Vision Research 62 (2012) 1–8

Contents lists available at SciVerse ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns

Michelle R. Greene [a,b,c,*], Tommy Liu [a], Jeremy M. Wolfe [a,b]

[a] Brigham & Women's Hospital, United States
[b] Harvard Medical School, United States
[c] Stanford University, United States

ARTICLE INFO

Article history:
Received 14 July 2011
Received in revised form 22 March 2012
Available online 2 April 2012

Keywords:
Eye movements
Multivariate pattern classification
Yarbus
Task

ABSTRACT

In 1967, Yarbus presented qualitative data from one observer showing that the patterns of eye movements were dramatically affected by an observer's task, suggesting that complex mental states could be inferred from scan paths. The strong claim of this very influential finding has never been rigorously tested. Our observers viewed photographs for 10 s each. They performed one of four image-based tasks while eye movements were recorded. A pattern classifier, given features from the static scan paths, could identify the image and the observer at above-chance levels. However, it could not predict a viewer's task. Shorter and longer (60 s) viewing epochs produced similar results. Critically, human judges also failed to identify the tasks performed by the observers based on the static scan paths. The Yarbus finding is evocative, and while it is possible an observer's mental state might be decoded from some aspect of eye movements, static scan paths alone do not appear to be adequate to infer complex mental states of an observer.

1. Introduction

If you are a vision researcher, Fig. 1 is most likely familiar to you. It is from Yarbus (1967) seminal monograph *Eye Movements and Vision*. In addition to being a landmark in the history of experimental psychology, this work has been widely cited in the fields of neuroscience, ophthalmology, and artificial intelligence (Tatler et al., 2010). Yarbus argued that changing the information that an observer is asked to obtain from an image drastically changes his pattern of eye movements. Moreover, the scan paths from this famous figure have been taken as evidence that eye movements can be windows into rather complex cognitive states of mind. However, as impressive as this demonstration appears to be, this figure represents the scan paths of only one observer examining only one image, using a very rudimentary eye tracking system. Several studies have improved on the original finding by using multiple participants and modern eye tracking systems (DeAngelus & Pelz, 2009; Tatler et al., 2010), by showing that different tasks, in general, alter patterns of eye movements (Castelhano, Mack, & Henderson, 2009; Tatler, Baddeley, & Vincent, 2006), or by showing task differences for eye movements related to real-world activities such as reading, sandwich making or fencing (Ballard, Hayhoe, & Pelz, 1995; Hagemann et al., 2010; Hayhoe et al., 2003; Kaakinen & Hyönä, 2010; Land & Hayhoe, 2001; Land, Mennie, & Rusted, 1999).

While these later studies have shown that an observer's task can change certain individual features of eye movement patterns, they have not shown whether these differences can be used to *identify* the task of the observer. This has been an untested, but popular inference from the Yarbus finding as the visual differences between scan paths appeared so different. Can human observers or machine classifiers predict the task of an observer from scan paths similar to those in Fig. 1?

Scan paths lend themselves to analysis using multivariate pattern analysis techniques such as linear discriminant analysis and linear support vector machine classification. These are widely deployed in functional neuroimaging and single-cell neural recordings (for a review, see Norman et al., 2006) to infer the representational content of neural activity patterns. Here, multiple features of the scan path are computed and treated as a pattern, rather than comparing single scan path features such as fixation duration. Although a large number of statistical models could be employed for classification, linear models tend to perform better and have fewer problems with overfitting data than more complex, non-linear models (Misaki et al., 2010). In principle, cognitive states can be decoded from single trials. Finally, similarities between cognitive states might be inferred from the patterns of errors made by the classifier.

In these studies, we test the hypothesis that an observer's eye movements carry information that is diagnostic of the task he is

* Corresponding author at: Brigham & Women's Hospital, United States.
E-mail addresses: mrgreene@stanford.edu (M.R. Greene), wolfe@search.bwh.harvard.edu (J.M. Wolfe).

0042-6989/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.visres.2012.03.019

