Atten Percept Psychophys (2017) 79:1415–1425
DOI 10.3758/s13414-017-1324-7

Human classifier: Observers can deduce task solely from eye movements

Brett Bahle[1] · **Mark Mills**[2] · **Michael D. Dodd**[2]

Published online: 10 May 2017

Abstract Computer classifiers have been successful at classifying various tasks using eye movement statistics. However, the question of human classification of task from eye movements has rarely been studied. Across two experiments, we examined whether humans could classify task based solely on the eye movements of other individuals. In Experiment 1, human classifiers were shown one of three sets of eye movements: Fixations, which were displayed as blue circles, with larger circles meaning longer fixation durations; Scanpaths, which were displayed as yellow arrows; and Videos, in which a neon green dot moved around the screen. There was an additional Scene manipulation in which eye movement properties were displayed either on the original scene where the task (Search, Memory, or Rating) was performed or on a black background in which no scene information was available. Experiment 2 used similar methods but only displayed Fixations and Videos with the same Scene manipulation. The results of both experiments showed successful classification of Search. Interestingly, Search was best classified in the absence of the original scene, particularly in the Fixation condition. Memory also was classified above chance with the strongest classification occurring with Videos in the presence of the scene. Additional analyses on the pattern of correct responses in these two conditions demonstrated which eye movement properties successful classifiers were using. These findings demonstrate conditions under which humans can extract information from eye movement characteristics in addition to providing insight into the relative success/failure of previous computer classifiers.

Keywords Categorization · Visual search · Eye movements · Cognitive

A primary function of the oculomotor system is to move the eyes around our visual world in order to place relevant objects into the fovea, the area of the eye with the highest resolution. To achieve this goal, the oculomotor system performs fixations and saccades. Fixations are short periods of time (200 to 300 ms) when the eyes minimize movement and focus on a relatively specific region in visual space to gain higher acuity. Saccades are when the eyes move rapidly between fixation locations.

Eye movements are controlled by a vast array of processes, including, but not limited to, low-level image salience (Itti & Koch, 2000; Wolfe & Horowitz, 2004) and higher-level expectation and knowledge (Henderson, Malcolm, & Schandl, 2009; Hollingworth, 2009). Previous research has demonstrated that task goals heavily influence where the eyes go (Einhauser, Rutishauser, & Koch, 2008), dating back to Yarbus (1967). More recent research has shown that even general tasks, such as searching for a contextually unrelated target in a natural scene (e.g., finding a small target letter embedded in scene), leads to differences in eye movement and fixation behavior compared with other tasks, such as memorization (Dodd, Van der Stigchel, & Hollingworth, 2009; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011). These differences have led researchers to ask whether task could be deduced solely from eye movements: that is, knowing where the eyes have been, can it be determined what the eyes were doing? In the present study, we investigate this issue by seeing if human observers can

✉ Brett Bahle
brett-bahle@uiowa.edu

[1] Department of Psychological and Brain Sciences, University of Iowa, W311 Seashore Hall, Iowa City, IA, USA

[2] University of Nebraska, Lincoln, USA

correctly classify eye movement patterns into one of three different tasks.

Broadly, the interpretation of eye gaze is critical across a number of fields, such as developmental psychology and psycholinguistics. Developmentally, interpreting another individual's gaze is important in social interactions. For example, children learn over time that an individual's gaze can be informative regarding their motives (Freire, Eskritt, & Lee, 2004). Similarly, in psycholinguistics, eye gaze acts as a paralinguistic cue, adding information to the meaning of an utterance. For example, when conveying a sarcastic response, characteristic eye movement patterns (such as eye rolls) are observed (Caucci & Kreuz, 2012).

In other fields, researchers have investigated whether eye movement patterns performed during a task, such as visual search, can be classified by computer systems. This has proven possible under certain conditions (Borji & Itti, 2014; Borji, Lennartz, & Pomplun; 2015; Haji-Abolhassani & Clark, 2014; Henderson et al., 2013). For example, Borji and Itti (2014) measured eye movements as subjects performed Yarbus' seven tasks on 15 natural scenes. The Yarbus instructions emphasized both general (e.g., free examination or memorization of a scene) and more specific (e.g., estimating properties of the scene such as the material wealth or age of the individuals pictured) task goals. They then trained a computer classifier to classify task from eye movements. Classification rate was successful in this task (24.21%, where chance was 14.29%). Similarly, Henderson et al., (2013) had subjects perform four tasks: natural scene memorization, text reading (of news reports), visual search for a "L" or "T" in a natural scene, and a pseudo-reading task where subjects scanned "text-like" stimuli with letters replaced by small blocks and were instructed to scan the "text" as if they were reading. A computer classifier was subsequently trained and, across all subjects, all four tasks were correctly classified above chance (all above 50%, with chance 25%).

Moreover, since Yarbus' seminal paper (1967), it has widely been assumed that eye movements differ between tasks (i.e., Yarbus' subjects' eye movements look different to us). However, whether observers notice these differences and have the ability to interpret between tasks has received little attention. Recently, Foulsham and Kingstone (2013) have demonstrated that human observers can classify eye movements as their own fixations. Subjects first memorized a set of images for 2.5 seconds each and then were shown their own fixations against either a random set of fixations, fixations from a different subject, or their own fixations from a different image. Subjects correctly identified their own fixations above chance in all conditions in a two-alternative forced choice task, implying that human observers can deduce where they have looked in retrospect. In evaluating their findings, the above chance ratings in the random compared with own fixation condition may have simply been a result of the task given that random fixations are less likely to fall on locations in a scene a real observer would typically fixate (e.g., objects). However, comparing fixations from a different observer to one's own fixations are a particularly strong test of eye movement awareness and leads to the suggestion that people may have a general sense of which specific objects/locations they did or did not look at.

One recent study has reported correct classification by human observers for a category classification task. Zelinsky, Peng, and Samaras (2013) had searchers look for either a teddy bear or a butterfly among distractors that varied in similarity to the target items. "Decoders" were then shown the searchers' scanpaths and asked to determine, using a confidence scale, what object the searchers were attempting to identify. Decoders were successful at this task. However, they received substantial training regarding target categories and what the displayed fixation information meant. This instruction included telling decoders to look at both the first fixated object and the object fixated the longest, properties previous analyses had shown were critical to defining the differences between tasks. This evidence strongly suggested that aspects of eye movement classification can be performed by a human observer, at least following some form of training.

