim was to compare the gaze of participants walking in the real environment with fixations on pictures of the same scene. The data show that picture order affects inter-observer fixation consistency and changes looking patterns. Critically, while fixations on the static images overlapped significantly with the actual real-world eye movements, they did so no more than a model that assumed a general bias to the centre. Remarkably, a model that simply takes into account where the eyes are normally positioned in the head—independent of what is actually in the scene—does far better than any other model. These data reveal that viewing patterns to static scenes are a relatively poor proxy for predicting real world eye movement behaviour, while raising intriguing possibilities for how to best measure attention in everyday life.

Introduction

A central goal of cognitive psychology is to understand natural information processing—including everyday attention—through controlled experiments with laboratory stimuli. However, it has often been acknowledged that this method may fail to capture everyday human functioning due to low ecological validity (Neisser, 1976; Gibson, 1979; Kingstone et al., 2003, 2008; Risko et al., 2012). For example, Gibson worried that our choice of laboratory stimuli did not always preserve the specific information and environmental context in which people operate. In the present study, we investigated this issue with regard to experiments attempting to measure attention in natural scenes.

In everyday life, humans typically align their body, head and eyes with items they wish to pay attention to. While this alignment is not mandatory (we can covertly attend to locations away from fixation, or focus on internal thoughts), measuring the distribution of fixations has become a standard way of assessing attention. In particular, fixations can be measured during the viewing of complex, naturalistic stimuli as a way to understand everyday attention. A large number of experiments have been conducted which use eyetracking to determine attended locations in natural images (see Foulsham, 2015, for a review). The data from these experiments have been used to inform a variety of models of human visual attention. For example, researchers have asked how particular locations are prioritized when searching for an object (e.g., Zelinsky, 2008; Eckstein et al., 2006), and how image features from the periphery may be processed preattentively to drive attention in a bottom-up manner (Itti & Koch, 2001; Peters et al., 2005).

There are many advantages of eyetracking for testing such models. The eyes move frequently, providing behavior which can be directly observed and measured, unlike covert attention. The series of fixated locations can provide both a summary and a time-course of how attention moves over a stimulus. This allows experiments using complex images, and as there is excellent evidence that covert attention moves to a target location prior to the execution of a saccade (e.g., Deubel & Schneider, 1996), it is a reasonable assumption that fixation and attention are routinely coupled. Finally, considerable progress has also been made in understanding the brain mechanisms involved in generating and guiding saccadic eye movements (e.g., McDowell et al., 2008).

Despite these advantages, several researchers have acknowledged the fact that most experimental data on this topic are derived from experiments with static images that are quite different from real scenes (Tatler et al., 2011; Henderson, 2007). This raises issues not dissimilar from those identified by Gibson (1979) and Neisser (1976) in their criticisms of artificial laboratory science. In a typical experiment, participants are shown a series of unrelated images for less than 10s each. This "picture viewing paradigm" is convenient for experimenters, and there is no doubt that it allows researchers to evaluate how different features within a scene are prioritized. By acting as a surrogate for the real world, pictures of natural scenes can be viewed and interpreted like the real environment, and there remain interesting questions about how attention is deployed in this situation. However, Tatler et al. (2011) argue that picture viewing is unlikely to tell us much about natural visual attention, and that a reliance on this paradigm has led to models of attention that over-emphasize pixel and feature-based guidance. Their critique points out that static scenes are typically smaller than our view of the real world, and that they do not replicate

dynamic cues. Crucially, images do not provide a space in which participants can act, and therefore the attentional demands are likely to be very different from actual natural behavior. From an ecological point-of-view, pictures of a scene do not "afford" the same behaviours as the real world itself. Images may also introduce systematic biases due to framing (such as a tendency to look in the centre of an image) and their sudden onset in an experiment (Tatler et al., 2011; Tatler & Vincent, 2009; Foulsham & Underwood, 2008).

The present study tests the validity of picture viewing as a surrogate for natural gaze, in a way that has not been done in previous investigations. We describe results from a new study, comparing participants looking at static images to those who observed the same scenes while walking in the real world in Foulsham et al. (2011). There is a growing amount of research documenting gaze in active, natural tasks including walking, driving, sports and carrying out sequential everyday routines (Land et al., 1999; Land, 2009; Hayhoe & Ballard, 2005). However, there are far fewer investigations comparing behavior in these cases to attention in picture viewing.

In one exception to this, t' Hart et al., (2009) recorded gaze with a mobile eyetracker while 6 participants "behaved naturally" in a range of environments. They then replayed head-centred videos to 8 participants viewing a screen in laboratory conditions. Participants in the lab saw either a continuous video replay of real-world exploration or a selection of static frames in a random order. Analysis focused on comparing fixations in the lab to those in the natural environment, and t'Hart et al. demonstrate a modest correlation, with fixations in the continuous replay condition being a better predictor of gaze in the real world than the static scenes. On the other hand, static

frames (presented for 1s) yielded the highest inter-observer consistency, which was mainly due to a central bias triggered by the rapid onset of the scene. A natural question from this study concerns the relative importance of dynamic presentation and temporal contiguity. It appears from t'Hart et al. that a dynamic presentation is more likely to predict fixations in the real environment. However, because the static frames were presented in a random order the static condition would not have reflected the contiguity and context that is present in the natural environment. Moreover, due to the analysis used by t'Hart et al., it remains unclear how well static presentations—which remain the most commonly used in experiments—can predict fixations over and above general spatial biases.

In Foulsham et al. (2011), we also replayed videos from a head-mounted eyetracker to participants in the lab. The mobile data came from participants walking unconstrained in an outdoor, campus environment. We found both similarities and differences in the way that gaze was distributed when walking in the real world or watching it unfold in the lab. Both conditions showed a central bias (which, in the case of the mobile eyetracking data, indicated that participants tended to fixate near the centre of the head frame-of-reference). Participants walking in the real world spent more time paying attention to the path, and less time looking at pedestrians who were close to the observer, but there were some objects and pedestrians that were fixated equally in both walking and watching. Thus, unlike the results from t'Hart et al. (2009), Foulsham et al. (2011) placed an emphasis on the differences between real-world gaze and fixations on video. However, because we looked at aggregate gaze times on objects of interest and not the pixel-based fixation coordinates at particular times, and because we did not test the

commonly-used static scene presentation in the lab, it is not straightforward to apply the results to attention in scene viewing.

In the present study, static scenes from the environment in Foulsham et al. (2011) were presented to participants in laboratory conditions. An important research approach for cognitive psychologists is to reduce real world behavior to model tasks in the laboratory (see Kingstone et al., 2008). In the domain of attention in scene perception, our analysis aims to identify the benefits and costs of simplifying environmental stimuli to static scenes, as well as to understand whether aspects, such as temporal order, are significant contributors to this process. We evaluated the degree to which the fixation distributions from a number of people looking at a picture could predict the actual gaze location of the person walking in the scene. In addition, we manipulated the order in which the static frames were presented so that they were either sequential, reinstating the temporal contiguity from the original walk, or randomly ordered. The vast majority of picture viewing studies consider isolated images presented in no particular order. However, the video evidence from t'Hart et al. (2009), as well as studies investigating movie watching (Smith et al., 2012) and comic-strip reading (Foulsham et al., 2016), suggest that presenting images in a coherent sequence will change the way that people allocate their attention. We therefore tested the effect of the minimal temporal context of providing images in a coherent order. Because the real world is often temporally predictable, particularly when we move through it at our own pace, a sequential order of static images should provide better conditions for predicting natural gaze.

Method

Mobile eyetracking data

Measurements of natural gaze come from Foulsham et al., (2011), where they are described in full. In brief, fourteen participants (nine females) were recorded during a self-guided walk across the UBC campus in Vancouver, Canada, walking outdoors in a mostly pedestrianized area. Their instructions were to walk from the laboratory to the students' union building to purchase a snack or a beverage, and then to walk back. No other explicit goals were given, and the route was not specified. Most participants took a similar journey, although slight variations (e.g., which side of the road was walked on) and day-to-day differences in the environment (due to weather, presence of vehicles, et cetera) meant that the visual environment was not identical between different participants.