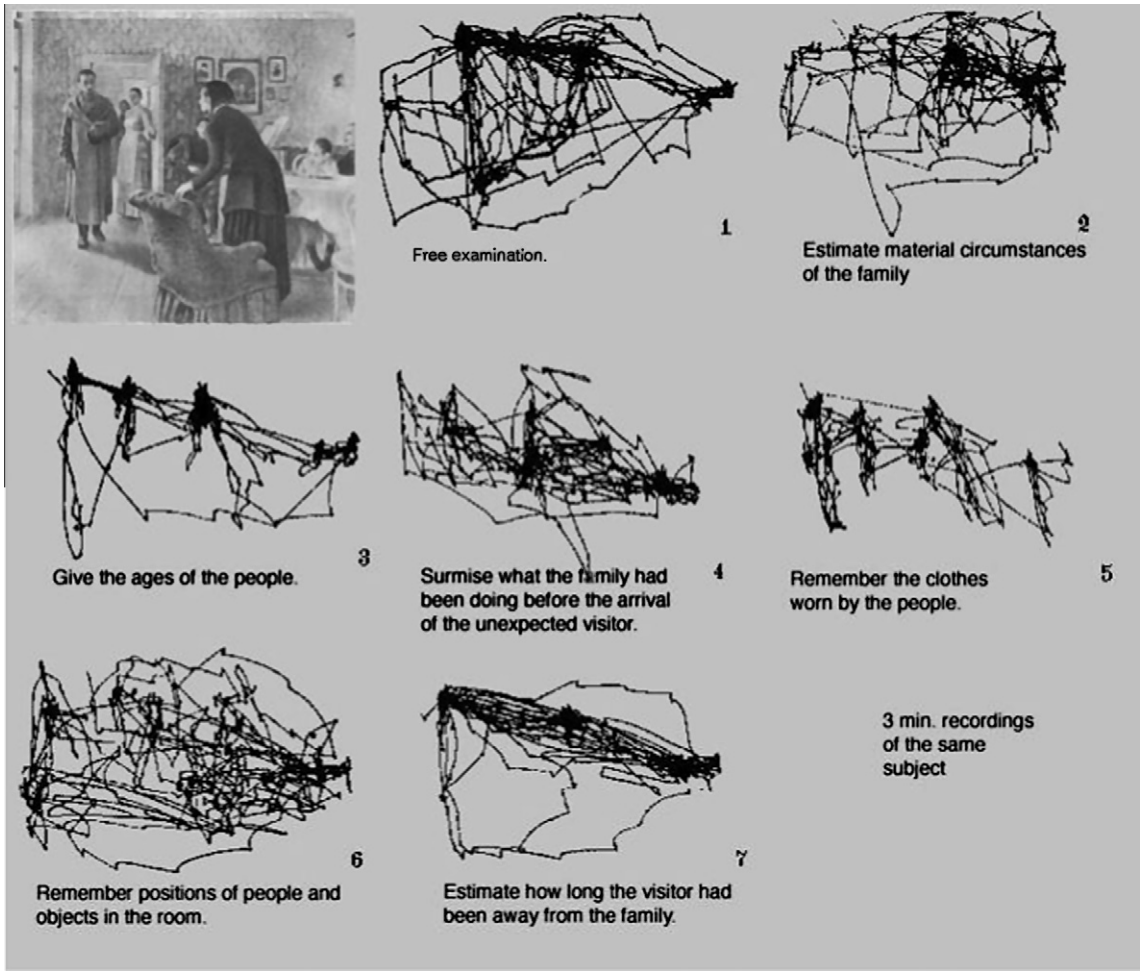

Fig. 1. Figure 109 from Yarbus (1967).

trying to perform. To test this hypothesis, we train a linear classifier with the eye movements recorded from observers viewing a set of images. We then test this classifier with images not used in training to see if task information is available in these static scan paths. We compare the performance of the classifier to the performance of human observers who were also trying to predict task based on scan paths.

2. Experiment 1

In this experiment, we tested the ability of a linear discriminant classifier to determine what information human observers had been asked to glean from a photograph.

2.1. Methods

2.1.1. Materials

Though it is difficult to replicate the cultural importance of the Repin painting for modern observers outside of Russia, we tried to use rich and informative images. Stimuli for this experiment consisted of 64 grayscale photographs taken from the Time Life archive on Google (http://images.google.com/hosted/life). Scene selection was constrained to photographs taken between 1930 and 1979 that contained at least two people. The image set was also chosen to represent a wide variety of social situations in several cultural contexts. Care was taken to avoid photographs of celebrities (politicians, musicians or actors) who could be known to the observers. Although wide ranges of emotional contexts were portrayed, photographs containing gory or traumatic scenes were also avoided. Examples of images used in Experiment 1 are shown in Fig. 2. Images were displayed on a 21-in. CRT monitor, and subtended 24° by 24° of visual angle at 57.4 cm viewing distance.

Eye movements were recorded with an Eyelink 1000 eye tracker (SR research), which sampled eye position at 1000 Hz. Although viewing was binocular, only the right eye was tracked. A nine-point calibration procedure was done at the beginning of the experiment. The experimenter did not accept calibration unless the average spatial error was 1° or less. Participants were re-calibrated after every eight images to ensure continuously accurate calibration. The experiment was run with MATLAB using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997).

