Classification of visual and linguistic tasks using eye-movement features

Moreno I. Coco

Faculdade de Psicologia, Universidade de Lisboa, Lisboa, Portugal

Frank Keller

School of Informatics, University of Edinburgh, Edinburgh, UK

The role of the task has received special attention in visual-cognition research because it can provide causal explanations of goal-directed eye-movement responses. The dependency between visual attention and task suggests that eye movements can be used to classify the task being performed. A recent study by Greene, Liu, and Wolfe (2012), however, fails to achieve accurate classification of visual tasks based on eye-movement features. In the present study, we hypothesize that tasks can be successfully classified when they differ with respect to the involvement of other cognitive domains, such as language processing. We extract the eye-movement features used by Greene et al. as well as additional features from the data of three different tasks: visual search, object naming, and scene description. First, we demonstrated that eye-movement responses make it possible to characterize the goals of these tasks. Then, we trained three different types of classifiers and predicted the task participants performed with an accuracy well above chance (a maximum of 88% for visual search). An analysis of the relative importance of features for classification accuracy reveals that just one feature, i.e., initiation time, is sufficient for above-chance performance (a maximum of 79% accuracy in object naming). Crucially, this feature is independent of task duration, which differs systematically across the three tasks we investigated. Overall, the best task classification performance was obtained with a set of seven features that included both spatial information (e.g., entropy of attention allocation) and temporal components (e.g., total fixation on objects) of the eye-movement record. This result confirms the task-dependent allocation of visual attention and extends previous work by showing that task classification is possible when tasks differ in the cognitive processes involved (purely visual tasks such as search vs. communicative tasks such as scene description).

Introduction

Visual attention actively serves the cognitive system in a wide range of different tasks and everyday activities. Each task entails a well-defined sequence of steps to be accomplished. In order to inform this process, specific task-related information has to be extracted by the visual system from the visual percept.

The role of the task in visual attention has attracted the interest of vision researchers from very early on. Buswell (1935) was the first one to investigate eye movements with complex scenes and to show that expertise in a certain task (being an artist or not) influences the associated eye-movement responses. A few decades later, Yarbus (1967) confirmed that indeed task plays a key role in the eye-movement patterns observed. Different task instructions, such as "estimate the material circumstances of the family shown in the picture" versus "give the ages of the people shown in the picture," resulted in qualitatively different eye-movement trajectories, often referred to as *scan paths* (Noton & Stark, 1971) or scan patterns (Henderson, 2003).

The key message of this seminal work was that eye-movement patterns provide evidence about a possible causal model of the task being performed. Thus, it should be possible to infer the underlying attentional mechanisms by comparing eye-movement patterns across tasks.

More recent work in visual cognition is motivated by the aim of understanding the visual system in ecologically valid real-world tasks, such as making tea or washing hands; sport activities, such as playing table tennis or driving; as well as in computer-simulated, virtual scenarios (Ballard & Hayhoe, 2009; Ballard, Hayhoe, & Pelz, 1995; Hagemann, Schorer, Cañal-Bruland, Lotz, & Strauss, 2010; Land & Furneaux, 1997; Land & Hayhoe, 2001; Land & McLeod, 2000;

Citation: Coco, M. I., & Keller, F. (2014). Classification of visual and linguistic tasks using eye-movement features. *Journal of Vision*, *14*(3):11, 1–18, http://www.journalofvision.org/content/14/3/11, doi:10.1167/14.3.11.

doi: 10.1167/14.3.11 Received August 2, 2013; published March 7, 2014 ISSN 1534-7362

Land, Mennie, & Rusted, 1999; Pelz & Canosa, 2001; Rothkopf, Ballard, & Hayhoe, 2007). This research has demonstrated that eye-movement responses are launched preferentially to task-relevant objects during precisely time-locked stages of the task, e.g., looking at the spout of a kettle when pouring during a tea-making task (Land et al., 1999). The memorability of attended information has also been shown to depend on its task-relevance. Triesch, Ballard, Hayhoe, and Sullivan (2003), for example, showed that participants became aware of changes occurring to an attended object only when such object was relevant, at that particular moment, for the task.