Gaze was recorded from one eye using the ASL MobileEye (Applied Science Laboratories; Bedford, MA, USA), a glasses-mounted mobile eyetracker with a scene camera capturing the environment in front of the wearer. Eye position was sampled at 30Hz and synchronized to the 30fps video of the scene. The result was a set of videos showing the first person view of the walk, along with coordinates indicating the point of regard at each moment in time, relative to the head frame-of-reference.

Participants

Twenty-nine participants took part in the static scene viewing experiment. The participants were recruited from students at the University of British Columbia in Vancouver. All participants volunteered in return for payment or course credit, had

normal or corrected-to-normal vision, and had not taken part in the mobile eyetracking experiment (although they would have been familiar with the local environment featured in the scenes). Participants were pseudorandomly allocated to either the Sequential or Random viewing condition (resulting in n=15 and n=14, respectively).

Stimuli and apparatus

In order to represent a range of content from the real walk, while still collecting enough data for each scene, we selected six participant's videos from the mobile eyetracking experiment. These videos were representative of the others in the data set. Videos with poor lighting, considerable data loss, or variable weather conditions were not included. Videos were approximately 15 minutes in duration, and did not include the part of the original walk where the participant made a purchase from a coffee shop. Each video was further divided into 4 blocks of equal duration, which formed the basis of the viewing blocks in the scene experiment. We extracted a frame from the scene camera videos every 10s (i.e., every 300th frame), from each of the blocks. This interval was chosen so that consecutive frames would not be identical, but would still reflect the temporal progress of the walk.

This resulted in a total of 567 image frames for use in the static scene viewing experiment. Each participant was pseudorandomly assigned to one of the six original walker's videos, such that an approximately equal number of observers saw each video (2-3 observers per condition, per video). During the experiment, each participant saw all of the frames from the assigned video. The exact number of frames varied due to differences in the length of the original walk, but ranged from 70 to 116 images per video.

Frames were presented on a colour monitor at the original resolution (640 x 480 pixels), centred on a grey background, where they subtended 25˚ by 19˚ of visual angle.

Eye movements were recorded using the EyeLink II system (SR Research), which is a head-mounted, video based eyetracker. Eye position was recorded from the pupil image of one eye at 500 Hz, and saccades were detected automatically using velocity (30°/s) and acceleration (8000°/s^2) thresholds. Participants viewed the scenes using a chinrest, which ensured that they were a fixed distance of 60cm from the screen.

Static scene viewing procedure

The procedure is summarized in Figure 1. Participants were asked to freely view the series of images as if they were present in the scene. They were not explicitly informed about the goal of the original walk, although they may have intuited it (because it is a common route with which they would have been familiar). Following calibration of the eyetracker, each scene was presented for 5s, followed by a blank screen for 1s. The presentation of one scene was considered a trial, and prior to each trial a drift correct marker was presented in the centre of the screen, at which point the participant confirmed that they were looking at this point by pressing a key on the keyboard.

--Figure 1 about here--

In the Sequential condition, image frames were presented in the original order, giving 10s interval snapshots of the journey. In the Random condition, frames from the

same video were presented in a randomized order. The experiment was divided into 4 equal blocks (giving a pause every 18-29 trials), and the participants had the option to take a break between blocks, at which point the eyetracker was also recalibrated.

Analysis and results

General eye movement behaviour

The data from the static scene viewing task consisted of more than 19,000 fixations in each condition. Participants made an average of approximately 14 fixations per trial. First, we asked whether there were any overall differences in the eye movements made when viewing static scenes in a coherent versus a random order. Interestingly, there were no differences between sequential and random orders (independent samples t-tests, all ts(27) < 1.4, ps > .19) in either the number of fixations per trial (M = 14.4, 95% CIs [13.4, 15.3]; and M = 14.4, 95% CIs [13.4, 15.4], for sequential and random, respectively), the average fixation duration (M = 311 ms, 95% CIs [296, 325]; M = 298 ms, 95% CIs [285, 312]) or the average saccade amplitude (M = 4.3, 95% CIs [3.8, 4.7]; M = 4.4, 95% CIs [3.9, 4.9]). Thus on a surface level overt attention was deployed similarly in each case.

Comparison of spatial fixation distributions

This study aims to compare where people look in static scenes and the real world. Comparing multiple fixations ("scanpaths") can be complex, particularly if sequential

patterns are important (see Foulsham & Underwood, 2008; Le Meur et al., 2013). Here, we focused on the overall spatial similarity between regions inspected in the different conditions by using fixation density maps (Wooding, 2002) and the area under the Receiver Operating Characteristics (ROC) curve (t'Hart et al., 2009; Ehringer et al., 2009; Tatler et al., 2005). This method has become a standard approach for testing computational models of fixation location, and it has a number of advantages because it makes no assumptions about differences in the underlying distributions. In the present case, this method determines the degree to which one set of fixation locations can be predicted by a spatial model formed from another set.

The process for this analysis is shown in Figure 2. First, a given set of predictor fixations is transformed into a fixation density map by adding Gaussian patches centred on the current fixation location. This creates an "attentional landscape", the peaks of which show the regions of space that were looked at most frequently. The Gaussian patch had a standard deviation of approximately 1˚, and all maps were scaled to a fixed range of 0-1, with 1 indicating the most inspected point in the image. This step therefore transforms the predictor fixations into a continuous spatial distribution.

Then, a set of criterion locations are compared to the fixation density map. If the map is a close match with these criterion locations then most of the locations will coincide with high values and therefore with peaks on the map. The map is thresholded and used as a binary classifier to discriminate the criterion locations from those that are not fixated. For example, a threshold could be chosen which selects only the top 5% of values as those which will be fixated (a "High threshold" example, see Figure 2). The proportion of "hits" is calculated from the number of criterion fixations captured by this

threshold. The proportion of "false alarms" is calculated from the number of non-fixated pixels which are incorrectly included by the threshold. For a good classifier such as that in Figure 2, the above-threshold peaks will successfully capture most of the criterion fixations, leading to a large proportion of hits and a low false alarm rate. When a less stringent level is chosen (e.g., the "Low threshold" in Figure 2 which selects the top 20% of values), more fixations will be captured. However, the higher hit rate will be offset by the higher false alarm rate due to the larger area selected.

Rather than using a single arbitrary threshold, this process is repeated across the full range of possible thresholds. In each case the proportion of hits is compared to the false alarm rate. This results in an ROC curve showing how well the fixations can be predicted by the map. The area under the curve (AUC) provides a measure of the sensitivity of this prediction, ranging from 0 to 1, with a value of 0.5 indicating chance performance. This method provides a robust way to quantify the similarity between looking patterns in sequential and random presentations, as well as whether these can predict looking in the real world. Moreover, because the method essentially relies on ranking the values in the predictor map, it is unaffected by changes in the range or distribution of the map (which are expected with highly variable image content).

All analyses were conducted at the level of the particular image, pooling across all participants who viewed a given scene. We calculated the AUC for each image and report the mean (and 95% CIs) across images, testing between different predictions using related-samples t-tests. The first fixation was constrained to the centre of the screen and overlapped in time with the onset of the scene and so this fixation was excluded, along with any that were outside the image frame. Of the 567 images, 9 were excluded due to

excessive data loss resulting in fewer than 10 fixations being available. Data and code for analysis are available online from the first author's website.

--Figure 2 about here--

Comparing sequential and random presentations

Our first analysis compared attentional landscapes in the sequential and random conditions. The aim here is to determine whether the minimal sequential context on offer altered where participants focused their attention. If so, then looking patterns within the sequential condition should be more similar than those compared between conditions. In contrast, if fixations are mostly determined by the contents of the image, then random and sequential conditions will be similar. We used a "split half" method where we attempted to predict a random subset of half of the fixations on each image using (1) the remaining subset of fixations from the same condition; (2) fixations from the same image but the other condition; or (3) fixations from all other images. Table 1 shows the results of these comparisons.