2.1.2. Observers

Sixteen observers took part in Experiment 1. All were between the ages of 18–55, had normal or corrected-to-normal vision and had no history of eye or muscle disorders. All provided informed consent and were compensated $10/h for their time.

2.1.3. Design and procedure

Participants were seating 57.4 cm away from the display monitor, using a chin rest. After initial calibration, participants viewed

Fig. 2. Example scenes from the 64-image database. Each scene contained at least two people.

each of the 64 images one time in four blocks of 16 images. Each image was viewed for 10 s. In each block, participants were given one of the following viewing instructions (tasks):

1. Memorize the picture (*memory*).
2. Determine the decade in which the picture was taken (*decade*).
3. Determine how well the people in the picture know each other (*people*).
4. Determine the wealth of the people in the picture (*wealth*).

The order of tasks was counterbalanced across participants. Thus, each participant viewed each image under only one set of viewing instructions. Since four participants were needed to create a complete set of observations, our 16 observers produced four full sets. After the 10 s viewing, participants provided a numerical response using a visual analog scale that was shown on the screen (except for memorization task).

2.1.4. Eye movement measures

For each trial, we computed the following seven measures from the eye movement scan paths: (1) number of fixations, (2) the mean fixation duration, (3) mean saccade amplitude, and (4) percent of image covered by fixations assuming a 1° fovea. These summary statistics are commonly used features for scan path analysis (Castelhano, Mack, & Henderson, 2009; Mills et al., 2011).

In addition, we computed the proportion of dwell-time on various regions of interest: (5) faces, (6) human bodies, and (7) objects. Objects are strong predictors of eye fixations (Einhäuser, Spain, & Perona, 2008). For each image, all regions of interest were defined with rectangular bounding boxes. Face regions of interest encompassed front and profile views of human heads. Body regions of interest were defined as the bodies (but not heads) of people. Object regions of interest were defined as any discrete artifact not making up the boundary of the scene. Hence, items such as paintings, cups and candles counted as objects, but sky and walls did not.

2.1.5. Pattern classification

The seven eye movement measures described above were fed into a linear discriminant classifier.[1] For this experiment, we trained three separate classifiers: one to predict which image was being viewed, one to predict which participant produced a particular pattern of eye movements, and critically, one to predict which task was being done by a participant.

The classifier was iteratively trained and tested on all trials using leave-one-out cross validation. In this procedure, one trial is removed from the dataset as the test trial, while the remaining trials are used as training. Training trials are labeled with stimulus condition (the task being performed, in the case of task classification, for example), and a discriminant axis is fit that maximizes distance between classes. Then, the test trial is classified based on the discriminant function of the training set. This was done iteratively such that each trial in the dataset served as the test image (see Duda, Hart, & Stork, 2001).

2.2. Results and discussion

Participant agreement on image ranking was moderate for all tasks, ranging from $r = 0.45$ for the decade task to $r = 0.68$ for the wealth task.

Overall, the classifier was able to determine which image was being viewed at an above-chance level (33% correct, 95% CI = 30–36%, chance = 1.5%, see Fig. 4). This is not particularly surprising as images varied in the distributions of interest areas (i.e. some images had more faces, others had more bodies, etc.) Additionally,

[1] All reported analyses were also performed using correlational methods (Haxby et al., 2001) as well as a linear support vector machine (http://sourceforge.net/projects/svm/). Task prediction was performed at 26.7% correct using correlation methods ($p = 0.19$), and 26.9% correct using SVM ($p = 0.15$).

Fig. 3. Sample scan paths from 10 s (top) and 60 s (bottom) viewing. Tasks are: memory, decade, people, wealth (left to right, top row) and wealth, memory, decade, people (left to right, bottom row).

