Journal of Experimental Psychology:
Human Perception and Performance
2015, Vol. 41, No. 6, 1502–1514

0096-1523/15/$12.00 http://dx.doi.org/10.1037/a0039673

Classifying Mental States From Eye Movements During Scene Viewing

Omid Kardan and Marc G. Berman
The University of Chicago

Grigori Yourganov and Joseph Schmidt
University of South Carolina

John M. Henderson
University of California, Davis

How eye movements reflect underlying cognitive processes during scene viewing has been a topic of considerable theoretical interest. In this study, we used eye-movement features and their distributions over time to successfully classify mental states as indexed by the behavioral task performed by participants. We recorded eye movements from 72 participants performing 3 scene-viewing tasks: visual search, scene memorization, and aesthetic preference. To classify these tasks, we used statistical features (mean, standard deviation, and skewness) of fixation durations and saccade amplitudes, as well as the total number of fixations. The same set of visual stimuli was used in all tasks to exclude the possibility that different salient scene features influenced eye movements across tasks. All of the tested classification algorithms were successful in predicting the task within a single participant. The linear discriminant algorithm was also successful in predicting the task for each participant when the training data came from other participants, suggesting some generalizability across participants. The number of fixations contributed most to task classification; however, the remaining features and, in particular, their covariance provided important task-specific information. These results provide evidence on how participants perform different visual tasks. In the visual search task, for example, participants exhibited more variance and skewness in fixation durations and saccade amplitudes, but also showed heightened correlation between fixation durations and the variance in fixation durations. In summary, these results point to the possibility that eye-movement features and their distributional properties can be used to classify mental states both within and across individuals.

Keywords: eye movements, scene viewing, multivariate analysis, classification, linear discriminant

Whenever we perform a visual task or explore a scene, our eyes move in a series of fixations and saccades (Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013; Rayner, 1998). Fixations are brief periods of time in which the high-acuity fovea settles on the features of interest. Saccades are high-velocity movements that step fixations from one object to another through a scene. Because eye movements are critical to scene understanding, the nature of the processes that control these movements has been a focus of intense research (see Henderson, 2011).

Systematic differences in eye movements across tasks shed light on the underlying cognitive operations involved in each task and on changes in the allocation of attention used to perform each task (Henderson, 2011; Rayner, 2009). For example, in a classic demonstration of the relationship between eye movements and cognition, Yarbus (1967) asked a viewer to examine *The Unexpected Visitor*, a painting by Ilya Repin depicting the homecoming of a political prisoner. The viewer was asked to look at the picture for a variety of purposes, and Yarbus showed that the viewer's eye movements changed systematically depending on the viewing task. For example, when the viewer was asked to determine the ages of the people in the painting, she concentrated her fixations on the faces in the scene, but when she was asked to determine the material circumstances of the family, she directed her eyes more generally over the objects. In this way, Yarbus demonstrated that cognitive processes and internal states of the viewer influence eye movements.

The influence of cognitive processes on eye movements during scene viewing was also shown by Buswell (1935), and has been demonstrated many times since (Borji & Itti, 2014; Castelhano & Henderson, 2008; Henderson, 2003; Henderson, Weeks, & Hollingworth, 1999; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011; Tatler, 2009). That is, eye-movement patterns differ as a function of the viewing task and other changes in cognitive state. The question then arises: Is the inverse also true? Can cognitive states be predicted from differences in eye-movement patterns? More specifically, can we classify cognitive states from the descriptive statistics of fixations and saccades? If it

This article was published Online First September 7, 2015.

Omid Kardan and Marc G. Berman, Department of Psychology, The University of Chicago; Grigori Yourganov and Joseph Schmidt, Department of Psychology, University of South Carolina; John M. Henderson, Center for Mind and Brain, Department of Psychology, University of California, Davis.

Joseph Schmidt is now at Department of Psychology, University of Central Florida.

This work was supported in part by a grant from the TKF Foundation to Marc G. Berman, an internal grant from the Department of Psychology at the University of South Carolina to Marc G. Berman, and National Science Foundation Grant BCS-1151358 to John M. Henderson.

Correspondence concerning this article should be addressed to Marc G. Berman, Department of Psychology, The University of Chicago, 5848 South University, Chicago, IL 60637, E-mail: bermanm@uchicago.edu or John M. Henderson, Center for Mind and Brain, Department of Psychology, University of California, Davis, 267 Cousteau Place, Davis, CA 95618. E-mail: johnhenderson@ucdavis.edu

is the case that eye movements generally reflect the cognitive processes active during a given cognitive task, then the prediction is that such classification should be possible. Furthermore, we can ask whether eye-movement differences across tasks are consistent enough that we can classify the task that one participant is engaged in from other participants' eye-movement data. If the mapping of cognitive processes to eye-movement behavior is generalizable across people, as it should be if this mapping is lawful, then the prediction is that it should be possible to recover one viewer's task from the similarity of their eye movements to those of other people engaged in the same task.

Three recent studies (Borji & Itti, 2014; Greene, Liu, & Wolfe, 2012; Henderson et al., 2013) have investigated this question using classification analysis. Classification analysis is an alternative to hypothesis testing; rather than estimating the significance of the task-driven differences for each statistical feature of eye movements, a classification analysis uses a combination of such features to (a) train a classifier to classify the task, and (b) test this classification on data that have not been used during training (Rosa, 2010). The accuracy of predicting the task for previously unseen data is a measure of the replicability of the effect; this is one advantage of classification analysis over null-hypothesis significance testing, which does not evaluate replicability (Cohen, 1994).

In the present study using classification analysis, we set out to examine several eye-movement phenomena that have theoretical and practical significance in terms of how individuals perform various visual tasks. The first goal of this study was to determine whether eye-movement features could be used to predict visual task performance, even when the same stimuli are used for each visual task. Henderson et al. (2013) demonstrated that one can distinguish visual search, scene memorization, reading, and pseudoreading from each other based on the statistics of eye movements. However, the stimuli for several of the tasks differed (although they were the same for the two scene-viewing tasks), suggesting that differing scene features may have contributed to classification accuracy. A question, then, from the Henderson et al. (2013) study is whether similar classification accuracy can be observed when the visual stimuli are held constant across several tasks and only the task varies, as in the Greene et al. (2012) study. To address this question, the current study used the same scene stimuli in all three tasks (scene memorization, visual search, and aesthetic preference).

A second goal of this study was to examine whether individuals perform different visual tasks in idiosyncratic ways, or whether there are generalizable eye-movement patterns that characterize how all people perform different visual tasks. To test this question, we performed two levels of classification analysis: within-subject classification (in which classifiers were trained on and tested within a subject) and between-subjects classification (in which classifiers were trained on a subset of subjects and tested on another). Results from this analysis have theoretical implications for understanding the cognitive control of overt attention. If differences in the manner in which attention is allocated over a scene as a function of task are functionally related to information processing in that task, then we would expect these differences to be consistent across individuals. In that case, classification based on training data from one set of individuals should generalize to a new set of individuals.

A third goal of this study was to examine whether classification accuracy would differ depending on the nature of the classifier. Differences across classifiers provide information about the best model for visual task prediction in the context of our chosen visual tasks. This helps to inform how different eye-movement features are combined across tasks, providing information about how attentional deployment changes across tasks. For this purpose, we used four classifiers that make different assumptions about the populations from which the data are sampled. The linear Gaussian naive Bayesian classifier (GNB-L) and the nonlinear Gaussian naive Bayesian classifier (GNB-N) assume that the statistical features of eye movements are independent of each other. The linear discriminant (LD) and the quadratic discriminant (QD) are multivariate classifiers and model the covariance between features. In addition, GNB-L and LD make the assumptions of linearity, whereas the GNB-N and QD classifiers are nonlinear classifiers and do not impose the assumption of linearity on the data. It is also of note that the success of each classifier in predicting mental states may provide information about the nature of eye movements for these different tasks. For example, eye-movement features may correlate differentially with each other for different tasks, and some of these classifiers would capitalize on this feature to improve classification accuracy. Such information provides additional evidence about how eye-movement features are related to different visual tasks.