--Table 1 about here--

All of the AUCs are far greater than 0.5, demonstrating that there is consistency in where people look across conditions and images. However, there are a number of robust differences. We shall consider the three sets of comparisons in turn.

Firstly, the “Same condition” comparison results in a lower mean AUC in the Random condition than in the Sequential condition. This demonstrates that inter-observer consistency was reduced when images were shown in a shuffled order –fixations were literally more random and less likely to cluster together.

Secondly, the mean AUC was reliably lower when fixations were predicted between conditions compared to within a condition, but only within the Sequential condition (where the confidence intervals do not overlap and a paired t-test across images showed a significant difference: $t(555) = 5.2$, $p<.001$, Mean difference = 0.019, 95% CIs on the difference [0.012, 0.026]). Indeed, the only comparison where consistency is notably greater is when fixations are compared within the same condition and with sequential presentation.

Thirdly, the prediction using fixations from other images is weaker in both conditions. This is to be expected, and demonstrates that while general spatial biases can predict fixations better than chance (AUC = ~0.72), where other people have looked in the same image is a better model.

This analysis confirms, in a data driven manner, that there are differences between looking patterns in the Sequential and Random presentations. Figure 3 gives an example of this difference.

--Figure 3 about here--

Comparing static viewing to gaze in the real world

The same comparison method was used to evaluate whether gaze of the original participant who was actually walking in the real world could be predicted by the fixation distributions from static scenes. This analysis relied on moments where there was a valid point of regard from the mobile eyetracker, as well as at least ten fixations from the equivalent static image. This depended on a single participant's data, and due to data loss in the original study (due to blinks, saccades, lighting artefacts and other tracking issues), we included 354 frames of data. Using the AUC, we attempted to predict the real-world gaze location from (1) the fixations made on this image during static viewing in the Sequential condition; (2) the fixations made on *all other images* in the Sequential condition; (3) the fixations made on this image during static viewing in the Random condition; and (4) the fixations made on *all other images* in the Random condition.

The other image models control for image-independent biases (chiefly the central bias) that were present throughout the static scene viewing experiment. Recently, a comparison of spatial biases across existing image datasets has proposed a standard baseline model which can account for the central bias (Clarke & Tatler, 2014). We also implemented the model recommended by Clarke and Tatler (5), which is an anisotropic Gaussian function, centred on the image and with a covariance matrix of $[\sigma^2, 0; 0, v\sigma^2]$, where σ^2 (the variance across the horizontal axis) is 0.23 and v (which scales the variance in the vertical axis) is 0.45. Of course, such models are designed to account for fixations on 2D static images presented on computer screens. Spatial biases relative to head direction are also a factor in real-world viewing but there have been few attempts to characterize them and it is possible that they might differ from screen-based data. Finally,

therefore, we tried to predict the real-world gaze location from the empirically-observed distribution of real-world gaze from all the other frames (6). Figure 4 shows the baseline model alongside the empirically-observed distributions from the static conditions and the real-world gaze points. Because these distributions are independent of any particular image, they should be outperformed by models capturing information about the important features of a scene.

--Figure 4 about here--

Table 2 shows the results of all six model comparisons. Although the average AUCs are all greater than 0.5, this predictive power could come from image-independent biases as well as information about attentional priority in a particular scene. There is a modest difference between the prediction from fixations on the same image and those pooled across other images, in both Sequential and Random conditions. This difference was statistically reliable across images (Sequential: $t(353)=3.1$, $p=.002$, 95% CIs on the difference [0.01, 0.06]; Random: $t(353)=2.6$, $p=.009$, 95% CIs[.01, .05]). Predicting gaze from the Sequential condition was not significantly better than doing so from the Random condition ($t(353)=1.7$, $p=.097$, [-0.003, 0.04]).

--Table 2 about here--

Critically, these values should be interpreted relative to the performance of the two baseline models, which contain no information about the particular image being inspected. Surprisingly, the predictions from static viewing failed to reliably outperform the formula-derived baseline central model ($t(353) = 1.09$, $p=.27$ and $t(353) = 0.212$, $p=.83$, for Sequential and Random conditions, respectively). Moreover, the empirical central model, which captures the overall tendency for walking participants to look in a particular location, is far better than any model generated from the static scene data. Because we only have gaze from a single participant in each real world scene, it is not possible to calculate inter-observer consistency in the same way as within the static scene data. However, it is clear that the predictive power of the data from static scene viewing is worse than we would expect given the results in Table 1 and the baseline models.

Figure 5 gives two examples of the gaze location observed in the real world and the distribution provided from Sequential static scene viewing. In some cases (e.g., Figure 5, left column), the gaze behaviour in lab viewing is a close match with where the actual walker was looking. In other cases, there is a large disparity (e.g., Figure 5, right column).

--Figure 5 about here--

In scene viewing experiments, it has been suggested that attentional priorities may change over time and thus that the initial fixations in a scene may be guided in a different way from those later in viewing (Parkhurst et al., 2002; Tatler et al., 2005; Unema et al., 2005). For example, it might be the case that early fixations are drawn towards the most salient items, or that after the first few seconds observers begin to inspect less salient regions. If this is the case then the first few fixations on a static scene might be a better predictor of gaze in the real world than those fixations drawn from a longer viewing period. This was probably particularly true in the present study, where we forced people to look at a static frame for several seconds, when in reality the scene would have changed more quickly. We therefore repeated the comparisons between static and real-world gaze, using only either the first 3 fixations on the scene from each lab participant or the last 3 fixations. In both conditions, early fixations produced a better prediction than late fixations. In Sequential viewing, the mean AUC was 0.691 (95% CIs [0.663, 0.719]) for the first three fixations and 0.631 [0.602, 0.659] for the last three. In Random viewing the mean AUC was 0.659 [0.630, 0.688] for early fixations and 0.634 [0.606, 0.663] for late fixations. However, there was no substantial improvement over predictions based on the whole trial, and in no case did the AUC approach the image-independent empirical central model.

Discussion

Attention experiments in general, and eye movement studies in particular, have often examined how individuals process 2-D static images presented on a screen. The present investigation tested the extent that data collected in such a manner predicts the

way people explore their visual environments in the real world. The match in behaviour between the real environment and pictorial stimuli is a specific example of the worries about ecological validity in perception and cognition that were discussed by Gibson (1979) and Neisser (1976). Importantly, we do not consider such issues to be all-or-nothing (e.g., Kingstone et al. 2008). It may well be that some aspects of attentional selection are preserved when showing participants pictures of the real world, while others are not. Our results can also shed light on some of the properties of stimuli and experiments that are important for the relationship between lab and life.

Our study utilised the following novel procedure. First, we collected real world eye movement behaviour as individuals walked on a university campus (Foulsham et al., 2011). Second, we systematically extracted static images of those walks at 10 second intervals and showed them to new participants in either sequential or random order. We found that looking at images shown in sequence yielded eye movements that were more similar than looking at images ordered randomly, demonstrating that viewing behaviour was sensitive to the temporal coherency depicted across the static images. There was a small but detectable difference in where people looked in the two conditions. More critically, however, eye movements for the static images shown in sequence or randomly did not differ in their ability to predict where people were looking when those scenes were actually encountered in the real world, and neither performed better than a formula-derived model that assumes that most fixations are directed centrally. By far the best predictor of where people were looking for those individual real world scenes was provided by taking where the eye was normally positioned in the head during a walk in real life. In other words, eye movements to static scenes that factor in the actual

information that one could look at do not do any better than a model that simply assumes people look centrally, and all of these do worse than a model that considers simply where the eye tends to be directed in real life.

Our study replicates some aspects of the work by Foulsham et al., (2011) where, using the same mobile eyetracking data, we found reliable differences between real-world gaze and the things looked at while watching a video. The empirical, real-world distribution in Figure 4 mirrors the more general pattern demonstrated throughout the walk and reported in Foulsham et al. (2011), with gaze being more frequent along the vertical midline. However, in that study we did not investigate the specific locations in the scene that were fixated, and we used a contiguous video presentation meaning that the results might not extend to static scene viewing.