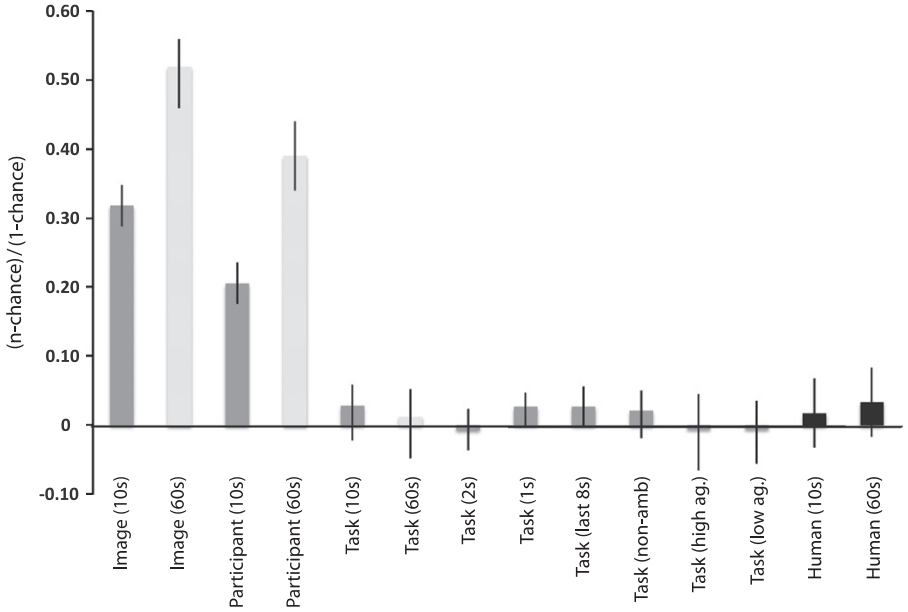

Fig. 4. Overall performance of classifiers and human observers on predicting viewer, image and task. Error bars represent 95% confidence intervals on the binomial test using the procedure described in Clopper and Pearson (1934). The *y*-axis ranges from 0 (chance) to 1 (perfect classification).

Table 1
Percentage of samples used as support vectors for each classifier used in Experiment 1.

Model	Percentage of samples used as support vectors (%)
Image	2
Participant	93
Task	86

the classifier was able to determine which participant was viewing an image (26% correct, 95% CI = 23–28%, chance = 6.3%). As different observers are known to have idiosyncratic patterns of eye movements (Andrews & Coppola, 1999; Boot, Becic, & Kramer, 2009; Castelhano & Henderson, 2008; Underwood, Foulsham, & Humphrey, 2009), this was also expected. However, the classifier was unable to determine which task observers were performing (27.1% correct, 95% CI = 24–31%, chance = 25%, see Fig. 4).[2] Table 1 shows the percentage of samples used as support vectors in each of these models, reflecting the relative difficulty of each learning problem. The super majority of data points were used as support vectors for both task and participant prediction. This could reflect possible over-fitting in the case of participant prediction. For task prediction, the large percentage of samples used as support vectors likely indicates that no discernable pattern existed to classify task, as classifier performance was at chance.

Was the poor performance of the classifier driven by a subset of the tasks? Fig. 5 (upper left panel) shows the confusion matrix from the pattern classifier. Perfect performance would be represented as red along the diagonals and dark blue elsewhere. We can see that the *people* task had the best classification performance

[2] SVM results: 26.9% correct (95% CI = 24–30%, $p = 0.15$). Correlation results: 26.7% correct (95% CI = 24–30%, $p = 0.19$).