Lastly, we examined the relative importance of each eye-movement feature to successful classification by excluding that feature and measuring any decrease in classification accuracy. This analysis also provides information about how differences in cognitive processes can be reflected in specific (isolated) variables that describe the shape of distribution in the eye-movement behaviors (mean, standard deviation, and skewness, which are first, second, and third moments of the distribution, respectively). Our hypothesis was that differences in the strategies to achieve the goals of these visual tasks would be reflected in their overt attentional deployment, and those variables that summarize the shape of the distribution of fixation durations and saccade amplitudes during each task should suffice to reliably distinguish the tasks from each other. In summary, we set out to determine whether cognitive states, operationalized as the viewing task type during scene viewing, can be predicted from eye-movement data within participants and across participants. Our results have theoretical implications as to how cognitive processes control overt attentional deployment in complex tasks and how specific features of cognitive control generalize across viewers.

Method

Participants

Seventy-two Edinburgh, Scotland, undergraduate students with normal or corrected-to-normal vision participated in the experiment as part of a large eye-movement corpus study. All participants were naive concerning the purposes of the experiment and provided informed consent.

Apparatus

Eye movements were recorded via an SR Research Eyelink 1000 eye tracker (spatial resolution of 0.01°) with a sampling rate

of 1000 Hz. Participants were seated 90 cm away from a 21-in. CRT monitor. Head movements were minimized with chin and head rests. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with SR Research Experiment Builder software.

Stimuli and Tasks

We used 135 unique full-color 800 × 600 pixel (32 bit) photographs of real-world scenes from a variety of indoor and outdoor scene categories in the experiment. The scenes were split into blocks of 45 images, and participants were instructed to perform one of three tasks during each block: search for an object (specified by a word cue), memorize the scene in preparation for a later memory test, or provide an aesthetic judgment (4-point scale, 1 = *dislike* to 4 = *like*). All scenes were presented for 8 s. During the search block, search responses were logged during the trial but the scene remained visible after the response until the full 8 s had elapsed. During the aesthetic preference task, a preference judgment was made after each scene's presentation. Memorization performance was assessed with a separate memory test following all three main tasks.

Classification Analysis

Features. For each trial, we computed seven features from the eye-movement data: Three described the distribution of fixation durations, three described the distribution of saccade amplitudes, and one coded the number of fixations. Specifically, we used the mean, standard deviation, and skewness of the fixation durations; the mean, standard deviation, and skewness of the saccade amplitudes; and the number of fixations per image. We used the Matlab functions mean, stdev, and skewness to compute these features.

Classifiers. Throughout this study, we used four distinct classifiers: LD, QD, GNB-L, and GNB-N. Implementation of the classifiers was provided by the classify function in the Statistics toolbox in Matlab (the classifier type was set to linear, quadratic, diagLinear, and diagQuadratic, respectively). These classifiers use a multivariate Gaussian distribution to model the classes and classify a vector by assigning it to the most probable class. The LD classification model contains an assumption of homoscedasticity; that is, all classes are sampled from populations with the same covariance matrix. For our purposes, this assumption means that (a) the variance of each feature does not change across tasks, and (b) the covariance between each pair of features is the same for all tasks. QD makes no such assumption, and instead estimates the covariance matrices separately for each class (that is, the variances of and the covariances between features are allowed to differ across tasks). GNB classifiers impose a constraint that the covariance matrices are diagonal (in our case, this implies that the eye-movement features are uncorrelated); furthermore, GNB-L makes a further assumption of homoscedasticity, whereas GNB-N does not make this assumption (that is, GNB-L assumes that the variance of each feature is the same across tasks, and GNB-N allows it to differ). Given that the covariance between features is ignored by both versions of GNB, these two classifiers are univariate, whereas LD and QD are multivariate classifiers because they use the feature's covariance to classify the task. Another distinction between linear and nonlinear classifiers is that the assumption of homoscedasticity is equivalent to separating the classes with a linear plane; otherwise, the classes are separated with a nonlinear curved surface (therefore, both QD and GNB-N are nonlinear classifiers, whereas LD and GNB-L are linear).

Cross-validation procedure. All classifiers were evaluated using a cross-validation approach. A subset of trials was used to train the classifier, and the task was predicted for the trials that were not included in the training set. We used two approaches for separating the data into training and test sets. The first approach used a within-participant classification, in which training and testing were performed on data within the same participant in an iterative fashion. In this analysis, 10% of the trials (13 trials) were randomly left out for testing. We produced 1,000 unique training and test sets for each participant in which the classifier was trained on 90% of the trials and tested on the remaining left-out 10% of the trials. This was done for each of the 1,000 unique training and test set combinations. The percentages of correctly classified tasks in each of the 13 left-out trials was calculated in all 1,000 training–test set combinations and were then averaged for each participant to obtain within-subject accuracy.

The second approach used an across-participant classification in which all trials for a particular participant were iteratively tested using all of the trials from the remaining 71 participants for training. This process was iterated until all trials for all participants had been tested.

Evaluation of classifiers. For both within- and across-participants classification procedures, we computed, for each participant, the proportion of trials when the task was predicted accurately; this proportion was our measure of classification accuracy for a given participant. In addition, we computed the confusion matrices: The (*i*th, *j*th) cell of the confusion matrix specifies the proportion of trials that were recorded while performing task *i* and were assigned to task *j* by a classifier. The diagonal values of the confusion matrix correspond to the proportion of trials in which the classifier correctly predicted the task, whereas the nondiagonal values correspond to the proportion of trials incorrectly classified as another task.

Feature loadings. The relative importance of each of the seven features (the unique contribution of each feature) was evaluated by removing a feature from all participants' data and computing the difference in task classification accuracy relative to classification using the full set of seven features. This difference was computed for all 72 participants and for both within-participant and across-participants classification.

Results

Seven eye-movement features were computed for each trial. Table 1 lists the mean value and the associated standard error of each feature for each of the three tasks across participants. Simple paired *t* tests showed a significantly greater standard deviation of fixation durations for the visual search task than the scene memorization, $t(71) = 6.64$, $p < .001$, and the aesthetic preference tasks, $t(71) = 8.14$, $p < .001$, and also a greater standard deviation of fixation durations in the scene memorization task compared with the aesthetic preference task, $t(71) = 2.78$, $p = .007$. The skewness of fixation durations in the visual search task was significantly greater than those of

Table 1
Features of Eye Movements Used to Classify Task Performed During the Trial

Feature	Task		
	Visual search	Scene memorization	Aesthetic preference
Mean fixation duration	256.59 ± 3.28	256.88 ± 3.35	251.43 ± 3.77
Average standard deviation of fixation durations	132.75 ± 2.12	117.46 ± 2.17	113.52 ± 2.36
Average skewness of fixation durations	1.14 ± 0.03	1.01 ± 0.03	1.00 ± 0.03
Number of fixations	22.42 ± 0.48	25.02 ± 0.35	27.47 ± 0.56
Mean saccade amplitude	4.46 ± 0.09	4.63 ± 0.07	4.85 ± 0.08
Average standard deviation of saccade amplitudes	3.88 ± 0.06	3.65 ± 0.05	3.69 ± 0.06
Average skewness of saccade amplitudes	1.19 ± 0.03	1.05 ± 0.02	1.00 ± 0.02