Effects of task have also been observed in other visual activities, such as search or memorization (Castelhano, Mack, & Henderson, 2009; Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011; Tatler, Baddeley, & Vincent, 2006), in which photo-realistic 2-D scenes were mainly used as contexts. Castelhano et al. (2009), for example, compared several measures, such as the area of the scene inspected, for eye-movement data collected during a visual search task (find a MUG in a kitchen scene) and during a memorization task (memorize the scene in preparation for a later recall). They found significant differences between the two tasks, e.g., more regions of the scene were inspected during memorization than during search. In a memorization task, participants attempt to inspect as many objects as possible within the preview time with the aim of maximizing the number of items that could be recalled whereas, in a search task, participants focus on contextually relevant regions of the scene to maximize the likelihood of finding the target object (see also Castelhano & Heaven, 2010 and Malcolm & Henderson, 2010 for evidence of top-down contextual-target guidance in visual search). Moreover, also in purely visual tasks, the task-relevance of scene information exerts direct control on eye-movement responses, for example, on the duration of the first fixation (Glaholt & Reingold, 2012; see, however, Salverda & Altmann, 2011, in which task-irrelevant information is also shown to impact fixation duration). The influence of direct cognitive control on visual attention is a strong indicator that task differences should be observed in the associated eye-movement pattern.

The causal dependence of task and eye-movement responses extends also to other cognitive activities, such as reading. Research in this area has clearly shown that eye-movement responses are strongly modulated by linguistic properties of the text, such as word frequency (Inhoff & Rayner, 1986), and more general task demands, such as silent reading versus reading aloud. Moreover, eye-movement patterns in reading significantly differ from those observed during scene perception. The average length of a saccade is, for example, longer in scene perception than in reading (see Rayner, 2009, for a review on the topic).

Task is therefore a major factor that needs to be considered when interpreting eye-movement responses. In fact, by understanding the properties of a task and its underlying goals, we should be able to accurately estimate which objects (or words) are attended and when this should happen during the task (Ballard & Hayhoe, 2009). If this is correct, then the inverse inference should also be possible: Eye-movement responses should be informative of the task being performed. In particular, the statistics of eye-movement responses should be predictive of the task.

A recent study by Greene et al. (2012) addressed this question by explicitly testing whether the task performed by the participants could be accurately determined from the associated eye-movement information (see also Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013 for another example of eye movement–based task classification). This study followed up on work by DeAngelus and Pelz (2009), in which Yarbus's (1967) qualitative findings were successfully replicated and more stringently quantified, using modern eye-tracking technology, a larger set of participants (25 instead of just one) by comparing task differences with and without self-termination (3 min of fixed viewing per scene).

Participants in Greene et al.'s (2012) study were asked to perform four types of visual tasks (memorize the picture, determine the decade the picture was taken in, determine people's wealth and social context); the study used black-and-white historical scenes selected from the Time Life archive on Google (http://images.google.com/hosted/life). From the eye-movement data collected during the different tasks for each individual trial, Greene et al. extracted seven distinct features (e.g., area of the scene fixated), reported also in previous studies (Castelhano et al., 2009; Einhäuser, Spain, & Perona, 2008; Mills et al., 2011). This set of features was used to train different regression-based classifiers (e.g., support vector machines) in order to automatically determine which of the four tasks was performed in a given trial.

Surprisingly, Greene et al.'s (2012) results show that none of the classifiers they utilized was able to detect the task performed using eye-movement features with an accuracy above chance. This result seems at odds with DeAngelus and Pelz (2009), in which task differences were clearly observed (e.g., the distance of scan-path trajectories between tasks was significantly larger than between observers). More generally, the result of Greene et al. undermines previous claims about task-dependent allocation of visual attention and challenges the widespread assumption that each task gives rise to a distinct pattern of eye-movement responses.