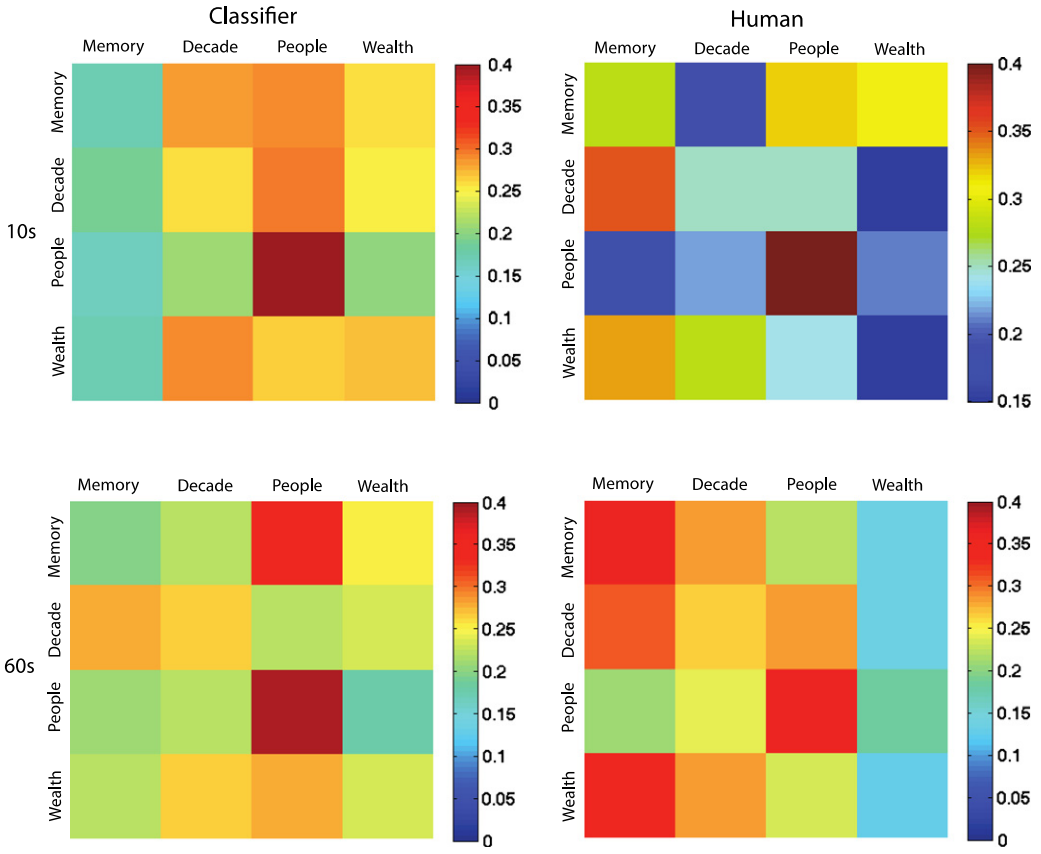

Fig. 5. Confusion matrices between the four experimental tasks for classifiers (left) and human observers (right), for 10 s viewing (top) and 60 s viewing (bottom).

(42% correct), and *memory* had the least (18% correct). In fact, a set of eye patterns from a participant performing the *memory* task was more likely to be classified as one of the other three tasks than to be correctly classified as *memory*.

Why did the classifier fail? As the perception of a scene's category is incredibly rapid (Potter, 1976), as is the perception of a scene's emotional context (Maljkovic & Martini, 2005), perhaps the most diagnostic task information came from the earliest epochs, with idiosyncratic eye movements following. To test this hypothesis, we trained and tested the classifier on the same eye movement features, but restricted analysis to the first 1 or 2 s of each trial.

We found that the classifier was still at chance performance at predicting task using either 1 (26.9% correct, 95% CI = 24–30%),[3] or 2 s (24.5% correct, 95% CI = 22–27%)[4] of information, suggesting that the failure of the classifier to predict observers' task is not due to the lengthy viewing duration of the images.

Alternatively, perhaps the earliest fixations are driven by saliency or general interest, and task-related fixations are made *later* in viewing (e.g. Ballard, Hayhoe, & Pelz, 1995; Mannan, Ruddock, & Wooding, 1997). To test this idea, we trained and tested the classifier on the last 8 s of viewing (excluding the first 2 s). This did not improve performance. Again, the classifier was at chance when predicting task (27.4% correct, 95% CI = 24–30%).[5]

Perhaps eye movement patterns reliably differ across tasks, but not in the same way across images. This could account for the striking difference between the Yarbus result and our own. To test this account, we trained and tested the classifier on each image individually. If image variability explains our poor classification performance, then performance should be above chance in this analysis. This was not the case – the classifier was not able to classify the tasks of any of the 64 images at an above-chance level (range 13–38%).[6]

Despite the compelling nature of the Yarbus (1967) figure, our results indicate that an observer's task cannot be predicted from summary statistics of eye movement patterns from viewing an image for 10 s. This negative result is not due to the insufficiency of these features for classification, or to the inadequacy of the classifier since this technique could successfully predict which image was being viewed and which observer was viewing an image. Nor is the classification failure due to the observers being given too long a glance at the images as reducing the analyzed time from 10 to 1 or 2 s did not ameliorate the performance. Last, classification failure was not due to image variability as testing each image individually yielded the same pattern. Although task prediction for all images was at chance, some images were trending towards significant classification performance. In Experiment 2, we examine the extent to which participant agreement about the information being obtained in the tasks influences classification performance.

3. Experiment 2

In Experiment 1, our pattern classifier failed to predict the observers' tasks from their patterns of eye movements. Perhaps not all of the 64 images used in Experiment 1 contained useful information for each of the four tasks. Conceivably observers have

[3] SVM results: 26.9% correct (95% CI = 24–30%, $p = 0.15$). Correlation results: 23.5% correct (95% CI = 21–26%, $p = 0.30$).

[4] SVM results: 24.8% correct (95%CI = 22–28%%, $p = 0.91$). Correlation results: 24.2% correct (95% CI = 22–27%%, $p = 0.59$).

[5] SVM results: 23.9% correct (95% CI = 21–27%, $p = 0.43$). Correlation results: 24.5% correct (95% CI = 22–27%, $p = 0.77$).

[6] There was not enough data to test LD classifier. We are reporting SVM results for this analysis.

eye movement patterns that are characteristic of the viewing instructions, but only for some subset of images for which the task makes sense. If observers do not agree on, for example, how well the people in a picture know each other, does this lead to higher or lower classification performance? Although participants ranked images after viewing in Experiment 1, we lack the power to quantify subject agreement from these data. Furthermore, we wanted to know if it does not make sense to rank some tasks for some images. Here, we employed a ranking procedure to determine how much agreement there was between subjects for the three non-memory tasks for each of the 64 images.