Note. Values are means ± standard errors ($df = 71$) across participants computed for each task. Fixation durations were measured in milliseconds, and saccade amplitudes were measured in degrees of visual angle. The numbers of trials were 3,132 for visual search, 3,239 for scene memorization, and 3,131 for aesthetic preference. The first fixation in each trial and fixations that were above or below the mean ± 3.5 *SD* of fixations in all trials ($n = 247{,}017$) were removed from data. The features were calculated in each trial (8 s) in each task and then averaged over all trials in that task for a participant. Then the means and standard errors across participants were calculated.

both the scene memorization task, $t(71) = 3.17$, $p = .002$, and the aesthetic preference task, $t(71) = 3.40$, $p = .001$. As for the number of fixations, visual search had fewer fixations than both scene memorization, $t(71) = -6.07$, $p < .001$, and aesthetic preference, $t(71) = -9.30$, $p < .001$. Also, scene memorization had fewer fixations than aesthetic preference, $t(71) = -6.98$, $p < .001$. In terms of saccade amplitudes, mean saccade amplitude for visual search was smaller than those for both scene memorization, $t(71) = -2.78$, $p = .007$, and aesthetic preference, $t(71) = -6.38$, $p < .001$, and also was smaller in the scene memorization task compared with the aesthetic preference task, $t(71) = -5.47$, $p < .001$. The standard deviation of saccades in visual search trials was larger than those in both scene memorization, $t(71) = 4.81$, $p < .001$, and aesthetic preference, $t(71) = 3.69$, $p < .001$. Finally, the skewness of saccades was larger in visual search compared with both scene memorization, $t(71) = 5.22$, $p < .001$, and aesthetic preference, $t(71) = 6.15$, $p < .001$, and skewness of saccades in memorization was marginally larger than that in aesthetic preference, $t(71) = 2.02$, $p = .047$. To see how these variables that describe the shape of the distribution of fixation durations and saccade amplitudes are distinguishable for different tasks, one can inspect the visualizations of these distributions across tasks, which are plotted in Figure 1.

In addition, we computed the correlations between features, which are displayed in Figure 2. The correlation matrix shows a clear grouping, that is, the features of saccade amplitudes were relatively uncorrelated with the features of fixation durations (Castelhano & Henderson, 2008; Henderson & Luke, 2014). The number of fixations was negatively correlated with mean fixation duration, as long fixations would reduce the total number of possible fixations in the fixed length trials. The correlations displayed in Figure 2 were computed by pooling the trials for all three tasks. The same general pattern of correlations was observed in all tasks, but a small set of correlations differed by task. Significantly different correlations were assessed with paired *t* tests, computed using Fisher's *r*-to-*z* transformation, with the threshold of significance set to .002 (.05/21) after adjustment for Bonferroni correction for multiple comparisons. The correlation between mean and standard deviation of fixation durations was significantly larger in the visual search task than in the aesthetic preference task, $z(69) = 3.89$, $p < .001$, and was also larger in the visual search task than the scene memorization task, $z(69) = 2.34$, $p = .019$, $df = 69$, but did not survive the Bonferroni-corrected threshold.

Classification Results

The fact that some features were strongly correlated (see Figure 2) suggests that the multivariate classifiers may outperform the univariate classifiers because a multivariate model incorporates the covariance between features, whereas the univariate classifiers assume that the features are independent. However, it is not immediately clear whether a linear or a nonlinear multivariate model should perform better. The linear (i.e., homoscedastic) model ignores the task-driven differences in correlations mentioned above (i.e., it assumes that the eye-movement features covary with each other in the same way for each task type). A nonlinear multivariate classifier, such as QD, models the covariances between features separately for each task. This increases the number of model parameters by one and, therefore, requires a larger amount of training data relative to LD. Therefore, QD is expected to outperform LD only if the feature interactions are sufficiently different across tasks and a sufficient amount of data is available for training (Seber, 2004).

The accuracy of the four classifiers applied to the within-participant classification is shown in Figure 3, where a box-and-whisker plot summarizes the distribution of each classifier's accuracy over the 72 participants. The plot was created using the Matlab boxplot function; any point outside the $[q_1 - 1.5^*(q_3 - q_1), q_3 + 1.5^*(q_3 - q_1)]$ range, where q_1 and q_3 are the 25% and 75% percentiles, respectively, was considered an outlier. The asterisk-marked lines that connect classification models indicate significant differences in the performance of the two corresponding classifiers, as estimated by the nonparametric Wilcoxon signed-ranks test with the threshold of significance set to a Bonferroni-corrected threshold of .008. For the use of Wilcoxon's tests in comparative evaluation of classifiers,