However, the fact that Greene et al. (2012) did not observe task-specific patterns of eye movement might not mean that such task differences do not exist but rather that the tasks performed by Greene et al.'s participants were not distinct enough to produce separable patterns of visual attention. One, mostly technical, explanation for the null result of Greene et al.'s study might be that all tasks had a fixed termination time (10 s per trial). The fixed viewing time might have flattened any implicit temporal variability between tasks. This hypothesis can be deduced from the study of DeAngelus and Pelz (2009), in which it is clear that different tasks trigger different self-termination times. However, if this is the only explanation of the null result, then it should be impossible to accurately classify tasks using eye-movement features that are insensitive to self-termination, such as initiation time (i.e., the time it takes to launch the first eye movement) or the average length of a saccade.

Another, perhaps stronger, alternative hypothesis for the null result is that all tasks performed in Greene et al. (2012) demanded only visual processing and did not require the involvement of other cognitive modalities (e.g., language processing). Thus, it seems likely that similar strategies of attention allocation are adopted when participants perform similar visual tasks. Greene et al.'s results therefore leave open the possibility that eye-movement patterns become distinctive when tasks differ substantially with respect to the cognitive processes they involve. This hypothesis would also better align with the literature on natural gaze control during real-world tasks, during which visual attention always occurs jointly with motor actions and attention allocation strongly depends on the task goals (e.g., Pelz & Canosa, 2001). Visual attention does not only co-occur with motor actions, but it also often co-occurs with language in tasks such as describing the function of a device or giving directions on a map. Thus, we can hypothesize that reliable task differences can be observed between eye-movement patterns for purely visual tasks (e.g., visual search) and communicative tasks (scene description), which require the concurrent processing of visual and linguistic information.

Psycholinguistic research on linguistic tasks situated in visual scenes has, in fact, convincingly demonstrated that the allocation of visual attention and the processing of linguistic information are mutually interdependent: The mention of a visual referent occurs time-locked with fixations on the relevant object in a scene in both language comprehension and language production (e.g., Gleitman, January, Nappa, & Trueswell, 2007; Griffin & Bock, 2000; Tanenhaus, Eberhard, & Sedivy, 1995). In scene-description tasks, for instance, the time-locking of speech and eye movements means that sentence similarity correlates with the similarity of the associated scan patterns (Coco & Keller, 2012). This suggests that eye-movement responses carry detailed information about the task that is performed; in fact, Coco and Keller (2012) show that it is possible, with accuracy above chance, to determine which sentence was spoken based on the scan pattern that was followed when speaking it.

The aim of the present study is to demonstrate that different tasks are characterized by distinct patterns of eye-movement responses and that the accurate classification of tasks is possible provided that the tasks differ substantially with respect to the cognitive processes involved in accomplishing them. In particular, we compare a visual search task with object-naming and scene-description tasks, both of which require the concurrent processing of visual and linguistic information. The three tasks considered vary by the amount of cross-modal processing involved: (a) Search is expected to mostly require visual processing, (b) naming also demands the activation of linguistic information (the names of objects need to be retrieved and uttered), and (c) description requires visual and linguistic processing to be integrated fully as sentences and scan patterns correlate closely (Coco & Keller, 2012). This key difference is predicted to lead to distinct eye-movement patterns, which would make it possible to classify all three tasks with an accuracy greater than chance. However, we also predict classification accuracy to degrade when more cross-modal processing is required. In that case, eye-movement features need to be used in conjunction with linguistic features to achieve good task classification performance.

The present study

We conducted three eye-tracking experiments involving visual search, object naming, and scene description; each task was performed by a different group of participants. From the eye-movement data of each trial, we extracted the seven features used by Greene et al. (2012) (we will refer to these features as *GF*) as well as other eye-movement features (referred to as *OF*); these will be explained in more detail in the Features section below.

We compare how visual search, object naming, and scene description affect eye movements using linear–mixed effect modeling. This analysis allows us to infer how the goals of a given task determine which objects need to be attended to perform the task, thus explaining why different eye-movement patterns emerge across tasks.

We then train three different types of regression models (multinomial regression, least-square angle regression, and support vector machines) on the eye-

movement features in order to automatically classify the task used. When training the classifiers, we use either only *GF*, only *OF*, or all features. We test the accuracy of the classification models using a tenfold cross-validation procedure. The results show that all models, using all three feature sets, are able to predict the task with an accuracy well above chance.