3.1. Methods

3.1.1. Observers

Eight observers took part in Experiment 2. All were between the ages of 18–55, had normal or corrected-to-normal vision and had no history of eye or muscle disorders. All provided informed consent and were compensated $10/h for their time. None of the eight participants took part in Experiment 1.

3.1.2. Design and procedure

Participants viewed all 64 scenes in random order. Participants ranked each image along each of the three non-memory tasks (people, wealth and decade) on a continuous 1–6 scale using a sliding bar for each of the tasks. The interface also provided a "?" button alongside each of the sliders that observers could push if a task was too ambiguous for a particular image. Participants were given unlimited time to perform this task.

3.2. Results

Twenty-seven of the images received a "?" rating for at least one of the tasks from at least one of the participants. For images never receiving a "?" ranking, participants had a fair degree of agreement in their scores: inter-observer correlations were 0.74, 0.62 and 0.49 for *wealth*, *people* and *decade* tasks, respectively. The three different tasks are not strongly correlated with each other: the *wealth* and *people* tasks ($r = 0.10$), the *wealth* and *decade* tasks ($r = 0.13$) and *people* and *decade* tasks ($r = 0.18$), suggesting an amount of independence between the tasks.

3.2.1. Classification results

First, we trained and tested the classifier on the eye movement data from Experiment 1 using only the 37 images receiving non "?" rankings. All classifier details were the same as in Experiment 1. As in Experiment 1, the classifier could predict the participant, albeit, not well (10% correct, 95% CI = 8–12%, $p < 0.001$, chance level = 6.3%) and the image (38% correct, 95% CI = 35–42%, $p < 0.0001$, chance level = 2.7%). However, as in Experiment 1, the participants' task could not be predicted at an above-chance level (26.6% correct, 95% CI = 23–30%, $p = 0.31$, chance level = 25%).[7]

Next, we wanted to test whether agreement in participants' ranking affected classification performance. Perhaps the classifier would work if its task were limited to images where participants were in the greatest agreement. Alternatively, it is possible that a small degree of ambiguity in the image-task combination would force participants to look harder and longer for the relevant information, making eye movement patterns more distinct for the task. In this case, images with lower ranking agreement would be classified above chance.

The 37 images were divided into two groups of most and least ranking agreement based on the summed variance of ranks for the three tasks. However, neither group supported classification at above chance levels (most agreement: 24.6% correct, 95% CI = 20–30%, $p = 0.95$.[8] Least agreement: 24.6% correct, 95% CI = 20–30%, $p = 0.95$. Chance level = 25%).[9]

3.3. Discussion

Experiment 2 demonstrated that the failure to predict an observer's task from eye movement patterns is not due to poor image choices. Removing images that were rated as ambiguous for a task did not improve classification performance, and classification performance did not depend on the degree of subject agreement on any of the tasks.

4. Experiment 3

In Experiments 1 and 2, the classifier failed to identify a participant's task from eye movement patterns obtained from 10 s of viewing. Perhaps 10 s did not provide enough time for characteristic patterns to develop. In the original Yarbus demonstration, the viewing time was a full 3 min, but with this amount of viewing time, observers tended to examine the same regions over and over "Additional time spent on perception is not used to examine the secondary elements, but to re-examine the most important elements. The impression is created that the perception of a picture is usually composed on a series of 'cycles' each of which has much in common" (Yarbus, 1967, p. 193). In DeAngelus and Pelz (2009) modernization of the Yarbus result, observers were allowed to self-terminate image viewing. Mean viewing time varied according to task (9–50 s). In Tatler et al. (2010), participants viewed a human figure for 50 s. In Experiment 3, we repeated Experiment 1 with a new group of observers, giving them a full 60 s of viewing time on each image.

4.1. Methods

4.1.1. Materials

In order to reduce strain on observers, the number of images used was reduced to 20 images from the 64 images that were used in Experiment 1. These images all received numerical rankings (not "?") in Experiment 2.

4.1.2. Observers

Sixteen observers participated in Experiment 3. All were between the ages of 18–55, had normal or corrected-to-normal vision and had no history of eye or muscle disorders. All provided informed consent and were compensated $10/h for their time. None of these observers had taken part in Experiment 1. Recording difficulties occurred for one observer who was replaced with one additional observer.

4.1.3. Design and procedure

Images in Experiment 3 were viewed for 60 s each. All other experimental details were identical to Experiment 1.