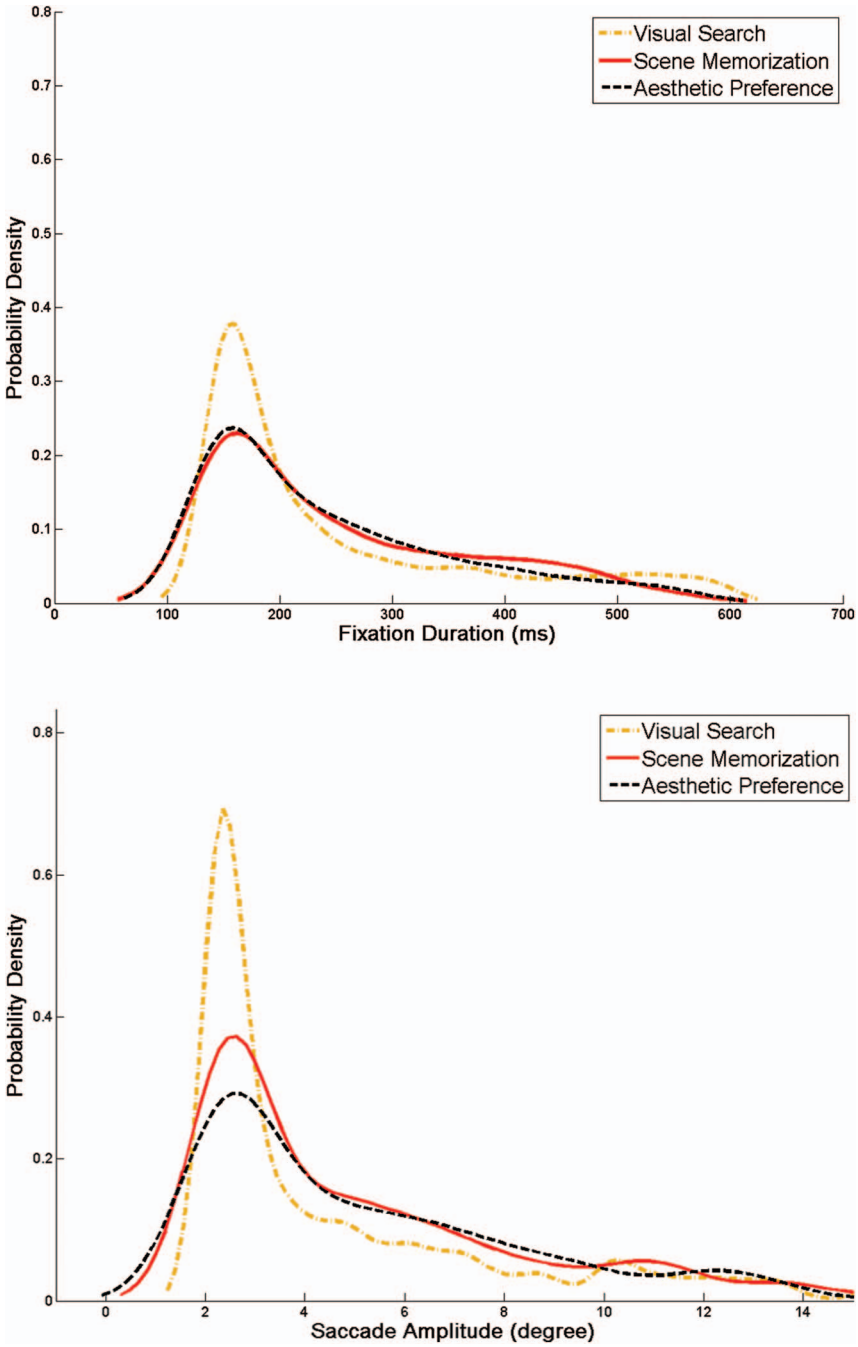

Figure 1. Comparison of the shapes of the probability distributions of fixation durations (top) and saccade amplitudes (bottom) for visual search, scene memorization, and aesthetic preference tasks. (Top) Greater standard deviation of fixation durations and skewness of fixation durations are observed in the visual search task compared with the scene memorization and the aesthetic preference tasks. Greater standard deviation of fixation durations is also observed in the scene memorization task compared with the aesthetic preference task. (Bottom) Mean saccade amplitudes for the visual search task are smaller than those for both scene memorization and aesthetic preference tasks. Mean saccade amplitude is also smaller in the scene memorization task compared with the aesthetic preference task. The standard deviation of saccade amplitudes and the skewness of saccade amplitudes in the visual search task are larger than those in both the scene memorization and aesthetic preference tasks. See the online article for the color version of this figure.

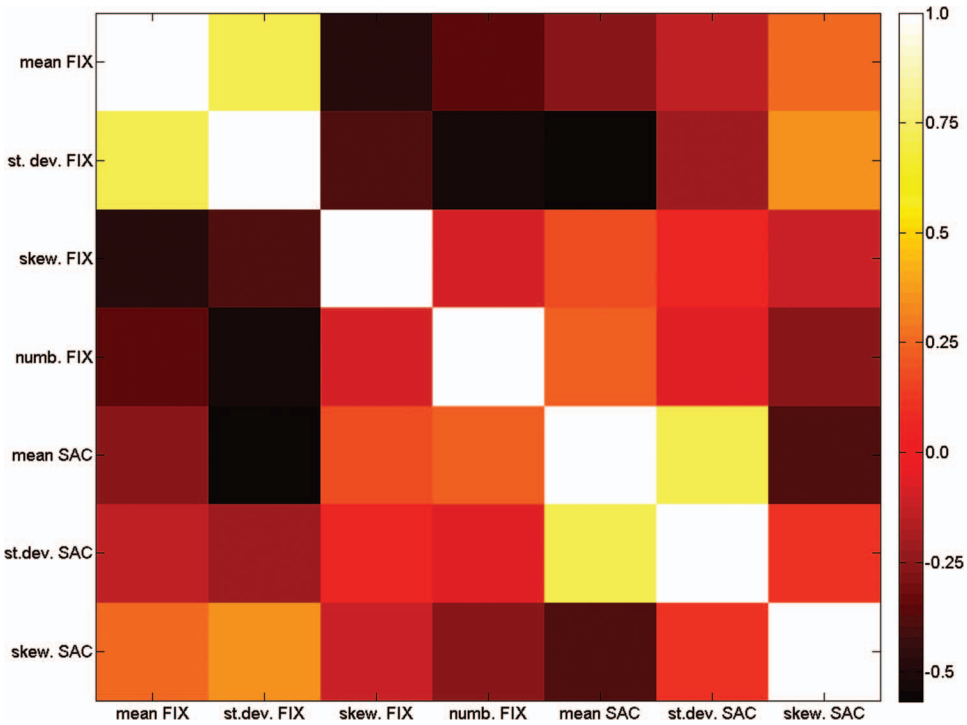

Figure 2. Correlation matrix exhibiting the correlation between all seven features of the eye movements. See the online article for the color version of this figure.

see Demsar (2006). All four classifiers accurately classified the task types. However, the multivariate classifiers (LD and QD) consistently outperformed the two univariate GNB classifiers, suggesting that the covariance between eye-movement features is particularly important for correct within-participant task classification. The mean within-subject classification accuracies across participants were 66.4% for LD, $t(71) = 23.3$, $p < .001$ (compared with chance note: chance is 33.3%), 65.5% for QD, $t(71) = 20.7$, $p < .001$ (compared with chance), 61.2% for GNB-L, $t(71) = 23.2$, $p < .001$ (compared with chance), and 59.1% for GNB-N, $t(71) = 20.6$, $p < .001$ (compared with chance). All classifiers showed above-chance levels of accuracy (as indicated by the dashed line in the top panel of Figure 3) in both median level of classification accuracy and classification accuracy of each participant (i.e., the error bars do not cross the chance line, indicating that all participants were classified with above-chance accuracy). The high within-participant task classification accuracy is also reflected in the confusion matrices (see Figure 3, bottom panel). In each confusion matrix, the highest value for each row is on the diagonal, indicating that for each task the number of correct classifications is higher than the number of mistakes.

Figure 4 displays the task classification accuracy across participants. Median classification accuracy was again above chance for all classifiers (55.6% for LD, 45.8% for QD, 53.3% for GNB-L, 45.5% for GNB-N), suggesting a reasonable degree of overlap in the task-specific eye-movement characteristics across participants. However, for all classifiers, across-participant classification was significantly less accurate than within-participant classification ($p < .001$, Wilcoxon signed-ranks test). Nevertheless, the LD classifier did provide above-chance across-participant classification in all participants. For the three remaining classifiers, classification was at or below chance for a small subset of participants (five participants for QD, two for GNB-L, and six for GNB-N), and yet the means of accuracies across subjects were significantly more than 0.33 for all classifiers ($df = 71$, $p < .001$, for all classifiers). A Wilcoxon signed-ranks test showed that LD was significantly more accurate than all other classifiers ($df = 71$, $p < .001$, for all classifiers), suggesting that the homoscedastic model used by LD (which accounts for covariance between the features, but ignores the task-driven differences in the covariance between features) most accurately captures the task-specific information that generalizes across participants.

In addition, GNB-L was significantly more accurate than both the QD, $t(71) = 9.91$ $p < .001$, and GNB-N, $t(71) = 9.54$ $p < .001$, which were not significantly different from each other, $t(71) = 1.18$, $p = .238$. The confusion matrices for the two linear classifiers (LD and GNB-L) have high diagonal entries, meaning that each task is likely to be accurately predicted. This is not the case for the two nonlinear classifiers; only the third task (aesthetic judgment) is likely to be predicted correctly by the nonlinear classifiers.

Finally, we evaluated the contribution of each feature to classification accuracy using the LD classifier (see Figure 5). The number of fixations contributed the most to classification accuracy: After its removal, the mean classification accuracy dropped by 8.85% (within-participant) and 4.42% (across-participants). In some participants, the drop in accuracy was as high as 23%. The