Moreover, in order to test whether certain features are more predictive than others and to investigate how many features are needed to obtain a classification accuracy above chance, we run a stepwise forward model-building procedure in which features are ordered by their classification performance (best first) over the 10 cross-validation folds. We track how classification accuracy changes as a function of the features used and find that already a single feature is enough to distinguish among tasks above chance.

We conclude the study by demonstrating that eye-movement features insensitive to self-termination, such as initiation time and saccade length, are sufficient to classify the tasks well above chance. This additional analysis rules out the possibility that fixed termination is what caused the null result observed by the Greene et al. (2012) study, hence suggesting that tasks can be accurately classified as long as they substantially differ in the types of cognitive processes needed to perform them.

Tasks

Visual search

Participants were asked to count how many instances (between one and three) of a target object are present in the scene.[1] The target object was cued prior to scene onset for 750 ms, using a word naming the object displayed in the center of the screen. The target object was either animate or inanimate (each half of the time). Participants could freely inspect the scene without any time constraints and then self-terminate the trial by pressing a response button to indicate the number of targets seen. Once every four trials, a comprehension question about the number of target objects present in the scene was asked. This data set was published by Dziemianko, Keller, and Coco (2009).

Scene description

The same scenes and the same target objects were used as in the visual search task, but now participants were asked to generate a spoken description of the target object. The target object was cued in the same way (a word naming the object was displayed for 750 ms prior to scene onset). Again, participants could freely inspect the scene without any time constraints; once they had finished speaking, they pressed a response button to trigger the next trial. This data set was published by Coco and Keller (2012).

Object naming

The same scenes as in the previous two tasks were used, but now no cue was presented to participants. Instead, participants were asked to name at least five objects in the scene by speaking words describing them. Participants had 1500 ms preview of the scene after which a beep was played to prompt them to start naming. The scene was visually available during the whole trial, i.e., for preview and naming. Again, they were under no time pressure and had to press a response button to proceed to the next trial. This data set was collected as filler trials in Coco, Malcolm, and Keller's (2013) study.

Materials

We created 24 photo-realistic scenes drawn from six different indoor scenarios (e.g., bathroom, bedroom), four scenes per scenario, using Photoshop (see Figure 1 for an example scene[2]). Scene clutter was manipulated: There were two versions of each scene with either low clutter or high clutter as estimated using the Feature Congestion measure of Rosenholtz, Li, and Nakano (2007). Note that the clutter manipulation was part of the original studies that generated our data sets (Coco & Keller, 2012; Dziemianko et al., 2009); it is not of relevance for the present study. However, in order to make sure that classification performances were consistent in both versions of the scene, we trained and tested the classifiers on scenes split by clutter (high and low) and showed that the classification results were consistent between the two sets.

The cued object used in the visual search and scene-description task was either animate (GIRL in this example) or inanimate (TEDDY in this example). The cue was always referentially ambiguous with respect to the scene: In this case, two GIRLS and two TEDDIES are depicted. (Again, this feature was of interest in the studies that originated the data but will be ignored in the following.) A Latin square design was used to make sure that each scene was only seen in one of the four conditions (cue either animate or inanimate, clutter either low or high) by each participant.

We use LabelMe (Russell, Torralba, Murphy, & Freeman, 2008) to fully annotate each scene with the objects of which it is made. Objects at the border of the scene were annotated using background generic labels, such as wall or floor.[3] Low-clutter scenes had a mean density of 3.10 ± 0.22 and mean number of objects 27.42 ± 9.93 whereas, in high-clutter scenes, the mean

Figure 1. An example of a scene used in the different tasks: low-clutter condition on the left, high-clutter condition on the right. The cued target objects were GIRL and TEDDY. The face is blanked out to protect the identity of the photographed character. The image is an original created with PhotoshopC2 using components that are in the public domain sources (e.g., Flickr).

density is 3.90 ± 0.24, and the number of objects is 28.65 ± 11.30. We mapped x-y fixation coordinates onto the corresponding objects. However, as objects can be nested, e.g., the TEDDY polygon is embedded into the GIRL polygon, we use the size of the object in pixels squared to assign the fixation to the smallest object, i.e., TEDDY in this working example. This makes sure that features at fixation are not redundantly computed over nested objects.