4.2. Results and discussion

As in Experiment 1, we trained classifiers to predict three things: the task of the observer, the image being viewed and the

[7] SVM results: 24.9% correct (95% CI = 22–28%, $p = 0.99$). Correlation results: 25.8% correct (95% CI = 23–27%, $p = 0.61$).

[8] SVM results: 27.7% correct (95% CI = 22–33%, $p = 0.28$). Correlation results: 27.1% correct (95% CI = 22–33%, $p = 0.41$).

[9] SVM results: 22.9% correct (95% CI = 18–28%, $p = 0.45$). Correlation results: 19.8% correct (95% CI = 15–25%, $p < 0.05$).

participant doing the viewing. The classifier had above chance performance in classifying the participant (42.8% correct, 95% CI = 37–48%, $p < 0.0001$, chance = 6.3%) and the image (54.4% correct, 95% CI = 48–60%, $p < 0.0001$, chance = 5%), as in Experiment 1. However, prediction of task still remained at chance (25.9% correct, 95% CI = 21–31%, $p = 0.70$, chance = 25%).[10] Thus, neither 10 s nor 60 s of eye movement information permitted the classifier to identify the observers' task.

5. Experiment 4

Perhaps the classifier's failures in Experiments 1–3 reflect a problem with the classifier and not a lack of information in the eye movements. After all, the classifiers used were relatively simple, and the Yarbus figure (Fig. 1) remains compelling because those scan paths look different to us. Perhaps human observers (still the best pattern classifiers) would succeed where the classifier failed. Perhaps they can predict the task of the previous observers from viewing eye movement scan paths overlaid on the images. We tested this in Experiment 4.

5.1. Methods

5.1.1. Materials

Scan path images were created by plotting a previous observer's eye movements on top of the image being viewed, similar to images shown in Fig. 3. Fixations were plotted as single points, and saccades as lines between the points. As an exercise for the reader, Fig. 3 shows four example images for both viewing times. Try to classify them as memory, decade, people, or wealth.

5.1.2. Observers

Sixteen observers took part in Experiment 4. All were between the ages of 18–55, had normal or corrected-to-normal vision and had no history of eye or muscle disorders. All provided informed consent and were compensated $10/h for their time.

Ten observers were used for the 10-s and 60-s classifications, with four observers participating in both. None of these observers had participated in Experiments 1–3.

5.1.3. Design and procedure

Participants performed 100 trials. In each trial, a sample scan path image was drawn randomly from either the set of 10 s scan paths (Experiment 4a) or 60 s scan paths (Experiment 4b) and presented to the observer. Four screen locations indicated each of the four tasks. The participant clicked on the location representing the task he believed the other participant had been doing. No performance feedback was given.

5.2. Results and discussion

Participants examining the 10 s eye traces were at chance for predicting task (26.3% correct, $t(9) < 1$, $p = 0.63$, chance = 25%). Similarly, participants examining the 60 s eye traces could not predict what task the observer was doing (27.5% correct, $t(9) = 1.75$, $p = 0.11$, chance = 25%).

Confusion matrices for both pattern classifier and human experiments are shown in Fig. 5. For each matrix, we plotted the responses of the observers (classifier or human) against the ground truth for the test image. Correct performance is shown on the diagonals. Warmer colors on the off-diagonals represent pairs of tasks that were frequently confused. Correlations between human and classifier confusion matrices for both 10 s ($r = 0.27$) and 60 s ($r = 0.25$) were modest.

For both human and pattern classifiers, classifying the *people* was the easiest, perhaps because more dwell time was spent on faces in these images. Human observers, particularly when viewing the 60-s scan paths, defaulted to classifying a path as "memory", while the pattern classifier tended to default to classifying a trial as "people". Despite the modest similarity between the classification patterns of the pattern classifiers and the human observers, it is clear that human observers fare no better at classifying another observer's task from eye movements.

6. General discussion

On his well-known figure showing task differences in eye movements, Yarbus wrote "Eye movements reflect the human thought process; so the observer's thought may be followed to some extent from the records of eye movements" (Yarbus, 1967, p. 190). In other words, Yarbus believed that an observer's task could be predicted from his static patterns of eye movements.

In this study, we have sadly failed to find support for the most straight-forward version of this compelling claim. Over a range of observers, images and tasks, static eye movement patterns did not permit human observers or pattern classifiers to predict the task of an observer.

Our failure to predict observers' tasks was not due to classifier choice. Three different pattern classifiers (linear discriminant, correlation and linear support vector machines) all provided the same result. Furthermore, the pattern classifiers (and the features they used) were sufficient to predict the image and the observer at above-chance levels. Even more striking is the failure of human observers to classify the tasks being performed by other observers.