One other recent study investigated the question of human classification with a different approach. Greene, Liu, and Wolfe (2011) published results questioning whether the classic Yarbus tasks could be classified according to eye movements. First, participants were shown images similar to Yarbus (e.g., at least two people present in a scene doing various real-world tasks) and instructed to do one of four tasks during viewing: memorize the picture, determine the decade in which the picture was taken, determine the age of the people in the picture, and determine the wealth of the people in the picture. A linear classifier was tested on the collected data, and the results showed an inability to classify correctly. Although further research has questioned this result by noting both a lack of variable classifying techniques and a lack of descriptive tasks—in addition to showing a positive classification for Yarbus' tasks (Borji & Itti, 2014; Haji-Albolhassani & Clarke, 2014)—the key result from Greene, Liu, and Wolfe (2011) as it relates to the present work comes from their final experiment. Given that humans are the best pattern classifiers, they attempted to determine whether humans could succeed where their computer classifier failed. New human observers looked at the same images from the original data collection, now with scanpaths (displayed as lines) and fixations (displayed as dots) from old subjects laid over the scene. Subjects were shown 100 images of previous subjects' eye movements and were asked to classify them into the four tasks described above, with no feedback given. Subjects were at chance for predicting task in this experiment.

One critical difference between this human classification task and computer classifiers is that the humans saw eye

movements with the additional context of the original viewed scene. When computers classify, they usually classify task without explicit reference to the image being viewed (or vice versa). For example, Greene, Liu, and Wolfe (2011) trained three separate classifiers—one to classify based on image, one based on subject, and one based on task. When they showed human observers scanpaths laid over an image, they coupled the image and task information. This coupling may have had the effect of making the classification more difficult because of possibly irrelevant scene information. Thus, in the present study, we include a condition in which eye movements are overlaid on a black background, in addition to a condition in which they are overlaid on the original image. If the eye movement metrics themselves are what is important for classification, human classification may improve for some tasks when the background scene is absent compared with when it is present. This manipulation was one of our primary motivations for the present study. By assessing performance across conditions in which this coupling is present compared with when it is absent, we can add insight into the relative importance of scene background information for classifying certain tasks. Some tasks may aid from the coupling, and this could explain why classic computer classifiers have failed in the absence of image statistics. Other tasks may be harmed by the coupling, and this could explain why the human classification of Greene, Liu, and Wolfe (2011) was unsuccessful.

A final notable feature of the tasks often used by classifiers is that the tasks often bias locations to specific scene regions. For example, a search task in which one searches for a blender would bias eye movements to locations where blenders typically appear (e.g., countertops). This bias could present a problem for classification, because tasks dependent upon context information tend to be more similar than those that do not: a search task in which one looks for a blender would look quite similar to a memory task in which one remembers items on a countertop. Importantly, in other tasks (Dodd, Van der Stigchel, & Hollingworth, 2009), fixation locations are not biased by the task instruction: a search task in which the target could be either an "N" or "Z" located anywhere in the scene with equal probability is not limited to certain scene regions. In the present study, we use the eye movement data from Dodd, Van der Stigchel, and Hollingworth (2009) to determine whether human observers can classify task when the task does not contextually restrict eye movements towards particular locations of the scene. If under these conditions, humans classify at above chance rates, it suggests they may have intuitive knowledge of how the tasks used in the present study are performed.

Experiment 1

Observers in Experiment 1 were shown three separate sets of eye movements metrics, extracted from Dodd et al. (2009): *Fixations*, which were displayed as blue circles, with larger circles meaning longer fixation durations; *Scanpaths*, which were displayed as yellow arrows, giving an order to viewing; and *Videos*, in which a neon green dot moved around the screen, signaling where exactly a subject looked, giving viewers all eye movement information. In addition, observers were shown eye movements in two different sets—one in which the eye movements were laid over the scene the original task performers saw (*Scene*), and one in which they were overlaid on a black background (*NoScene*). This manipulation served multiple purposes. Because the Search task did not bias fixations to any particular location in the scene (i.e., the search target could be anywhere), the lack of scene content could allow for better classification if what is important for classifying search is the eye movement metrics themselves, not what was looked at. Moreover, the *NoScene* condition is more akin to computer classification in that classifiers are normally provided eye movement kinematics in the absence of information about the image, whereas the *Scene* condition is more similar to the human classification of Greene, Liu, and Wolfe (2011), which coupled this information.

The observers in the present experiment were naïve—no feedback or training was given. Although the role of feedback has been well established in aiding classification tasks (e.g., Zelinsky, Peng, & Samaras, 2013), in the present study, we were interested in whether observers have a priori knowledge of how common tasks (e.g., visual search) are performed by the oculomotor system. If under these conditions of no training and no feedback subjects are still able to classify tasks correctly, it will imply that observers have intuitive knowledge of how gaze is allocated in the tasks used here.

Methods

Participants

A total of 110 naïve undergraduate subjects from the University of Nebraska each participated in three of the conditions, receiving course credit for remuneration. Each session lasted approximately 60 minutes.

Stimulus and Procedure

For the current study, the eye movement data of three randomly chosen subjects from Dodd, Van der Stigchel, and Hollingworth (2009) were collected from three of the previously used tasks—Search, Memory, and Rating. That is, one subject was selected for the Search task, one for the Memory task, and one for the Rating task. Each subject had representative data of the data set as a whole. Although the original study also contained a Free View condition, it was not included in the present experiment because of the ill-defined nature of the task. The Search task was to find an "N" or "Z"