The present study is perhaps most similar to t'Hart et al., (2009), who also compared mobile gaze to viewers in the lab. In that study, lab participants either watched a continuous video replay of the environment or they saw a randomly-ordered sequence of still frames for 1 second each. Their conclusion was that a better prediction of real-world gaze could be made if the continuous replay was used. However, these conditions differed in the fact that one was both a dynamic scene and a stimulus that preserved temporal continguity.

Our first set of findings address whether sequential order alone might improve the model of real-world gaze provided by static scenes. Using a data-driven approach, we observed that looking patterns in randomly ordered scenes were different from those in sequentially ordered scenes, as well as having lower inter-observer consistency. This is an example of the context of an image changing the distribution of attention, despite the

fact that the very same visual information was present in each case. Obviously, any bottom-up or feature-based model is not going to be able to predict such differences (e.g., raw saliency models; Itti & Koch, 2001). Presumably, observers are making eye movements that reflect predictions or expectations about what is going to occur in the future based on what was observed in the past. In other words, eye movements to static visual images are sensitive to inter-image coherency and narrative. The same conclusion was reached in a recent study of fixations in comic-strip reading (Foulsham et al., 2016), and future research needs to determine what sorts of expectations affect attention and how. For example, in the present study, even though participants in the sequential condition did not know exactly what they were going to see next, they may have been more inclined to think about where the participant was moving and thus to look at things in the distance that might appear in the next trial. It should be stressed that the temporal context in this case is really quite minimal, but it might be one way for subsequent studies to increase the realism of their procedure. t'Hart et al (2009) actually reported that inter-observer consistency was higher in randomly-presented static frames than in a continuous replay, and they suggested that this was largely due to a central bias exacerbated by the sudden onset of each frame. Critically, the present results go beyond that by demonstrating that even within static presentation, consistency depends on the sequential context. While some of this consistency is due to spatial biases, both conditions were presented in the same fashion and our other image control comparisons demonstrated that information from a particular image is guiding fixations. However, this guidance seems to be less idiosyncratic when scenes are presented in context. It should also be noted that because there were minor differences in the videos from individual

walking observers, we are not able to properly assess inter-observer consistency in the real world. It would be useful to do this in future studies (for example by using a more controlled environment), since normative models of attention can only ever capture what is common amongst individual observers.

While eye movements to static images were better than chance at predicting where people will look when those scenes are encountered in real life, even sequentially ordered scenes were no better than a formula based model that assumes that people generally look centrally. This is problematic for the many studies that wish to make conclusions about real world attention from static scene scanning, and it validates a central claim of Tatler et al., (2011) who criticised the picture-viewing paradigm. This key finding also differs from t'Hart et al., (2009) who found that fixations in static frames (and particularly in continuous movies) could predict gaze in the real world better than spatial biases alone. However the data in that study came from only 4 lab observers and the empirical model based on lab participants was not much better than chance (mean AUCs between 55% and 63%, but using a different method to the present study). Here, we explicitly modelled the empirical distributions in both scene and real-world, and we also used the recently proposed formula-derived distribution from Clarke and Tatler (2014). In our study, there was only weak evidence that gaze-worthy features in static scenes reflected where people looked in real life. The numerical advantage for predicting real-world gaze from sequential rather than random presentation is overshadowed by the fact that a much better prediction can be made by assuming a general central distribution.

Why do participants not select the same details when looking at a pictorial representation of the real world as opposed to gazing in the true environment? Of course,

while the method in our static scene experiment was not unusual for this topic (e.g., Foulsham, 2015; Foulsham & Underwood, 2008; Parkhurst et al, 2002), there were many differences between stimuli and procedure compared to the real world walking data. These differences included the smaller visual angle of the scenes and their lack of motion. t'Hart et al., (2009) included a video condition in their study, finding that this condition led to better predictions than presenting static scenes. We would therefore expect improved AUC scores if we compared the same (sequential) frames but in a continuous dynamic context. However, in Foulsham et al. (2011), even with a video condition, there were differences in gaze targets between laboratory and the real world.

Another important difference between the conditions concerns the different task demands evoked by simply looking at an image as opposed to walking in a scene. As in Foulsham et al., (2011), it is likely that some of the poor predictability arose from participants in the laboratory not fixating features which were important for the active task of locomotion through an environment (such as the path, which is not typically informative or salient for participants merely looking at an image).

The role of task constraints and their match with particular visual information in the world ("affordances") deserves further attention. We did not attempt to closely match the task in the laboratory with that in the real scene. In future work it would be useful to experiment with instructions that might prime participants to pay attention to consistent features. Immersive displays and virtual reality could also generate conditions where the laboratory observer feels more of an active presence. Although additional work is necessary to address the role of particular route-finding goals, the present study did raise some interesting possibilities. For example, with a sequential set of images, participants

receive a minimal amount of information about route and movement through space, and we have observed examples of this changing gaze patterns (e.g., Figure 3). Context, in combination with task, can shift behaviour. Understanding this dynamic is an exciting future research avenue which promises to inform how we make inferences from artificial test conditions to natural behaviour.

It is clear that there are particular visual cues relating to locomotion and personal safety which are more important in a real-world scenario. It is interesting to note that some of these features are going to be important across a range of tasks, and might be visually "salient", prompting them to be attended to regardless of the current goals (such as a rapidly moving car). Others features are not salient to those looking at an image, but the only way that we can identify these is through experiments in the real world. On the one hand, feature-based models could be devised which prioritise important regions such as the path (in the same way that models of search can be built to predict biased attention towards targets with particular features; Zelinsky, 2008; Navalpakkam & Itti, 2005). On the other hand, the present results demonstrate that this enterprise requires a much better understanding of natural gaze and the "targets" that govern its allocation, and that this cannot be accomplished solely from experiments with static scenes.

The current results demonstrate that central biases, both on a screen and within real life, can potentially tell us just as much about where someone will look as can an analysis of the features in the image. We are not the first to point this out, and research has begun to describe systematic biases and their causes during scene viewing in laboratory conditions (Tatler & Vincent, 2009; Foulsham & Underwood, 2008; Clarke & Tatler, 2014; Tseng et al., 2008). Clarke and Tatler (2014) have recently argued that

choosing an accurate central baseline is particularly important when assessing the performance of different models for predicting fixations in scenes. Using the best average model proposed by Clarke and Tatler, we confirmed that it could predict a considerable amount of the variation in fixation position, even in real-world fixations.

However, the shape and interpretation of the central bias in mobile eyetracking is less well understood, and this provides a real opportunity for subsequent research. Although the current dataset comprised a relatively small amount of natural gaze data, the empirically-observed distribution (see Figure 4) is similar to that reported in Foulsham et al., (2011) where continuous clips were drawn from the same experiment. In both cases, participants tended to look above the horizontal midline, showing greater variation vertically than they do horizontally, which is quite different from the typical pattern that is found when viewing comparable images on a screen. Part of this pattern is probably due to the tendency for participants to look down while moving forwards through the environment, a scanning routine considered in more detail in t'Hart and Einhauser (2014). It is critical for those seeking to model patterns in gaze to consider these regularities, and how they should be related to image features. In the present study, the vertical spread of gaze should, perhaps, be seen as an embodied feature of the task (walking) rather than the result of targeting particular visual features. Conceptualising fixations in this way requires a good understanding of the functions of gaze in a particular task (Foulsham, 2015).

In mobile eye tracking, spatial biases also reflect the position of the head in the world, which tends to be angled slightly below the horizon (Foulsham et al., 2011). Indeed, the difference between eye movement sampling in the lab and gaze in the real

world is also partly due to the head-constrained conditions adopted in most experiments, where the field of view is pre-defined and fixed. In contrast, in the real world, participants seem to select more often by changing the field of view with head movements. Of course, if all selection was accomplished by the head (such that selected items were in the centre of the scene camera), there would have been no need for observers in the laboratory to move their eyes at all. It therefore seems likely that a difference in orienting with the head and the eye is a major determinant of the poor predictability in the present study. To our knowledge, nobody has compared head sampling in the laboratory and the real world (by capturing a wider view of the environment), but this would be a fruitful line of enquiry. The coordination of eye and head movements (see Freedman, 2008), as well as potential differences in their cognitive consequences (Solman & Kingstone, 2014; Solman, Foulsham & Kingstone, under review), are important to consider when using eyetracking in images to study attention in the real world.