Nor is the failure due to observers being given an inappropriate amount of time to view the images. Pattern classification failed with 1, 2, 10 and 60 s worth of viewing time. Similarly, the human observers did not substantially benefit from seeing 60 s of eye movements as compared to 10 s worth of data.

Finally, the failure does not appear to be due to the choice of images viewed. Experiment 1 demonstrated that no single image could be classified at an above-chance level. Furthermore, examining only images with high or low subject agreement on a task did not improve classification performance.

Although several studies have shown that some eye movement features such as fixation duration or time on a region of interest depend on task (Castelhano, Mack, & Henderson, 2009; Tatler, Baddeley, & Vincent, 2006), to our knowledge, this is the first study that tried to predict task from eye movements.

So why did it fail? How can the Yarbus figure look so compelling to us and yet be so misleading? It may be the case that within a single observer and image, task differences can be found. In our experiments, observers viewed a single image only once. Indeed, Noton and Stark (1971) noted idiosyncratic scanning patterns for particular observers examining particular stimuli. Second, the Yarbus figure might seem compelling because the task labels are shown with the scan paths, and those labels seem reasonable to us as observers in hindsight. Third, eye movements have temporal structure that is not captured in the Yarbus figure or in these experiments. Perhaps examining scan path properties over time would reveal task differences. Fourth, it could be that changes in viewing instructions are not very dramatic for the visual system. Perhaps we would see strong task-dependence in eye movements only if participants needed to use the visual information they are gaining (e.g. Ballard, Hayhoe, & Pelz, 1995). Thus, while the idea that complex cognitive processes can be inferred from a simple

[10] SVM results: 26.8% correct (95% CI = 22–32%%, $p = 0.44$). Correlation results: 29% correct (95% CI = 24–34%, $p = 0.09$, two tailed).

behavior such as eye movements is deeply attractive, it is a deeply attractive idea that has very substantial limitations.

Acknowledgments

This work was funded by F32 EY19815 to M.R.G. and O.N.R. N000141010278 and NEI EY017001 to J.M.W. Thanks to Erica Kreindel for assistance in running participants, Melissa Vo, Chris Baldassano and Marius C. Iordan for insightful discussions on this project, and to Barbara Hidalgo-Sotelo for helpful comments on the manuscript.

References

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research, 39*(17), 2947–2953. http://dx.doi.org/10.1016/S0042-6989(99)00019-X.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7*(1), 66–80. http://dx.doi.org/10.1162/jocn.1995.7.1.66.

Boot, W. R., Becic, E., & Kramer, A. F. (2009). Stable individual differences in search strategy? The effect of task demands and motivational factors on scanning strategy in visual search. *Journal of Vision, 9*(3), 1–16 (article no. 7).

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology, 62*(1), 1–14. http://dx.doi.org/10.1037/1196-1961.62.1.1.

Castelhano, M. S., Mack, M., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*(3), 1–15 (article no. 6).

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*, 404–413.

DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition, 17*(6), 790. http://dx.doi.org/10.1080/13506280902793843.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley-Interscience.

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

Hagemann, N., Schorer, J., Cañal-Bruland, R., Lotz, S., & Strauss, B. (2010). Visual perception in fencing: Do the eye movements of fencers represent their information pickup? *Attention, Perception, and Psychophysics, 72*(8), 2204–2214. http://dx.doi.org/10.3758/APP.72.8.2204.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425–2430. http://dx.doi.org/10.1126/science.1063736.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*(1). http://dx.doi.org/10.1167/3.1.6.

Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition,, 36*(6), 1561–1566. http://dx.doi.org/10.1037/a0020693.

Land, M., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*, 3359–3565.

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.

Maljkovic, V., & Martini, P. (2005). Short-term memory for scenes with affective content. *Journal of Vision, 5*, 215–229.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception, 26*(8), 1059–1072.

Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage, 53*(1), 103–118. http://dx.doi.org/10.1016/j.neuroimage.2010.05.051.

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*(8). http://dx.doi.org/10.1167/11.8.17.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*(9), 424–430. http://dx.doi.org/10.1016/j.tics.2006.07.005.

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research, 11*(9), 929–942. http://dx.doi.org/10.1016/0042-6989(71)90213-6. IN3–IN8.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology-Human Learning and Memory, 2*(5), 509–522.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research, 46*(12), 1857–1862. http://dx.doi.org/10.1016/j.visres.2005.12.005.

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, eye movements, and vision. *i-Perception, 1*(1), 7–27. http://dx.doi.org/10.1068/i0382.

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition, 17*(6–7), 812–834. http://dx.doi.org/10.1080/13506280902771278.

Yarbus, A. L. (1967). *Eye movements and vision*. Springer.