embedded anywhere in the scene (the target was present but nearly indecipherable as the goal of the original investigation was to keep subjects searching for the entire duration of the trial, because eye movements would be impacted if a target was found early—only four subjects reported seeing a target in the initial investigation and none of these subjects reported seeing more than three targets over the duration of the experiment). The Memory task was to memorize the scene for a memory test at the conclusion of the experiment (which never occurred). The Rating task was to judge the pleasantness of each scene on a 7-point scale. This data set was used for the current experiment because previous analyses indicated differences in eye movement and refixation properties as a function of task (see also Mills et al. 2011). In the original Dodd et al. (2009) experiment, a given subject participated in only one of the tasks continuously for the duration of the experiment (i.e., task set comparisons were between subject). Although tasks differed, all subjects saw the same 67 scenes, which were mostly indoor locations (e.g., kitchen, bedroom) with no people present. For the current study, the first 60 images from the original subjects were selected and split into three sets of 20. *Fixation*, *Scanpath*, or *Video* Stimuli were created in conjunction with the natural scene presented to each subject. That is, for each selected subject, 20 of their trials were used to construct the *Fixation* Stimuli, 20 of their trials were used to construct the *Scanpath* Stimuli, and the remaining 20 trials were used to construct the *Video* Stimuli. A different set of 20 trials was chosen from each prior subject for each Stimulus and Task. Specifically, for Subject A, who participated in the Search Task, the first 20 trials of their experiment were used to create *Fixation* Stimuli, the next 20 the *Scanpath* Stimuli, then final 20 the *Video* Stimuli. Then, for Subject B, who participated in the Memory Task, their first 20 trials were used to create the *Scanpath* Stimuli, the next 20 the *Video* Stimuli, and the last 20 the *Fixation* Stimuli. Then, for subject C, who participated in the Rating Task, their first 20 trials were used to create the *Video* Stimuli, the next 20 the *Fixation* Stimuli, and the last 20 the *Scanpath* Stimuli. This process created 60 different images with *Fixations* overlaid, 60 different images with *Scanpaths* overlaid, and 60 different *Videos* of eye movements overlaid. Only the first 4 seconds of viewing (the original trial duration was 8 seconds) was used to prevent clutter in the *Fixation* and *Scanpath* conditions. Additionally, previous work has shown that most task differences in this specific task set lie in the first 4 seconds of viewing (Mills et al. 2011), so only presenting the first four seconds led to presenting classifiers with the most useful information. In addition to creating stimuli with the original scene present, a Scene manipulation was introduced by overlaying the same eye movement data on simply a black background (Figure 1). Thus, in addition to the three stimulus presentation types, there were two scene conditions: *Scene* and *NoScene*.

Combining these manipulations (Stimuli and Scene) led to 6 conditions, each with 60 trials and comprised of 20 trials of each task (Memory, Rating, and Search). The images in each condition were presented in random order (i.e., each trial had an equal probability of being a Memory, Rating, or Search task trial). In the *Video* condition, video was played back in real time, meaning that displays were shown for only four seconds each. In the *Fixation* and *Scanpath* Stimuli conditions, displays were shown for 8 seconds each after pilot testing indicated that 4 seconds was an insufficient amount of time to process the stimuli presented. This is because in the *Video* condition, subjects can follow the moving eye in real time, whereas in the other conditions, the eye movement information was presented simultaneously, which prevents an efficient real-time examination of each location in the same way the eyes moved for the original subjects (e.g., participants need to decide how to process the fixations themselves, because no additional guidance is given as is the case with a single moving point in the videos). Despite the imbalance in presentation time, the same amount of eye movement information was displayed in all conditions (4 seconds). The different presentation times were deemed a necessary compromise as otherwise we would have been required to either a) present *Fixation* and *Scanpath* information for an insufficient amount of time or b) reduce real-time playback to half speed, which appears unnatural and creates fixation durations that are twice as long. At the beginning of the experiment, classifiers were briefed on their current task (i.e., to classify eye movement patterns into task) and were provided with information regarding what the original subjects had been instructed when they performed either the Search, Memory, or Rating task. Classifiers were unaware of the fact that each task was performed by only one subject and were simply told that the eye movement data came from participants who had performed these tasks for a previous study. After each trial, classifiers were asked to indicate via key press whether the observed eye movements corresponded to a task of Search, Memory, or Rating. They were encouraged to guess if they were unsure and no performance feedback was given.

Each classifier took part in the experiment with the following restrictions: 1) they participated once in the *Fixation* condition, once in the *Scanpath* condition (with the order of these initial two tasks counterbalanced), and once in the *Video* condition; 2) if they took part in the *Scene* condition for the *Fixation* stimulus, they took part the *NoScene* condition for the *Scanpath* stimulus (and vice versa); and 3) took part in the *Video* condition last, either in the *Scene* or *NoScene* condition (alternated between subjects). The *Video* stimulus was displayed last so not to confound interpretation of the other stimuli given that the *Video* contained both scanpath and fixation information.

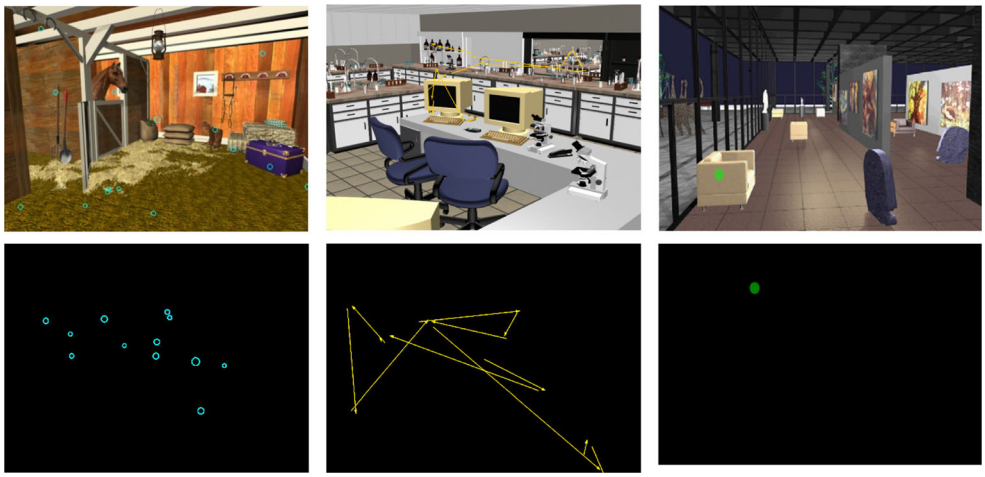

Fig. 1 Example stimuli for Experiment 1, with *Fixations* on the left, *Scanpaths* in the middle, and *Videos* on the right. The *Scene* condition is on top and the *NoScene* on bottom for all stimuli

Apparatus

The eye movements were originally collected using an SR Research Ltd. EyeLink II system with high spatial resolution and a sampling rate of 500 Hz. The present experiments were displayed on a Dell monitor in a testing room with soft lighting and sound.