Critically, the general empirically-observed distribution of real-world gaze proved much better at predicting the gaze location on individual frames than any other model for fixation. In a mobile eye tracking context, “centre” means head-centred. So simply assuming that gaze is pointed in the middle of a scene captured by a head camera is as good at predicting where people really look as actually showing them those images on a computer and tracking fixations. And if one simply includes where the eye is usually positioned in the head, one does better than any of those other methods. This means that one can do quite well predicting where people are looking by just looking at the head camera data and assuming that the eye is positioned in the centre. And one could do even

better than that by including where the eye is usually positioned in the head, and not being concerned at all about the content of any of the scenes.

We began our study by discussing the work of Tatler et al. (2011) who provided the sobering warning that looking behaviour in computer images may not be especially predictive of looking behaviour in real everyday life. The present study provides a test of that proposal and validates it. Recently, a series of investigations have converged on the conclusion that the way that we pay attention to other people is very different when those people are images on a computer screen, versus people in real life (Foulsham et al., 2011; Risko et al., 2012). The key difference in this case seems to be that real people can see where you are looking, providing a true communicative context, whereas images on a computer screen cannot look back (Risko, Richardson & Kingstone, 2016). It is possible that this effect was also present in the current study (see Figure 5, for example, where the observer in the real world is avoiding looking at other pedestrians). The context (laboratory versus real-world) changes the meaning of social stimuli, which in turn transforms the allocation of attention. The present study extends this principle to scene viewing in general, demonstrating that isolated static scenes do not stimulate attention in the same way as a real world environment.

Acknowledgements

We thank Esther Walker and Nicola Anderson for research assistance during the early stages of this project.

References

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*(23), 2577-2587.

Clarke, A. D., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41-51.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827-1837.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973-980.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, *17*(6-7), 945-978.

Foulsham (2015). Scene perception. In J. Fawcett, E. F. Risko, & A. Kingstone (Eds.), *The handbook of attention* (pp. 257–280). Cambridge, MA: MIT Press.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2), 6. doi:10.1167/8.2.6

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920-1931.

Foulsham, T., Wybrow, D., & Cohn, N. (2016). Reading Without Words: Eye Movements in the Comprehension of Comic Strips. *Applied Cognitive Psychology, 30*, 566-579.

Freedman, E. G. (2008). Coordination of the eyes and head during visual orienting. *Experimental Brain Research*, *190*(4), 369-387.

Gibson, J.J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188-194.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219-222.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194-203.

Kingstone, A., Smilek, D. & Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology, 99*, 317 - 345.

Kingstone, A., Smilek, D., Ristic, J., Friesen, C. K. & Eastwood, J. D. (2003). Attention, researchers! It's time to pay attention to the real world. *Current Directions in Psychological Science*, *12*, 176-180.

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

Land, M. F. (2009). Vision, eye movements, and natural behavior. *Visual Neuroscience*, *26*(01), 51-62.

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, *45*(1), 251-266.

McDowell, J. E., Dyckman, K. A., Austin, B. P., & Clementz, B. A. (2008). Neurophysiology and neuroanatomy of reflexive and volitional saccades: evidence from studies of humans. *Brain and cognition*, *68*(3), 255-270.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205-231.

Neisser, U. (1976). *Cognition and Reality: Principles and implications of cognitive psychology*. New York: Freeman.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107-123.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*(18), 2397-2416.

Risko, E. F., Laidlaw, K., Freeth, M., Foulsham, T., & Kingstone, A. (2012). Social attention with real versus reel stimuli: toward an empirical approach to concerns about ecological validity. *Frontiers in Human Neuroscience*, *6*, 143-143. doi: 10.3389/fnhum.2012.00143

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the Fourth Wall of Cognitive Science Real-World Social Attention and the Dual Function of Gaze. *Current Directions in Psychological Science*, *25*(1), 70-74.

Smith, T. J., Levin, D., & Cutting, J. E. (2012). A Window on Reality Perceiving Edited Moving Images. *Current Directions in Psychological Science*, *21*(2), 107-113.

Solman, G. J., & Kingstone, A. (2014). Balancing energetic and cognitive resources: Memory use during search depends on the orienting effector. *Cognition*, *132*(3), 443-454.

't Hart, B.M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., Koenig, P., & Einhäuser, W. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition*, *17*(6-7), 1132-1158.

’t Hart, B.M., & Einhäuser, W. (2012). Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research*, *223*(2), 233-249.

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.

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

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*(5), 643-659.

Tseng, P. H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7), 4. doi:10.1167/9.7.4.

Unema, P. J., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, *12*(3), 473-494.

Wooding, D. S. (2002). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 518-528.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787.

	Predictor		
	Same condition	Other condition	Other images
Sequential	0.826 [0.820, 0.833]	0.808 [0.801, 0.815]	0.717 [0.707, 0.725]
Random	0.808 [0.799, 0.817]	0.806 [0.798, 0.814]	0.722 [0.714, 0.731]

Table 1. Mean AUC values (with 95% CIs across images) comparing within and between conditions during static viewing.

Predictor	AUC across images	
	Mean	95% CI
(1) Sequential fixations	0.689	[0.660, 0.719]
(2) Sequential other image	0.652	[0.622, 0.683]
(3) Random fixations	0.672	[0.642, 0.703]
(4) Random other image	0.643	[0.612, 0.674]
(5) Baseline central (Clarke & Tatler, 2014)	0.675	[0.648, 0.702]
(6) Empirical central	0.736	[0.711, 0.761]

Table 2. Performance across frames when predicting the real-world gaze location on the basis of static scene viewing and the baseline models.

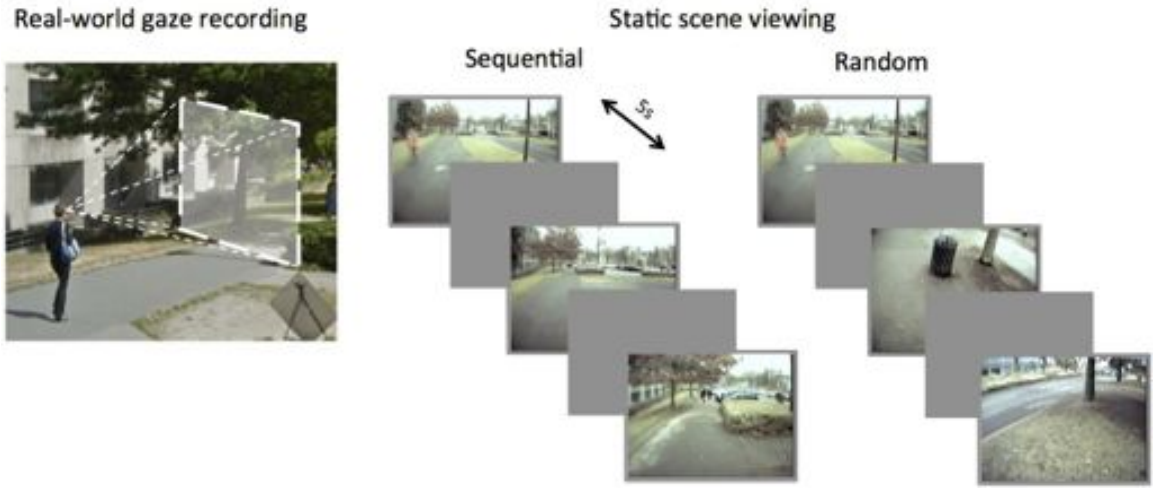


Figure 1. The procedure presented video frames from video recorded in the real world (left). These frames were presented one at a time in either the original sequence or a jumbled order (right).