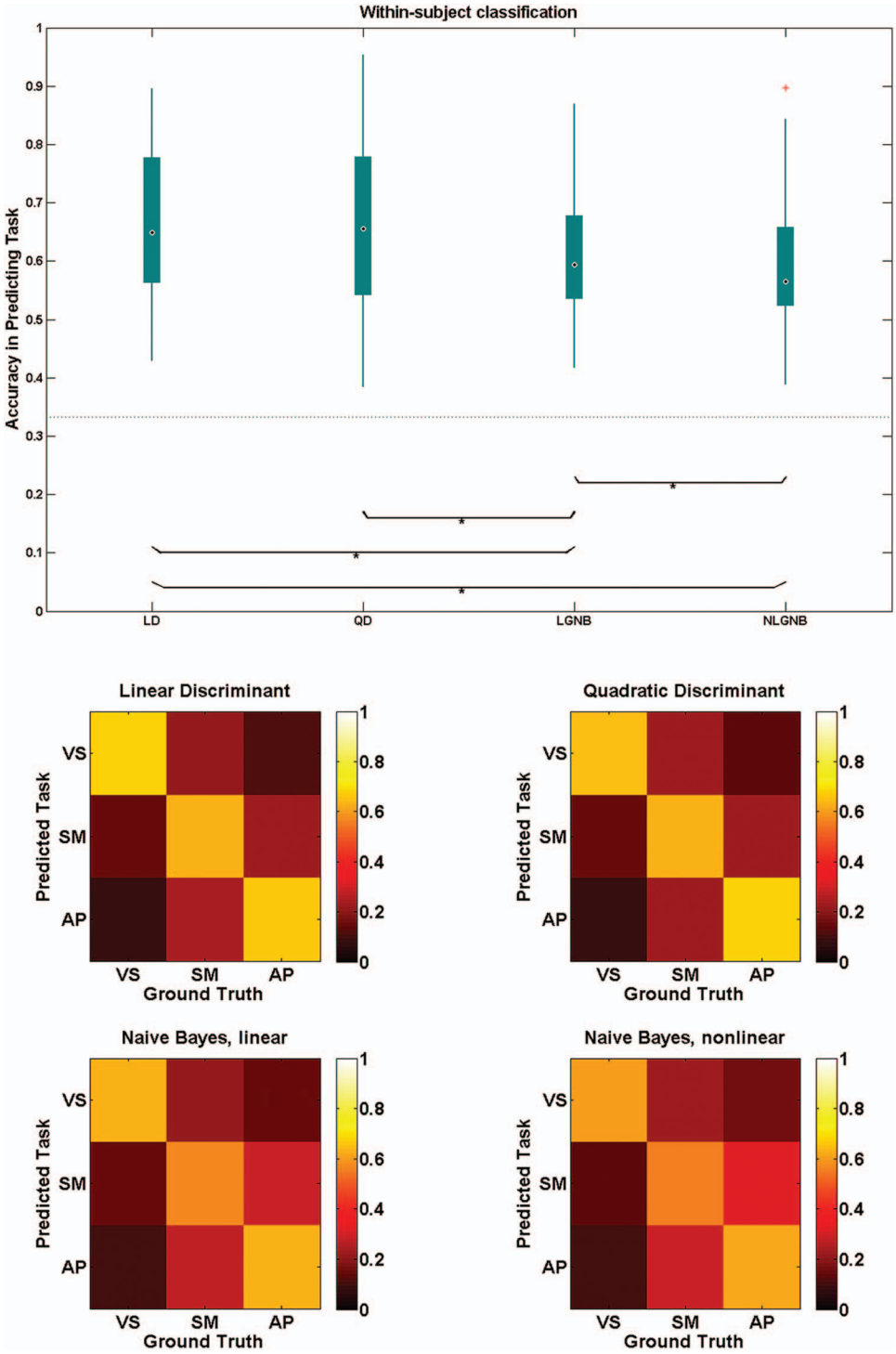

Figure 3. Performance of the four classifiers in within-participant classification. (Top) Accuracy of each classifier when predicting the task. From left to right, LD = linear discriminant, QD = quadratic discriminant, LGNB = linear Gaussian naive Bayes, NLGNB = nonlinear Gaussian naive Bayes classifiers. The dot represents the median classification performance of all participants. The box represents the middle two quartiles of classification performance, and the whiskers represent the full range of classification performance (outliers are marked with a +). The dashed line indicates the level of chance classification (0.333 . . .). The horizontal lines marked with an asterisk indicate a significant difference in performance between pairs of classifiers. (Bottom) The corresponding confusion matrices. The tasks are visual search (VS), scene memorization (SM), and aesthetic preference (AP). See the online article for the color version of this figure.

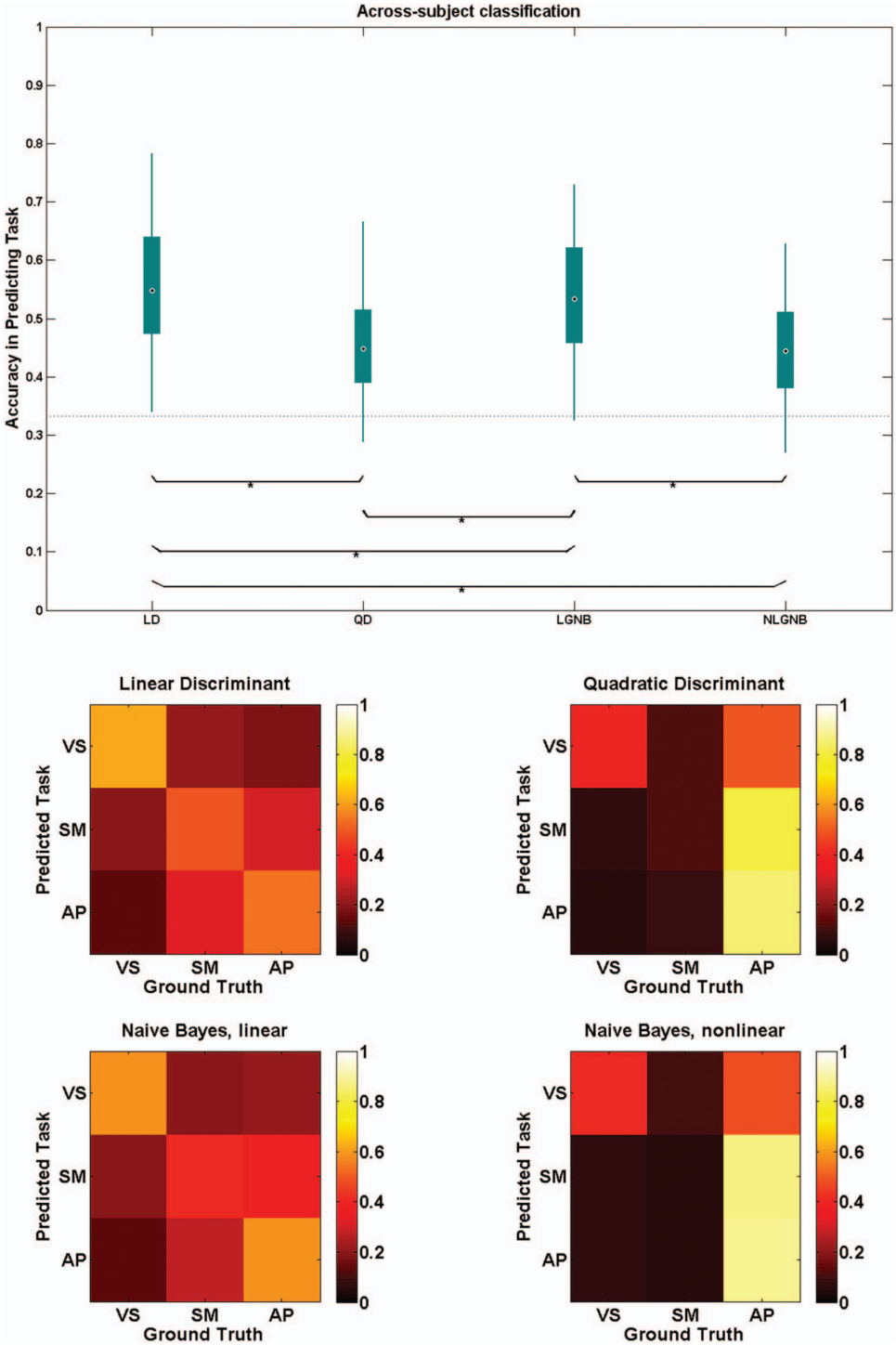

Figure 4. Performance of the four classifiers in across-participant classification. (Top) Accuracy of each classifier when predicting the task. The dot represents the median classification performance across participants. The box represents the middle two quartiles of classification performance, and the whiskers represent the full range of classification performance. From left to right, LD = linear discriminant, QD = quadratic discriminant, LGNB = linear Gaussian naive Bayes, NLGNB = nonlinear Gaussian naive Bayes classifiers. The dashed line indicates the level of chance. The horizontal lines marked with an asterisk indicate a significant difference in performance between pairs of classifiers. (Bottom) The corresponding confusion matrices. The tasks are visual search (VS), scene memorization (SM), and aesthetic preference (AP). See the online article for the color version of this figure.