Results and Discussion

Table 1 shows the observed proportion of responses by Task, Stimuli, and Scene. As shown, subjects were most accurate at classifying the search task. This was particularly true for the *Fixation* Stimuli and for the *NoScene* condition (correct classification of 62%, well above the chance level of 33% and considerably better than the 45% correct classification for the *Fixation-Scene* condition). In contrast to the Search Task, classification for the Memory Task was best in the *Video* condition, particularly for the *Scene* condition (correct classification of 41%).

To confirm these observations statistically, the extent to which Task (Search, Memory, Rating), Scene (*Scene*, *NoScene*), and Stimuli (*Fixation*, *Scanpath*, *Video*) predicted the likelihood of correctly classifying an observer's viewing task was examined in a sample of 19,800 responses, which were nested within 110 subjects and 180 items and where subjects and items were partially crossed given that half of subjects responded to one set of items and the other half responded to another set of items. To account for this cross-classified structure, multilevel models were estimated with random intercepts for subjects and items. The complexity of this structure precluded models from converging. Therefore, the model reported below contained a random intercept for items only. We note that we also estimated a model containing a random intercept for subjects only, and the results were the same.

A generalized logit model for multinomial data with unordered categories, in which the log odds of the outcome was modeled as a linear combination of the predictor variables was selected for analysis. This model contrasts each response level against a reference level via two submodels. Given the observed pattern of data in which experimental effects on memory and rating responses appear similar to each other but both different from search (Table 1), we specified search as the reference level. Thus, submodel 1 was the multinomial logit estimate for memory relative to search responses, whereas submodel 2 was for rating relative to search responses. Task (Search, Memory, Rating) was dummy coded, meaning the intercept term for each submodel directly tests the log-odds of correctly choosing search relative to memory (submodel 1) or rating (submodel 2). Note that for three outcomes, a multinomial logit of 0 is equivalent to a probability of 0.33 (i.e., chance of choosing any 1 of 3 possible responses), with a more positive logit reflecting a greater probability of a response *m* and a more negative logit reflecting a lower probability of response *m*. Thus, the test of whether classification was above chance was whether the intercept for each condition was greater than 0. The statistical package used for analyses was PROC GLIMMIX within SAS 9.4.

Table 2 shows the predicted probability of each response by Task, Stimuli, and Scene. When the observer's task was Search, subjects were more likely to classify the task as Search than as Memory ($Est = -0.343$, $SE = 0.036$, $p < 0.001$) or Rating ($Est = -0.593$, $SE = 0.042$, $p < 0.001$). These differences were significant for each eye movement Stimuli but were larger for *Fixation* than for *Scanpath* ($Est = 0.214$, $SE = 0.051$, $p < 0.001$) or *Video* ($Est = 0.138$, $SE = 0.051$, $p = 0.007$) stimuli. Moreover, this difference was significant for each Scene condition but was larger for the *NoScene* compared with the *Scene* condition ($Est = 0.270$, $SE = 0.069$, $p = 0.001$). The effect of *Scene* was larger for the *Fixation* stimuli than for the *Scanpath* ($Est = 0.221$, $SE = 0.103$, $p = 0.034$)

Table 1 Observed proportion of Search, Memory, and Rating responses by Task, Stimuli, and Scene

Task	Response	Fixation		Scanpath		Video	
		Scene	NoScene	Scene	NoScene	Scene	NoScene
Search	*Search*	*44.58%*	*61.96%*	*36.47%*	*41.19%*	*38.64%*	*40.98%*
Search	Memory	30.59%	22.84%	35.69%	32.88%	31.53%	33.82%
Search	Rating	24.83%	15.20%	27.84%	25.93%	29.83%	25.20%
Memory	Search	33.56%	25.39%	35.88%	30.42%	32.91%	39.61%
Memory	*Memory*	*39.92%*	*39.31%*	*34.12%*	*36.10%*	*40.63%*	*30.98%*
Memory	Rating	26.53%	35.29%	30%	33.47%	26.46%	29.41%
Rating	Search	33.39%	24.61%	37.16%	37.46%	39.58%	39.94%
Rating	Memory	36.44%	38.53%	35.78%	38.64%	35.34%	32.68%
Rating	*Rating*	*30.17%*	*36.86%*	*27.06%*	*23.90%*	*25.08%*	*27.38%*

Italicized values reflect correct classification

or *Video* stimuli (*Est* = 0.283, *SE* = 0.098, $p = 0.004$). In sum, *Search* was classified by far the best in the *Fixation-NoScene* condition.

When the observer's task was Memory, subjects were more likely to classify the task as Memory than as Search (*Est* = 0.454, *SE* = 0.051, $p = 0.003$) or Rating (*Est* = 0.586, *SE* = 0.061, $p < 0.001$). This difference was significant for each eye movement Stimuli but was larger for *Fixation* Stimuli than for *Scanpath* (*Est* = 0.274, *SE* = 0.072, $p < 0.001$) or *Video* stimuli (*Est* = 0.269, *SE* = 0.072, $p < 0.001$). The effect of Scene was significant only for the *Videos*, with correct classification of Memory more likely in the *Scene* compared with the *NoScene* condition (*Est* = 0.518, *SE* = 0.145, $p = 0.004$). In sum, Memory was classified the best in the *Video-Scene* condition.

When the observer's task was Rating, subjects were actually less likely to classify the task as Rating than as Search (*Est* = −0.364, *SE* = 0.059, $p < 0.001$) or Memory (*Est* = −0.354, *SE* = 0.058, $p = 0.002$). This was not true, however, for the *Fixation* Stimuli, in which case subjects were more likely to classify the observer's task as Rating than as Search (*Est* = 0.409, *SE* = 0.083, $p = 0.001$) or Memory (*Est* = 0.385, *SE* = 0.083, $p = 0.012$), although classification still was not above chance (*Est* = 0.068, *SE* = 0.097, $p = 0.487$). There was a significant effect of Scene for the *Fixation* Stimuli with correct classification of Rating more likely in the *NoScene* compared with the *Scene* condition (*Est* = 0.486, *SE* = 0.138, $p = 0.005$). In sum, the only condition for which correct classification of Rating was above chance was for the *Fixation-NoScene* condition.

There are two critical findings in Experiment 1. First, the Search Task was classified the best out of all three Tasks, followed by Memory, followed by Rating. Second, there was a differing effect of Scene on Search and Memory, with Search better in the *NoScene* (specifically *Fixation*) condition, and Memory better in the *Scene* (specifically *Video*) condition. Additionally, it is of note that the condition that was most like Greene, Liu, and Wolfe (2012), the *Scanpath-Scene* condition, replicates their results of poor classification.