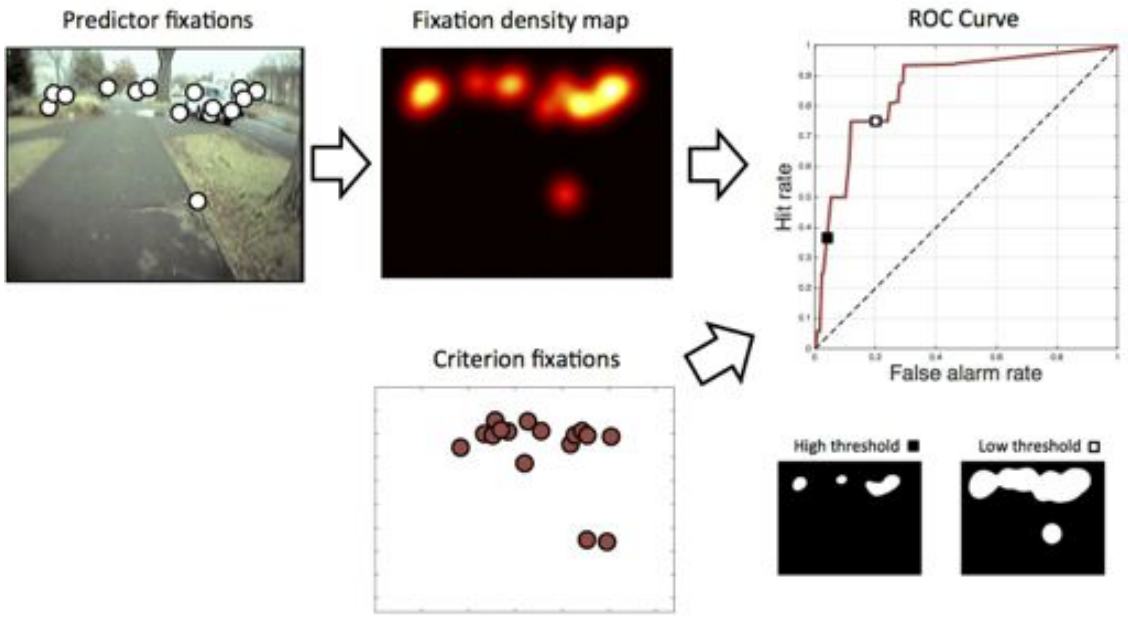


Figure 2. The analysis used to compare fixations. A fixation density map was generated, with high values (hotspots) indicating locations that were frequently fixated. This map was then used to classify a set of criterion fixations, by thresholding the map at progressively higher values. For each threshold, the map is evaluated by determining how many fixations it captures (hits) compared to non-fixated regions covered (false alarms; see lower right panels). Repeating for many thresholds gives an ROC curve depicting how well the map can discriminate criterion fixations from non-fixated locations. In this case, the area under the ROC curve (0.86) indicates that the map is a good predictor.

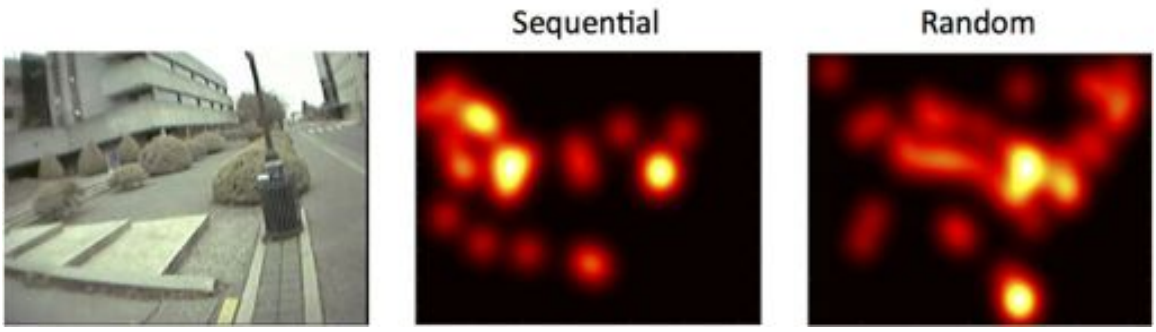


Figure 3. An example scene from the experiment, with fixation distributions from each condition. In this case observers in the Sequential condition spent more time looking at the left of the image than people in the Random condition, perhaps because this was where the walker was going to move next.

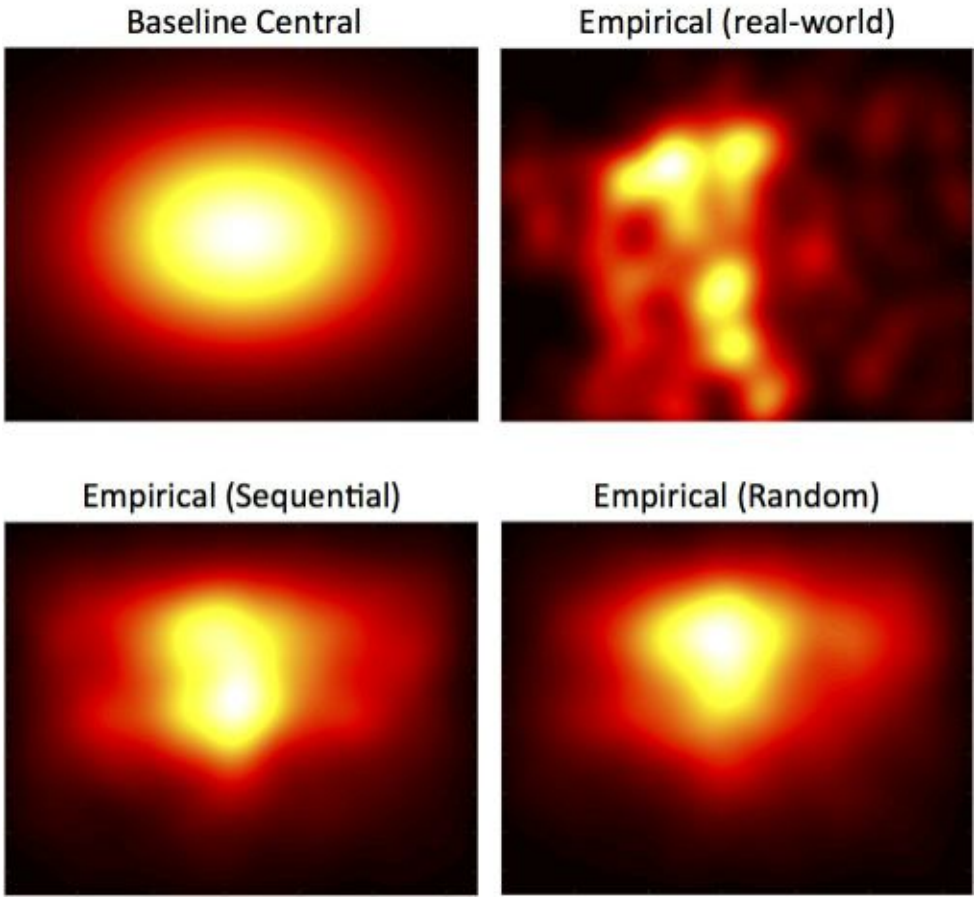


Figure 4. Image-independent fixation distributions, with brighter values indicating higher fixation density. The Baseline Central distribution was defined by a formula, while Empirical distributions were derived from all the real-world and static scene fixation data.