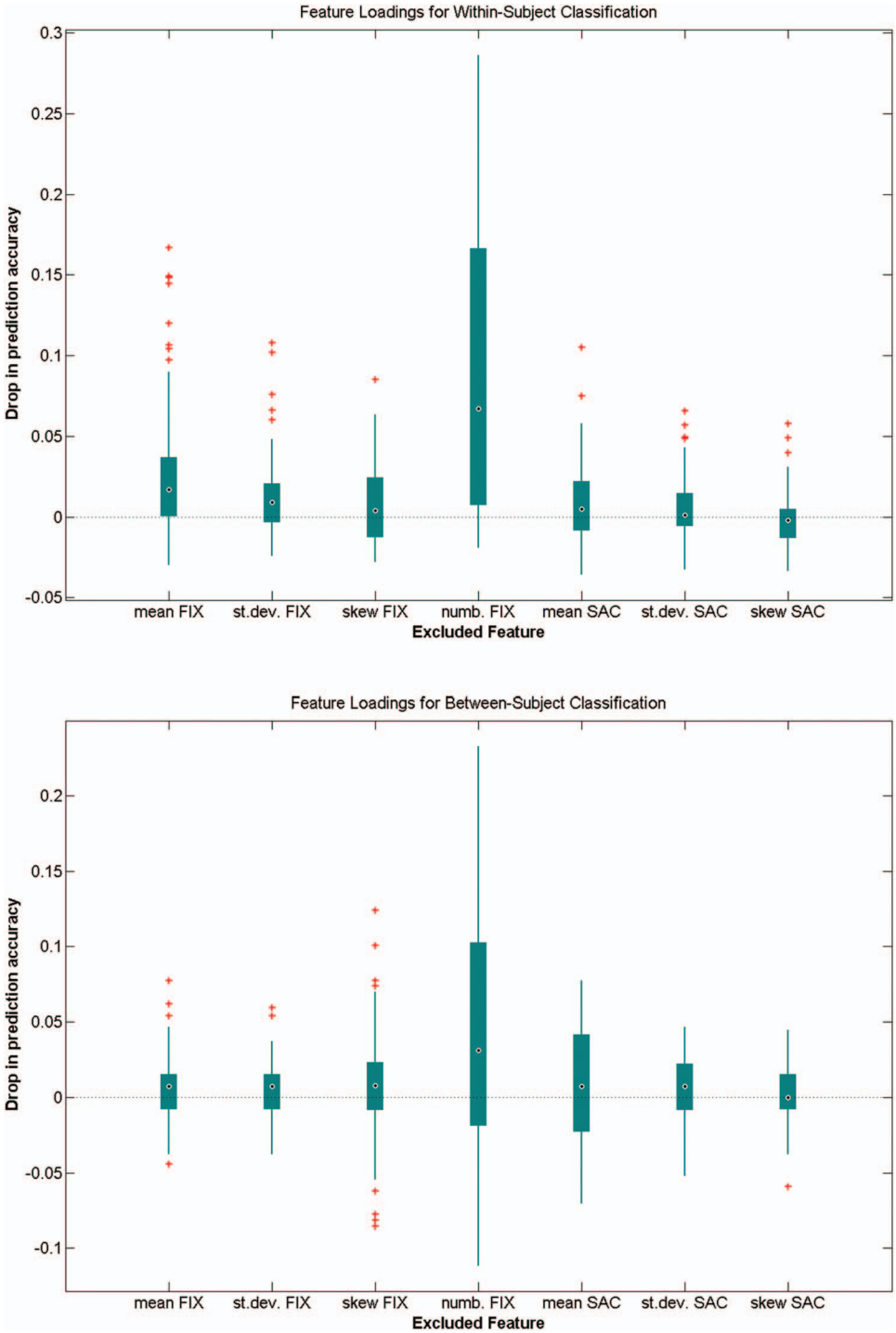

Figure 5. Relative influence of each measured feature on task classification using the linear discriminant classifier for within-participant classification (top) and across-participant classification (bottom). The vertical axis shows the drop in classification accuracy when the corresponding feature was excluded from the classification relative to classification using all features. The dot for each feature indicates the median value, and the box indicates the middle two quartiles; the whiskers represent the distribution, with + indicating the outlying values. The dashed line indicates no change in classification accuracy. VS = visual search; SM = scene memorization; AP = aesthetic preference. See the online article for the color version of this figure.

other six features, taken in isolation, were less important for classification, with skewness of saccade amplitudes having the least contribution for both within- and across-participants classification.

This result again suggests that the covariance between features is particularly important for classification accuracy because dropping individual features does not generally produce appreciable drops in classification accuracy. This suggests that the highly correlated nature of these variables (see Figure 2) fails to produce a large classification cost if only one variable is removed. Perhaps, if highly correlated feature pairs were removed, rather than single features, classification accuracy may have been impaired to a greater degree. Because the fixation features seem to be uncorrelated with the saccade features (see Figure 2), better classification accuracy could be achieved by a hybrid LD/GNB-L model, in which the fixation features are assumed to be independent of the saccade features. Evaluating this hybrid classifier is an interesting direction for future research.

Visual Search Caveats

Although including search target objects in the scenes helps to ensure that the participants are kept motivated throughout the search task, one drawback is the possibility that some of the viewing time during the visual search task may not be related to searching when the object has been found before the search time ends. Search trials would therefore include eye movements that were not due to search. It could be that these eye movements changed once search was complete in a way that artificially aided the classifier to distinguish the visual search task from the other tasks. Figure 6 compares the means, standard deviations, and skewness of saccade amplitudes and fixation durations for the eye-movement patterns for all three tasks with complete 8-s trials (labeled as VS, SM, and AP) as well as parts of trials in search before (pre-VS) and after (post-VS) each participant successfully found the target (sometimes the participant failed to find the target in 8 s). To check whether the differences in features before and after finding the target in the search trials inflated the accuracy of classifiers to distinguish this task from the others, we performed another classification analysis using only the predetection fixations for visual search and truncating the aesthetic preference and scene memorization trials to match predetection visual search trials. This was done by randomly assigning trial-by-trial search completion times from the visual search task to the aesthetic preference and scene memorization tasks and omitting the fixations in each trial that followed the assigned/artificial task completion times. This simulation was done 100 times.