Table 2 Predicted probability of Search, Memory, and Rating responses by Task, Stimuli, and Scene

Task	Response	Fixation		Scanpath		Video	
		Scene	NoScene	Scene	NoScene	Scene	NoScene
Search	*Search*	*36.16%*	*53.59%*	*34.30%*	*38.92%*	*37.49%*	*39.97%*
Search	Memory	33.75%	26.84%	37.06%	34.30%	30.30%	32.71%
Search	Rating	30.09%	19.58%	28.64%	26.78%	32.22%	27.32%
Memory	Search	37.98%	29.43%	38.86%	33.17%	29.61%	36.43%
Memory	*Memory*	*39.58%*	*39.92%*	*33.04%*	*35.22%*	*45.86%*	*35.69%*
Memory	Rating	22.44%	30.65%	28.10%	31.61%	24.53%	27.88%
Rating	Search	38.34%	28.84%	37.04%	37.46%	39.46%	39.91%
Rating	Memory	34.90%	37.71%	34.35%	37.24%	36.70%	34.00%
Rating	*Rating*	*26.75%*	*33.45%*	*28.60%*	*25.30%*	*23.83%*	*26.09%*

Italicized values reflect correct classification

Experiment 2

The results of Experiment 1 suggested that human observers were able to correctly classify some tasks under some viewing conditions. In particular, the *Fixation-NoScene* condition allowed for very high rates of classification of the Search task, and the *Video-Scene* condition allowed for high rates of classification of the Memory task. However, there exist methodological concerns for the generalizability of the findings.[1] For example, there was imprecise counterbalancing of conditions (e.g., the Video condition was always displayed last), and there were only three subjects eye movement patterns displayed to the classifiers. In Experiment 2, we used the same general method from Experiment 1 but with a few refinements. There were now only four classifying conditions: *Fixation-Scene, Fixation-NoScene, Video-Scene, Video-NoScene*. We focused on these conditions as they were the ones Experiment 1 suggested human classifiers could use to classify tasks correctly. By focusing on these tasks, we hoped to both replicate our original findings from Experiment 1 and possibly elaborate on the information classifiers were using to perform the task by analyzing their correct compared with incorrect responses. Moreover, in Experiment 2, we included eye movement stimuli from a much larger pool of the original observers.

Methods

Participants

A total of 100 naïve undergraduate subjects from the University of Nebraska were each subjected to two of the conditions, receiving course credit for remuneration. Seven subjects who participated in the *Fixation-Scene/Video-NoScene* conditions were eliminated from further analysis, as were three subjects from the *Fixation-NoScene/Video-Scene* conditions, because of either not completely finishing both conditions, or for not correctly participating in the task (e.g., guessing Search on every trial). Each session lasted approximately 60 minutes. Subjects eliminated from one condition were also eliminated from the other (e.g., if a subject did not finish the *Video-Scene* condition, their results from the *Fixation-NoScene* condition were not further considered).

Stimulus and Procedure

Classifiers saw 360 classification images in 2 blocks. In the first block, they were randomly chosen to participate in one of the four viewing conditions (*Fixation-Scene, Fixation-NoScene, Video-Scene, Video-NoScene*). In the second block, they participated in the other stimuli and scene condition. For example, if a classifier was randomly chosen to participate in the *Fixation-NoScene* condition for their first block, they then participated in the *Video-Scene* condition in the second block. Each block of 180 trials had 60 Search trials, 60 Memory trials, and 60 Rating trials, presented in random order.

As in Experiment 1, we collected eye movement data to present to the classifiers from Dodd, Van der Stigchel, and Hollingworth (2009). We collected data from 12 different subjects (none of the subjects were used in Experiment 1): 4 for the Memory images, 4 for the Search images, and 4 for the Rating images. As in Experiment 1, classifiers were unaware that multiple subjects were present in each task and were given no specific information regarding how many participants the eye movement data was drawn from. Each of the 12 subjects contributed 30 trials to the classification data set. Fifteen of the trials were chosen to be placed into the Fixation stimuli, with the other 15 placed into the Video stimuli. The same 15 trials were present in the both the *Scene* and *NoScene* version for each Stimuli condition. To give an example of how this process was performed, we took a subject who participated in the Memory task from Dodd, Van der Stigchel, and Hollingworth (2009) and selected 30 images from the original set's first 60 images. Fifteen images were chosen to be used for the *Fixation* stimuli, with the other 15 being used for the *Video* stimuli. Then, for a second subject who participated in the Memory task, the same 30 images were gathered as from the first subject, but the 15 images used for the *Fixation* stimuli from the first subject were now used for the *Video* stimuli, and vice versa. This process was repeated for two more participants from Dodd, Van der Stigchel, and Hollingworth (2009). In total, there were 60 Memory images for both the *Fixation* and *Video* stimuli. We repeated the same process for the Search and Rating images, creating the 180 trial data set for each of the four viewing conditions. Given the larger number of trials relative to the initial dataset, scenes could repeat multiple times though the eye movement data provided from each original participant was always novel. Because only 30 images were used from the original data set, each scene was repeated six times throughout the course of classification (twice in each task). Participants were provided this information.

Results

Table 3 shows the observed proportion of responses by Task, Stimuli, and Scene. As before, participants were most accurate at classifying the search task with *Fixation* stimuli in the *NoScene* condition (47% classification accuracy), whereas they were most accurate at classifying the Memory task with *Video* stimuli in the *Scene* condition (40% classification accuracy).

[1] We thank three anonymous reviewers for their helpful suggestions in developing Experiment 2.