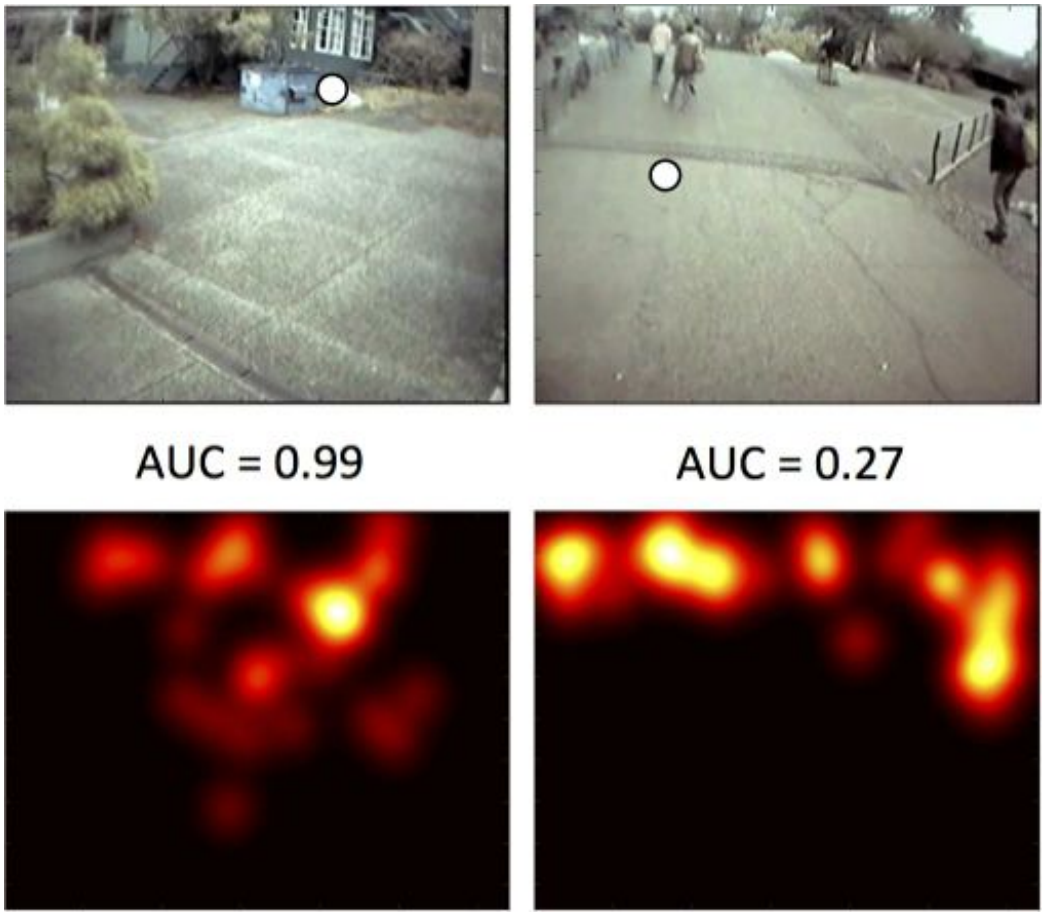
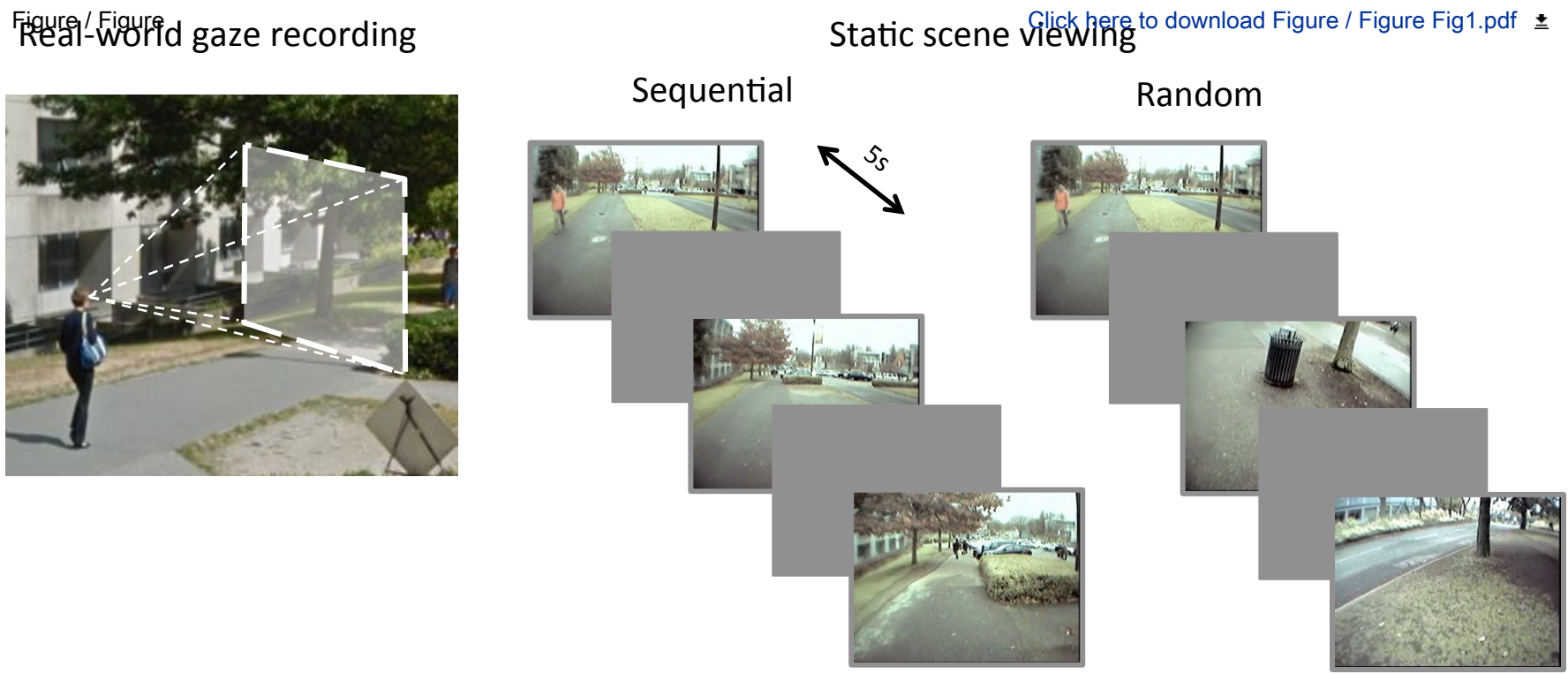
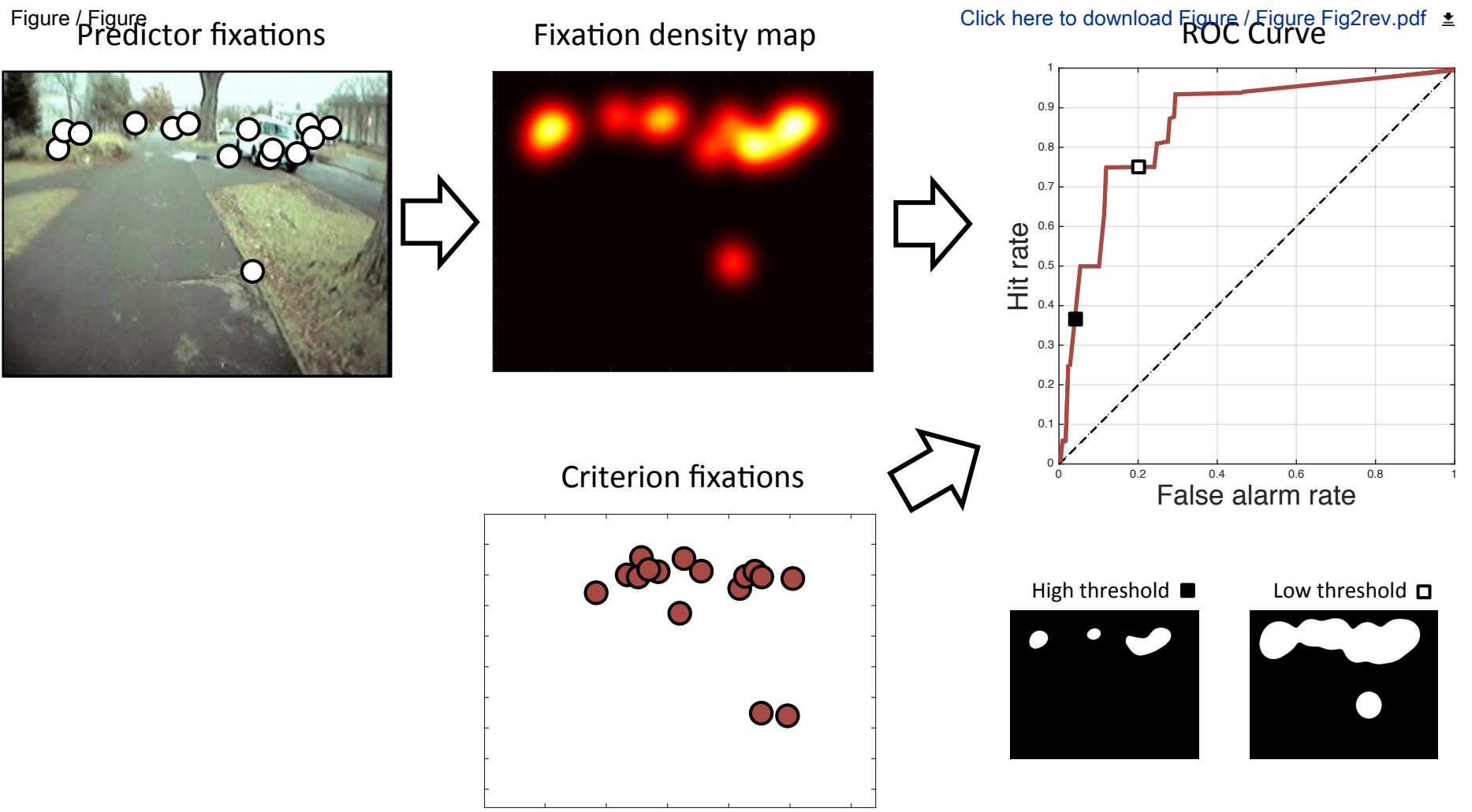
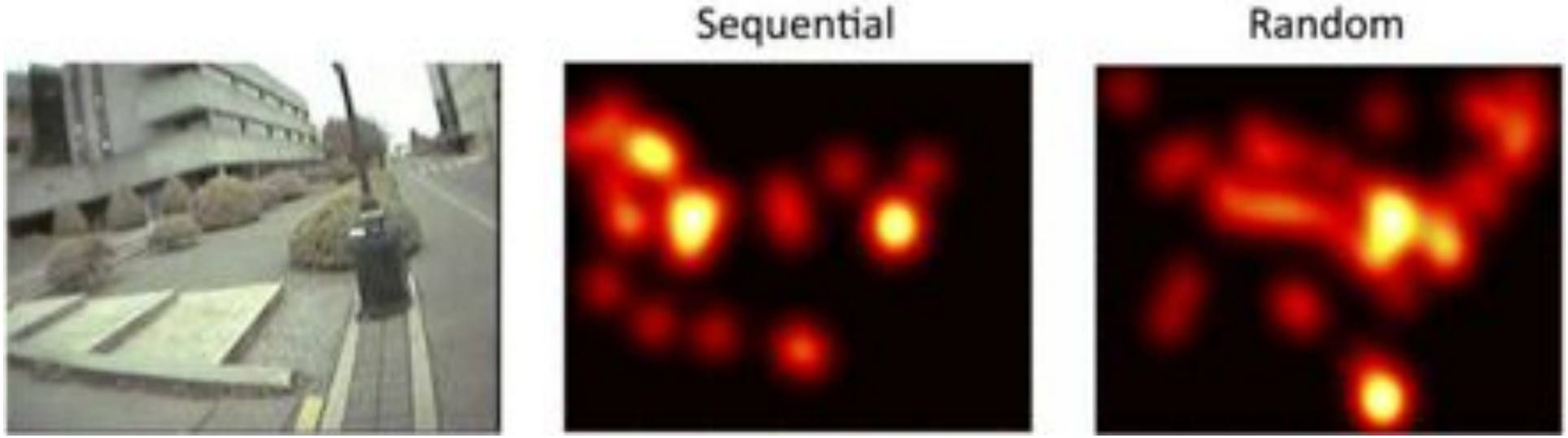
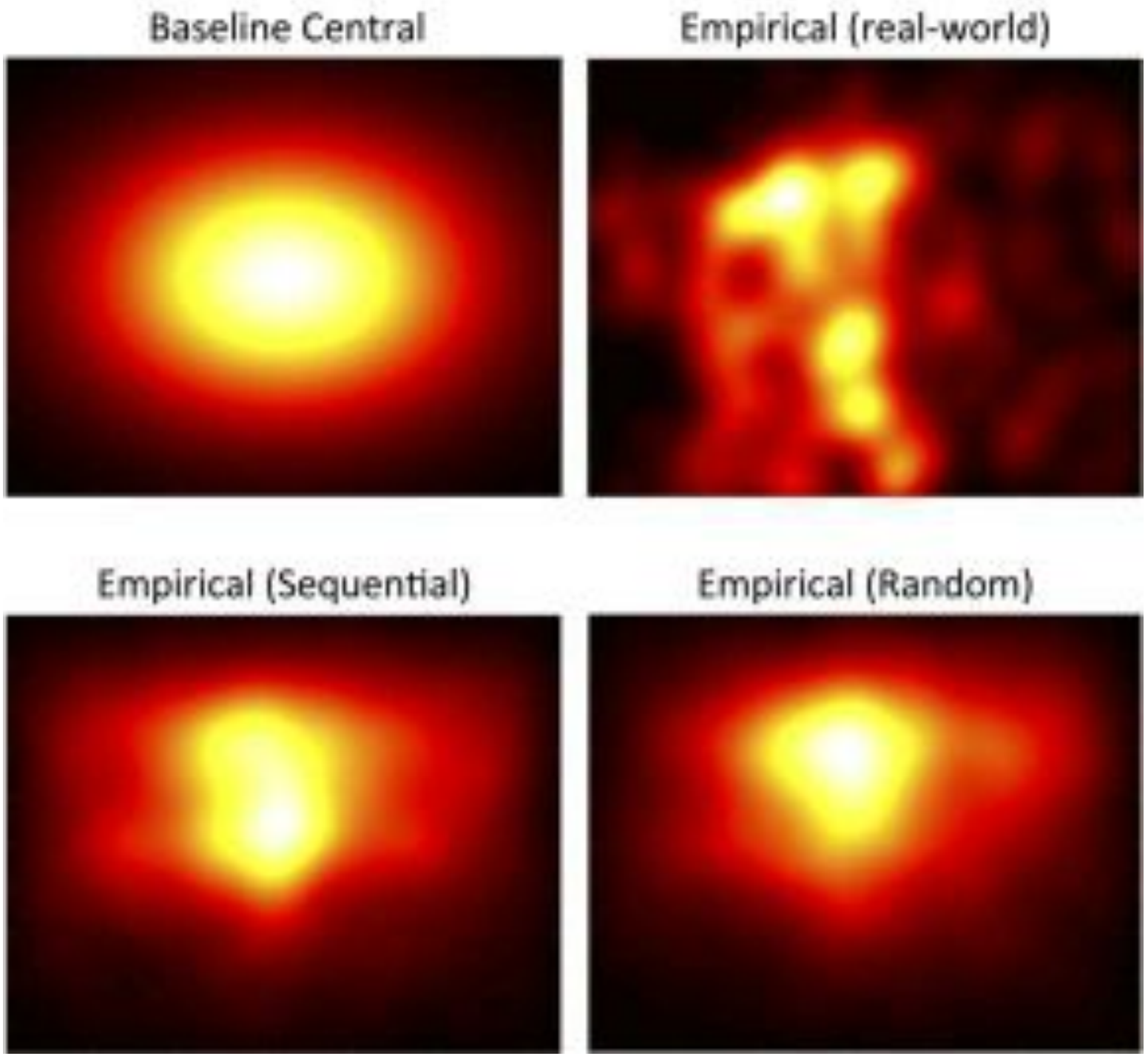


Figure 5. Examples of predicting the real world point of gaze (top) with fixations from static scene viewing (bottom, in this case from the Sequential condition). The computed AUC value is shown in each case.









AUC = 0.99

AUC = 0.27