For within-participant classification, all of the classifiers performed better than chance, and the median classification accuracies across 100 simulations were 58.6% for LD, 63.0% for QD, 51.4% for GNB-L, and 60.0% for GNB-N (note: chance is 33.3%),

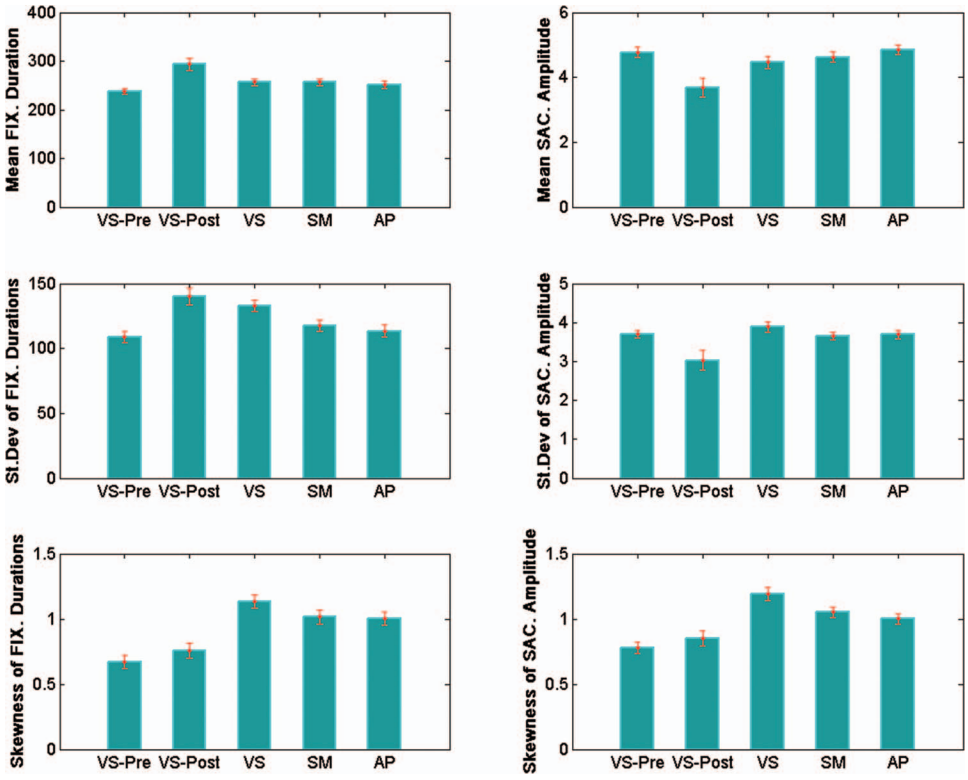

Figure 6. Eye-movement features for each visual task (VS = visual search, SM = scene memorization, AP = aesthetic preference) and for the visual search task separated by the eye-movement features for the intervals before the object was found (VS-Pre) and after the object was found (VS-Post). See the online article for the color version of this figure.

showing classification accuracy differences of −7.8% for LD, −2.5% for QD, −1.8% for GNB-L, and +7.7% for GNB-N compared with the original analysis. The mean accuracies across subjects for all classifiers were significantly higher than 0.33 ($df = 71$, $p < .001$, for all classifiers). For between-participants task classification, median classification accuracies across 100 simulations were again above chance for all classifiers (48.7% for LD, 54.0% for QD, 49.1% for GNB-L, 52.6% for GNB-N), showing classification accuracy differences of −6.9% for LD, +8.2% for QD, +4.2% for GNB-L, and +7.1% for GNB-N compared with the original analysis. Similar to the across-participant analysis on the original data, classification accuracy was below chance for a small subset of participants (four participants in LD and one participant for GNB-L), and yet the means of accuracies across subjects were significantly greater than 0.33 for all classifiers ($df = 71$, $p < .001$, for all classifiers).

Average classifier performance from this analysis thus showed similar levels of accuracy for both within- and between-participants classifications. However, there was a decrease in accuracy for the linear classifiers (specifically LD), which we believe was due to the amount of data that was lost when performing this truncation (on average, the last 51% of trials were removed in all tasks). Another explanation for this performance drop in LD could be that the number of fixations may have inflated accuracy for linear classifiers before equating task lengths (original analysis before truncating the trials). After truncating the trial lengths, however, the number of fixations may have become less informative for the classifiers because it was directly related to the total length of the tasks. To check this possibility, we calculated the accuracy drop in the LD classifier on the truncated trials after excluding the number of fixations. As before, the number of fixations had the highest isolated loading among features. The results showed an average accuracy drop of 3.26% in between-subjects classification and 7.67% in within-subject classifications after excluding number of fixations. These reductions in accuracy are similar to the results on the nontruncated data (4.42% between-subjects and 8.85% within-subjects). Therefore, the number of fixations' contribution to classification survives to a great degree even after equating task lengths.

Interestingly, we observed an improvement in classification accuracy in truncated trials versus the full data set for between-subjects task classification in the nonlinear classifiers. Improved performance of the nonlinear classifiers could be the result of removing some noise from the search task. The more limited data set may have made the covariance and variance structure of the visual search task more distinct from aesthetic preference and scene memorization, resulting in enhanced performance of QD and GNB-N. Performing classification analysis on only unsuccessful search trials was not possible given the lack of sufficient data for training (in only 15% of the trials the target had never been found). Therefore, we avoid drawing too many conclusions from the truncated data sets because of the large decrease in the amount of data, but these analyses demonstrate that our classification results were not artificially inflated based on the eye-movement patterns that occurred after targets were found in the visual search task.

Discussion

How eye movements reflect underlying cognitive processes during scene viewing has been a topic of considerable theoretical interest. One focus of research concerns the relationship between viewing task and eye-movement characteristics, with an emphasis on documenting differences in eye movements across viewing tasks (e.g., Castelhano & Henderson, 2008; Henderson, 2003; Henderson et al., 1999; Mills et al., 2011; Tatler, 2009; Yarbus, 1967). In the present study, we addressed this issue by asking the reverse question: Can viewing task be predicted from differences in eye-movement patterns? Specifically, we investigated whether it is possible to classify the viewing task from differences in the descriptive statistics of fixations and saccades that viewers make as they view pictures under different viewing conditions (Borji & Itti, 2014; Greene et al., 2012; Henderson et al., 2013). We hypothesized that if eye movements generally reflect the cognitive processes that are active during a given task, then it should be possible to classify the viewing task from participants' eye movements. In addition, we hypothesized that if the mapping of cognitive processes to eye-movement behavior generalizes across people, then it should be possible to classify the task that one viewer is engaged in from the eye-movement data of other viewers.

The first goal of this study was to determine whether eye-movement features could be used to predict visual task performance, even when the same stimuli are used for each visual task. To test this, we asked 72 participants to view pictures of natural scenes in three viewing task conditions: scene memorization, visual search, and aesthetic preference. The images were the same in the three conditions to ensure that any differences in viewing behavior were reflecting task differences rather than image content. We showed that it is indeed possible to successfully classify the viewing task using characteristics of eye-movement patterns, even when the stimuli are held constant. This success was due to differences in the distributional properties of fixation durations and saccade amplitudes. Therefore, for the visual tasks that we tested, unique eye-movement characteristics can be used to classify visual tasks even when the stimuli are held constant.

The second goal of this study was to examine whether the mapping of cognitive processes to eye-movement behavior generalizes across individuals. To investigate this question, we performed between-subjects classifications in which classifiers were trained on one subset of subjects and then tested on another. While within-participant classification produced more accurate results, between-participants classification accuracy was also significantly above chance, suggesting that there are lawful task-specific consistencies in eye-movement patterns that generalize across participants. This result demonstrates that there are aspects of eye-movement patterns that characterize how people in general perform different visual tasks. From the perspective of cognitive control of overt attention, the results are consistent with models in which cognitive processes common to each task across viewers influence the manner in which attention is allocated over a scene (e.g., Nuthmann, Smith, Engbert, & Henderson, 2010).