Table 3 Observed mean proportion of Search, Memory, and Rating responses by Task, Stimuli, and Scene

Task	Response	Fixation		Video	
		Scene	NoScene	Scene	NoScene
Search	*Search*	*34.65%*	*47.02%*	*42.31%*	*39.66%*
Search	Memory	32.21%	28.18%	31.39%	33.24%
Search	Rating	33.14%	24.80%	26.30%	27.09%
Memory	Search	35.66%	30.99%	31.31%	33.61%
Memory	*Memory*	*35.00%*	*36.85%*	*39.67%*	*34.21%*
Memory	Rating	29.34%	32.16%	29.02%	32.17%
Rating	Search	34.96%	29.73%	32.25%	35.66%
Rating	Memory	32.64%	39.47%	38.05%	34.19%
Rating	*Rating*	*32.40%*	*30.80%*	*29.70%*	*30.15%*

Italicized values reflect correct classification

Statistical analyses were the same as in Experiment 1. Table 4 shows the predicted probability of each response by Task (Search, Memory, or Rating), Stimuli (*Fixation* or *Video*), and Scene (*Scene* or *NoScene*). When the observer's task was Search, participants were significantly more likely to classify the task as Search than as Memory ($Est = -0.570$, $SE = 0.050$, $p < 0.001$) or Rating ($Est = -0.799$, $SE = 0.051$, $p < 0.001$. These differences were significant both with *Fixation* and *Video* stimuli ($ps < 0.001$) and did not differ significantly between the two Stimulus conditions (submodel 1: $Est = 0.096$, $SE = 0.099$, $p = 0.337$; submodel 2: $Est = -0.189$, $SE = 0.102$, $p = 0.062$). Moreover, this difference was significant for each Scene condition ($ps < 0.001$) and was more negative for the *NoScene* compared with *Scene* condition (submodel 1: $Est = -0.147$, $SE = 0.099$, $p = 0.140$; submodel 2: $Est = -0.236$, $SE = 0.102$, $p = 0.021$). The effect of Scene was significantly larger with *Fixation* versus *Video* stimuli (submodel 1: $Est = 0.598$, $SE = 0.199$, $p = 0.003$; submodel 2: $Est = 0.709$, $SE = 0.203$, $p < 0.001$). In sum, search was classified the best with *Fixation* stimuli in the *NoScene* condition.

Table 4 Predicted mean probability of Search, Memory, and Rating responses by Task, Stimuli, Scene

Task	Response	Fixation		Video	
		Scene	NoScene	Scene	NoScene
Search	*Search*	*42.75%*	*55.54%*	*51.58%*	*48.20%*
Search	Memory	28.81%	23.96%	28.41%	30.85%
Search	Rating	28.43%	20.47%	19.92%	20.93%
Memory	Search	35.32%	30.13%	29.48%	31.91%
Memory	*Memory*	*38.06%*	*40.53%*	*44.35%*	*38.34%*
Memory	Rating	26.58%	29.32%	26.12%	29.70%
Rating	Search	33.73%	28.14%	32.32%	35.51%
Rating	Memory	35.85%	43.28%	40.91%	37.10%
Rating	*Rating*	*30.32%*	*28.54%*	*26.68%*	*27.27%*

Italicized values reflect correct classification

When the observer's task was Memory, participants were significantly more likely to classify the task as Memory than as Search ($Est = -0.368$, $SE = 0.046$, $p < 0.001$) or Rating ($Est = -0.242$, $SE = 0.050$, $p < 0.001$). This result was true for both *Fixation* and *Video* stimuli, as well as in the *Scene* and *NoScene* condition, evident by nonsignificant effects of Stimuli and Scene ($ps > 0.094$). Importantly, the effect of Scene was significant only with *Video* stimuli, with correct classification of Memory more likely in the *Scene* compared with the *NoScene* condition ($Est = -0.336$, $SE = 0.141$, $p = 0.018$). In sum, memory was classified the best with *Video* stimuli in the *Scene* condition.

When the observer's task was Rating, participants were significantly less likely to classify the task as Rating than as Search ($Est = 0.137$, $SE = 0.051$, $p = 0.007$) or Memory ($Est = 0.327$, $SE = 0.046$, $p < 0.001$). This was true both for *Fixation* and *Video* Stimuli, as well as in the *Scene* and *NoScene* condition, evident by nonsignificant effects Stimuli and Scene ($ps > 0.070$).

Experiment 2 replicated the major findings of Experiment 1: the Search task was classified the best, followed by Memory, followed by Rating. Additionally, the Search Task was classified the best in the *Fixation-NoScene* condition and the Memory task in the *Video-Scene* condition.

Although observing that participants could classify these tasks in the two conditions particularly well, another question is, how are they performing the classification? Given the larger number of trials in each condition in Experiment 2, we were well positioned to answer this question. There were three eye movement metrics that our data set was particularly suited to evaluate: fixation duration (i.e., in *Fixation* condition, the radius of the circles carries this information, with a larger radius = longer fixation duration), number of fixations (i.e., in the *Fixation* condition, this is the number of circles), and saccade amplitude (a larger mean saccade amplitude acts as a proxy for distribution of fixations). Before reporting how these eye movement metrics differed between classifiers, we report descriptive statistics for the metrics themselves (Table 5). The most important difference to notice is that Memory and Search differed on all three metrics, with Search having shorter mean fixation durations, more fixations per trial, and larger saccade amplitudes compared with Memory. Interestingly, Rating also differed from Search in much the same way as Memory, but these differences did not lead to increased classification performance. This may be because the Rating task does not intuitively map onto eye movement strategies in the same way as Search or Memory.

In evaluating whether these eye movement metrics differed between successful and unsuccessful classifiers, we first considered the Search Task of the *Fixation-NoScene* condition. In general, search tasks are characterized by rapid scanning, which manifests as short, numerous fixations spread over the

Table 5 Observed mean fixation duration (ms), number of fixations, and saccade amplitude (visual degrees) during the first 4 seconds of viewing from the four original participants eye movement metrics for each task

Task	Fixation duration	Number of fixations	Saccade amplitude
Memory	240.6	13.2	8.1
Rating	241.4	13.5	7.6
Search	224.9	14.4	9.2

scene. Thus, those scenes with shorter fixation durations, more fixations, and larger saccade amplitudes may have been preferentially classified as Search if classifiers had intuitive knowledge of how a search task is performed.

To test this hypothesis, we further examined the subjects who responded significantly above chance to the Search Task (>30 correct responses out of 60, according to chi-squared test). Twenty-one of the 47 subjects who participated in the *Fixation-NoScene* condition met this criterion. We then investigated the mean fixation durations, mean number of fixations, and average saccade amplitude for each of the 60 classification images as a function of whether they were classified correctly or incorrectly. For these 21 subjects, there was a significant difference between incorrect and correct trials in both mean fixation duration, $t(20) = 4.01$, $p < 0.001$ (240 ms for incorrect vs. 219 ms for correct), and mean number of fixations, $t(20) = 4.98$, $p < 0.001$ (13.9 fixations for incorrect vs. 14.7 fixations for correct). Additionally, there was numerical difference in the average saccade amplitude (9.02 for incorrect vs. 9.28 for correct), although the effect did not reach significance, $t(20) = 1.69$, $p = 0.105$. Moreover, a significant difference in mean fixation duration was present for those 26 subjects who classified search below chance. However, they appeared to have considered this metrics in the opposite manner. That is, they classified those scenes with shorter fixation durations as a task *other than* search: $t(25) = 2.41$, $p = 0.024$ (223 ms for incorrect vs. 240 ms for correct). There was no significant difference for either number of fixations, $t(25) = 1.28$, $p = 0.211$ (14.4 for incorrect vs. 14.1 for correct) or average saccade amplitude, $t(25) = 1.04$, $p = 0.307$ (9.26 for incorrect vs. 9.14 for correct). In sum, these analyses suggest that successful classifiers were using both the number of fixations (as measured by mean fixation duration and number of fixations) and their spread (as measured by average saccade amplitude): more fixations and higher spread were signals to classify the task as search.

Next, we considered the high performance for the Memory Task in the *Video-Scene* condition. Mean fixation duration and mean number of fixations were predictive of Search in the *Fixation-NoScene* condition. We hypothesized that the same pattern may hold for the Memory Task in the Video-Scene condition but in the opposite direction; because Memory tasks typically have longer fixation durations and fewer number of fixations (possibly because of a need to more deeply encode objects or because fewer objects/locations may be task relevant in memory relative to search in which every area of the scene is potentially relevant), it is possible that subjects were able to use this to classify the Memory Task. To test this hypothesis, we collected those subjects in the classification task who responded significantly above chance to the Memory Task (>50%). Eleven of 47 subjects who participated in the *Video-Scene* condition met this criterion. For these 11 subjects, there was a significant difference between incorrect and correct trials in both mean fixation duration, $t(10) = 3.43$, $p = 0.006$ (242 ms for incorrect vs. 257 ms for correct), and mean number of fixations, $t(10) = 2.70$, $p = 0.022$ (12.8 fixations for incorrect vs. 12.3 fixations for correct), but no difference in average saccade amplitude $t(10) = 0.096$, $p = 0.925$. Unlike for the *Fixation-NoScene* condition, there was no significant difference for those subjects who did not classify Memory above chance levels. These results suggest that successful classifiers of memory were using the amount of fixations as a marker of memory.

In sum, Experiment 2 replicated the major findings from Experiment 1. The Search Task was classified the best and particularly well in the *Fixation-NoScene* condition. Additionally, it appeared subjects who classified the Search Task the best in this condition were selectively classifying scenes as Search when they had shorter mean fixation durations and a larger number of total fixations. Moreover, the Memory Task in the *Video-Scene* condition was classified well, as in Experiment 1. Those subjects who classified the Memory Task the best in this condition also were selectively utilizing eye movement information but the opposite pattern from the Search Task: these subjects selectively classified scenes as Memory when they had longer mean fixation durations and a smaller number of total fixations.

General Discussion

In the present study, we sought to determine whether human observers could correctly classify task (Memory/Search/Rating) from eye movements and whether their success/failure in this regard could inform the success/failure of similar classification attempts. We introduced a Scene information manipulation, wherein one condition had scene information (the eye movements from original subjects were overlaid on the original scene), and another condition had no scene information (the eye movements were overlaid on a black background). This manipulation was important given that it is more consistent with the manner in which computer classifiers have traditionally classified tasks, referencing only the eye movement characteristics independent of the image. In Experiment 1, subjects were shown three different sets of eye movement Stimuli (*Fixations*, *Scanpaths*, and *Videos*) on either the original background (*Scene*) or a black background (*NoScene*). In

Experiment 2, subjects were only shown two different sets of eye movement Stimuli (*Fixations*, *Videos*) on either the original background (*Scene*) or a black background (*NoScene*).

We observed three main results across these two experiments. First, the Search Task was classified the best out of all three Tasks, followed by Memory, followed by Rating (rating was not classified at above chance levels in either experiment). Second, there was a differing effect of Scene on Search and Memory, with Search better in the *NoScene* (specifically *Fixation*) condition, and Memory better in the *Scene* (specifically *Video*) condition. Third, it appeared successful classifiers were using specific eye movement information to guide their classification. For the Search Task, those subjects who classified the task particularly well in the *Fixation-NoScene* condition selectively classified scenes as Search if they had short fixation durations and many fixations. For the Memory task, the opposite was true: those subjects who classified the task particularly well in the *Video-Scene* condition selectively classified scenes as Memory if they had long fixation durations and few fixations.

Previous computer classification has shown that a search task typically can be classified above chance (Borji & Itti, 2014; Borji, Lennartz, & Pomplun, 2015; Haji-Abolhassani & Clark, 2014; Henderson et al., 2013). Similarly, we found a strong correct classification rate for Search, particularly in the *Fixation-NoScene* condition. This result holds in both experiments even when accounting for the human classifiers higher than chance rate of responding Search as the task. In previous research, when computers have classified, they have done so only with respect to the eye movement properties themselves, akin to the present experiment's *NoScene* condition. In this type of setup, there is no information for the classifier (whether human or computer) about the scene itself and where objects were located. In the case of the stimuli for the present experiment taken from Dodd, Van der Stigchel, and Hollingworth (2009), searchers were not necessarily more likely to fixate specific areas of a scene, as the search target was a small and difficult to discern letter "N" or "Z" that could have been located anywhere in the scene. Regardless of search target, however, search tends to elicit eye movement characteristics that differ from other tasks, namely a larger number of fixations and shorter individual fixation durations in various locations (Mills et al. 2011). This might provide an indication as to why subjects were actually better at classifying search in the absence of the accompanying scene, because it allows an easier examination of the eye movement metrics themselves. When we evaluated whether successful classifiers relied on particular eye movement metrics, we found that those subjects who were better classifiers of the Search Task selectively classified those scenes with a larger number of fixations and shorter individual fixation durations as Search.