Given that eye-movement patterns distinguished how participants performed different visual tasks, we were then able to examine which features were most critical in determining the cognitive task that the participants were performing. The number of fixations was a highly discriminating feature for both within-

and between-participants classification (see Figure 5). Potentially, more interesting is that whereas mean fixation durations and saccade amplitudes were only marginally important in distinguishing the visual tasks performed, parameters that capture the higher order shapes of their distributions (second and third moments) were more useful in distinguishing tasks. Specifically, we showed that the second (standard deviation) and third moment (skew) of fixation durations and saccade amplitudes distinguish search (with the most variability and skewness) from memorization (with variability higher than preference) and preference (being the most homogenously performed task in terms of duration of fixations and size of saccades during viewing) in a reliable way. One way to interpret this is that the time pressure of meeting an end is greatest for search and is the least for aesthetic preference judgment. Thus, adaptive alterations in eye-movement behavior are more drastic in search, more modest in memorization (with a less immediate goal), and relatively small in preference judgments during scene viewing. This could also be more generally thought of as a control system with feedback. If the gain of the feedback component is comparable to the gain of the feedforward component of the system, more drastic alterations are likely to be produced by the system, leading to more adaptive and less homogenous output (where attention deployments are reflected by eye-movement behavior). If, however, the gain of the feedback component is small relative to the feedforward component, the output will be less varied and less skewed. More research would be necessary to validate this feedback-gain-tuning control system model of visual task performance, but we believe that this may be a useful model to help predict eye movements for different visual tasks.

This finding is also interesting in light of the high test–retest reliability of fixation duration and saccade amplitude distributions across tasks and participants (Henderson & Luke, 2014). These results suggest that the distributional properties of these eye-movement characteristics, which reflect processing time (fixation duration) and attentional breadth (saccade amplitude), vary across tasks in reliable ways. From an empirical standpoint, these results indicate that attempts to model and understand overt attention and eye-movement control in scene viewing must take into account the shapes of the distributions as well as the means of these features (Nuthmann et al., 2010; see also Henderson & Luke, 2014; Henderson, Choi, & Luke, 2014). In the future, it will be important to determine whether the inclusion of distributional properties of additional eye-movement features will provide further discriminating information and boost classifier performance.

The third goal of this study was to determine whether one classifier was superior to the rest, as this information could be used to constrain theories of how eye movements relate to different cognitive states. The LD classifier was the most successful at predicting the task both within and across participants. This method uses a probabilistic model that considers the covariance between features but ignores task-related differences in the covariance of the features; that is, some covariance between eye-movement features could change with task. Specifically, the covariance between the mean and standard deviation of fixation durations differed between the visual search task and the scene memorization task, $z(69) = 2.34$, $p = .019$, and also between the visual search task and aesthetic preference task, $z(69) = 3.89$, $p < .001$. Interestingly, accounting for these differences (as in the QD model) did not improve task classification within a participant, and actually impaired classification across participants. This finding suggests that accounting for the covariance between eye-movement features is beneficial for accurate task classification, but task-specific modulations of these correlations do not generalize across participants. However, because the classification results were so similar between LD and QD in the full data set, and QD outperformed LD in the truncated data, it may be premature at this point to speculate about differences between them. What we can confidently conclude is that the covariance between eye-movement features is important in determining the task that individuals are performing. Treating each eye-movement feature as independent, as in the univariate classifiers, can achieve above-chance classification, but is not the optimal model in determining participants' cognitive states (i.e., the task they are performing).

In summary, if we want to determine which cognitive state a person is in from her eye-movement patterns, we need to model the covariance between eye-movement features. This finding suggests that there are relationships among the distributional features of fixation durations and saccade amplitudes across tasks, a finding that is not explicitly included in most current models of eye-movement control and should be accounted for in future models. An important caveat on this conclusion, however, is that our results also demonstrated a lack of correlation between fixation and saccade features, supporting the suggestion that two relatively independent mechanisms are involved in controlling saccades and fixations (Castelhano & Henderson, 2008; Henderson & Luke, 2014; Rayner, 2009; but see also Unema, Pannasch, Joos, & Velichkovsky, 2005, for an alternative view). Stronger correlations were observed within features (mean, standard deviation, and skewness) of fixation durations, and within the same features of saccade amplitudes. However, correlations between the mean and the standard deviation are expected in a population with non-zero skewness (Shanmugam, 2008). The distribution of fixation durations, and of saccade amplitudes has been shown to be highly skewed in previous work (Castelhano et al., 2009; Luke et al., 2013; Tatler et al., 2006).

In conclusion, our study supports three main conclusions. First, we demonstrated that it is possible to successfully classify the viewing task from eye movements. Second, we found that four common classifiers with different underlying assumptions were all successful to varying degrees, with classifiers that accounted for the covariance between features generally producing better performance. Third, we demonstrated that the relationships between the viewing task and the eye-movement patterns generalized across participants, such that we could classify a given participant's task from the similarity of her eye movements to those of other participants. This result is theoretically important because it suggests that task differences in eye movements are not simply due to idiosyncratic differences, but instead reflect common underlying cognitive mechanisms that are consistent across participants and reveal changes in the deployment of overt attention. As discussed by Henderson et al. (2013), this result also has practical implications for using eye-movement classification technologies to classify cognitive states in human–computer interactions. Specifically, the results suggest that changes in eye movements across tasks are lawful enough that they may be used to infer the task engaged. In the future, analysis of stimulus-driven visual features (such as saliency, entropy, etc.) and their distributions, as well as time-series analysis throughout visual tasks may provide additional

sensitive information, which is the next step for future work in our laboratories.

References

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision, 14,* 29. http://dx.doi.org/10.1167/14.3.29

Buswell, G. (1935). *How people look at pictures*. Oxford, UK: University of Chicago Press.

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

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9,* 6–15. http://dx.doi.org/10.1167/9.3.6

Cohen, J. (1994). The earth is round ($p <$ .05). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7,* 1–30.

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. http://dx.doi.org/10.1016/j.visres.2012.03.019

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7,* 498–504. http://dx.doi.org/10.1016/j.tics.2003.09.006

Henderson, J. M. (2011). Eye movements and scene perception. In S. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *Oxford handbook of eye movements* (pp. 593–606). New York, NY: Oxford University Press.

Henderson, J. M., Choi, W., & Luke, S. G. (2014). Morphology of primary visual cortex predicts individual differences in fixation duration during text reading. *Journal of Cognitive Neuroscience, 26,* 2880–2888. http://dx.doi.org/10.1162/jocn_a_00668

Henderson, J. M., & Luke, S. G. (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1390–1400. http://dx.doi.org/10.1037/a0036330

Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PLoS One, 8,* e64937. http://dx.doi.org/10.1371/journal.pone.0064937

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210–228. http://dx.doi.org/10.1037/0096-1523.25.1.210

Luke, S. G., Nuthmann, A., & Henderson, J. M. (2013). Eye movement control in scene viewing and reading: Evidence from the stimulus onset delay paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 10–15. http://dx.doi.org/10.1037/a0030392

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,* 17. http://dx.doi.org/10.1167/11.8.17

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review, 117,* 382–405. http://dx.doi.org/10.1037/a0018924

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422. http://dx.doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62,* 1457–1506. http://dx.doi.org/10.1080/17470210902816461

Rosa, G. J. M. (2010). The elements of statistical learning: Data mining, inference, and prediction by Hastie, T., Tibshirani, R., and Friedman, J. *Biometrics, 66,* 1315. http://dx.doi.org/10.1111/j.1541-0420.2010.01516.x

Seber, G. A. F. (2004). *Multivariate observations*. New York, NY: Wiley.

Shanmugam, R. (2008). Correlation between the sample mean and sample variance. *Journal of Modern Applied Statistical Methods, 7,* 6.

Tatler, B. W. (2009). Current understanding of eye guidance. *Visual Cognition, 17,* 777–789. http://dx.doi.org/10.1080/13506280902869213

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,* 1857–1862. http://dx.doi.org/10.1016/j.visres.2005.12.005

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition, 12,* 473–494. http://dx.doi.org/10.1080/13506280444000409

Yarbus, A. L. (1967). *Eye movements and vision*. New York, NY: Springer Science + Business Media.

Received May 30, 2014
Revision received June 23, 2015
Accepted July 14, 2015 ■