Moreover, when the scene image is present, as in the *Fixation-Scene* condition, subjects may attempt to infer task by examining which areas/objects are being fixated, which may in turn distract from the important information about the number of fixations and duration of each. In the absence of this information, however, subjects become better able to focus on fixation characteristics in much the same way a computer classifier would. In future work, it may be important to determine whether this lack of scene advantage for classification by humans is moderated when scene context actually dictates how search is performed. For example, when looking for a coffee pot in a kitchen scene, there are locations where the item is expected to appear, such as countertops, and a lower likelihood it would appear in other locations, such as the floor or ceiling. It seems possible that scene context would aid in classifying this type of search task: if fixations clustered around the search target (e.g., the coffee pot), it may be easy to classify the task as search. It also could be the case that search would become less dissociable from memory and rating under these conditions given that fixation locations may become more similar as scene context dictates the manner in which a task is performed. That is, as tasks become more constrained by scene context, their eye movement metrics become more similar. A search task in which one is looking for a coffee pot and a memory task in which one is memorizing objects in a kitchen will have more similar eye movement metrics than the tasks used in the current study that were not constrained by scene context.

Although the classification of the search task was aided by the lack of scene context, this *NoScene* benefit was not found for the Memory Task, which was classified at the highest percentage in the *Video-Scene* condition. Like the Search Task, this result holds even when accounting for a slightly higher than chance likelihood of human classifiers answering Memory as the task. Memory tasks tend to elicit eye movement characteristics that differ from Search, namely a fewer number of fixations and longer individual fixation durations (Mills et al. 2011). As in the Search Task from the *Fixation-NoScene* condition, this difference in eye movement characteristics may be a possible avenue for some subjects to classify the Memory task. When we tested this hypothesis, this is exactly what we found: those subjects who were above chance classifiers of the Memory Task in the *Video-Scene* condition selectively classified those scenes with a fewer number of fixations and longer individual fixation durations as Memory. However, it is important to note that the above chance classification cannot *only* be due to the fixation information; otherwise the *Fixation-Scene* condition should have had similar performance to the *Video-Scene* condition. Thus, we speculate that the increased performance in the *Video* compared with *Fixation* condition for classifying Memory is due to the order information present in the *Video* but absent in the *Fixation* condition. Although not explicitly tested here, previous analysis of the data sets used here has shown that people are faster to refixate the same objects in a memory task compared with a

search task (Dodd et al., 2009). In the *Fixation* condition, participants would be unable to differentiate refixations (separated in time by distant fixations) by subsequent fixations on the same object. However, in the *Video* condition, this distinction is apparent. Thus, it may be the case that the order information provided in the *Video* condition aids classification in addition to the fixation information.

One important consideration from the present experiments is that, although some subjects appeared to be using eye movement characteristics like mean fixation duration and number of fixations, they were only able to classify Memory above chance when the scene was present, i.e., when the fixation locations were tied to specific regions of a scene. Although not explicitly tested here, the present experiment may illuminate the difficulty of previous computer classifiers to classify memory above chance (Greene et. al, 2011), as it is indicative of scene information being important for classification of memory. When an observer attempts to memorize a scene, they must focus on the identity of objects within and the spatial reference frames between objects meaning that "what" they fixate is at least as important as the spatial locations they fixate. Given that computer classifiers classify task independent of scene statistics, they may be missing critical information that is required for a successful outcome. Thus, combining classifiers, such as an image and task classifier, may lead to better classification by computers for memory tasks.

One interesting supplement to the present classification data also can be obtained by considering subject's self-reports. At the end of each experimental session, subjects were asked to report any strategy they felt they may have employed for classifying. Many were able to verbalize certain characteristics that they had looked for in making their decision, and this informed our correct/incorrect analyses in Experiment 2. Search eye movements were identified as being "rapid and all over the image," memory tended to be focused "on specific objects," and rating was a "blending" of the two tasks. Human observers seem to have an implicit understanding of how these three tasks are performed generally given that these self-reports adhere to the trends in the data. The Search Task appears to not require the Scene, because the important information was that the eye movement patterns themselves. Hence, the Scene acted as an excess of irrelevant information, adding to perceptual load and hurting performance. The Memory Task, although also utilizing eye movement patterns, may have required the Scene for better classification in order to deduce that objects were preferentially fixated. Finally, the Rating Task was classified the least well as it was more difficult to dissociate from the other task types.

In the present experiment, we have shown that humans can classify task by eye movements under certain conditions. These findings further illuminate the results from previous classification research, showing why some classifiers have succeeded and others have failed. Together, these new findings, combined with the old, advance understanding of classification of task by eye movements.

Acknowledgements The study was partially supported by NIH Grant R01 EY022974. The authors thank Jordan Marshall and Alex Olsen for their assistance with data collection and Gerald McDonnell and Monica Rosen for helpful comments on an earlier version of this manuscript. They also thank Greg Zelinsky, Franco Amati, Joe Schmidt, and an anonymous reviewer for helpful comments on a previous version of this manuscript.

References

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision, 14,* 1–22.

Borji, A., Lennartz, A., & Pomplun, M. (2015). What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. *Neurocomputing, 149,* 788–799.

Caucci, G. M., & Kreuz, R. J. (2012). Social and paralinguistic cues to sarcasm. *Humor, 25,* 1–22.

Dodd, M. D., Van der Stigchel, S., & Hollingworth, A. (2009). Novelty is not always the best policy: Inhibition of return and facilitation of return as a function of visual task. *Psychological Science, 20,* 333–339.

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,* 1–19.

Foulsham, T., & Kingstone, A. (2013). Where have eye been? Observers can recognize their own fixations. *Perception, 42,* 1085–1089.

Friere, A., Eskritt, M., & Lee, K. (2004). Are eyes windows to a deceiver's soul? Children's use of another's eye gaze cues in a deceptive situation. *Developmental Psychology, 40,* 1093–1104.

Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research, 62,* 1–8.

Haji-Abolhassani, A., & Clark, J. J. (2014). An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision Research, 103,* 127–142.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16,* 850–856.

Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PLoS ONE, 8,* 1–6.

Hollingworth, A. (2009). Two forms of scene memory guide visual search: Memory for scene context and memory for the binding of target object to scene location. *Visual Cognition, 17,* 273–291.

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

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision, 11,* 1–15.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5,* 1–7.

Yarbus, A. (1967). Eye Movements and Vision. *Springer.*

Zelinsky, G. J., Peng, Y., & Samaras, D. (2013). Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision, 13,* 1–13.