Participants

Seventy-four (25 each for search and description, 24 for naming) native speakers of English, all students of the University of Edinburgh, gave informed consent to take part to the experiments and were each paid five pounds.

Apparatus and procedure

An EyeLink II head-mounted eye-tracker was used to monitor participants' eye movements with a sampling rate of 500 Hz. Images were presented on a 21-in. multiscan monitor at a resolution of 1024 × 768 pixels. Participants sat between 60 and 70 cm from the computer screen, which subtended approximately 20° of visual angle. A nine-point randomized calibration was performed at the beginning of the experiment and repeated halfway through the experimental session. Drift correction was performed at the beginning of each trial. The tasks were explained to participants using written instructions; the experiment took about 30 min to complete.

Features

The full data set contains a total of 1,756 unique trials, which are divided across the three tasks as follows: search (580), description (600), naming (576). Approximately 3% (20 trials) of the visual search data was lost due to machine error.

From the eye-movement data of each trial, we extract the seven features used by Greene et al. (2012): (a) number of fixations, (b) mean fixation duration, (c) mean saccade amplitude, and (d) percent of image covered by fixations assuming a 1° circle around the fixation position (Castelhano et al., 2009; Einhäuser et al., 2008; Mills et al., 2011). As two fixations can fall in close proximity to each other, the areas of the two circles may overlap. In this case, we subtract the area of the intersection. Following Greene et al. (2012), we also calculated the proportion of dwell time on (e) faces, (f) bodies, and (g) objects, i.e., any other region annotated in the scene that was not a human.

In addition to the Greene et al. (2012) feature set (GF), we extracted another set of 15 features (OF): (a) latency of first fixation; (b) first fixation duration; (c) mean fixation duration; (d) total gaze duration on faces, bodies, and objects (four features for three different regions, i.e., 12 features in total); (e) the initiation time, which is the time spent after scene onset before the first saccade is launched (a measure used by Malcolm & Henderson, 2009); (f) mean saliency at the

fixation location; and (g) the entropy of the attentional landscape.

For the mean saliency measure, we computed a saliency map of each scene using the model developed by Torralba, Oliva, Castelhano, and Henderson (2006)[4] and mapped each x-y fixation position onto the saliency value at that location. We then took the mean saliency value of the fixations in the trial. To calculate the entropy measure, we first computed attentional landscapes by fitting 2-D Gaussians on the x-y coordinates of each fixation with the height of the Gaussian weighted by fixation duration and a radius of 1° of visual angle (roughly 27 pixels) to approximate the size of the fovea (e.g., Henderson, 2003; Pomplun, Ritter, & Velichkvosky, 1996). The entropy of the map was then calculated as $\sum_{x,y} p(L_{x,y}) \log_2 p(L_{x,y})$, where $p(L_{x,y})$ is the normalized fixation probability at point (x, y) in the landscape L. Conceptually, entropy measures the spread of information and the "uncertainty" we have about it. Thus, the higher the entropy, the more spread out fixations in the scene are, i.e., the more distinct locations have been attended.

Methods for analysis and classification

First, we investigated how tasks differed in their associated eye-movement features and derived a causal task-based interpretation for it. In particular, we examined all GF features, and in the OF feature set, we focused on initiation time, saliency, and entropy, which are not region-specific. We built a linear–mixed effects model for each feature as a function of *Task* (Search, Naming, or Description) using the R package lme4 (Bates, Maechler, & Bolker, 2011). As *Task* is a three-level, categorical variable, we needed to create a contrast coding and chose one of the factors as a reference level. This factor could then be used to compare with the other two levels. We chose Naming as the reference level as it is the simplest linguistic task and contrast it with Search and Description. In building our models, we followed Barr, Levy, Scheepers, and Tily (2013) and chose a maximal-random structure, in which each random variable of the design (e.g., Participants) is introduced as intercept and as slope on the predictors of interest (e.g., Search vs. Naming). The random variables of our design are Participants (74) and Scenes (48 as we have 24 scenes in two conditions of visual clutter). We report tables with the coefficients of the predictors, their standard error; significance is provided by computing p values using the function pvals.fnc from the languageR package (Baayen, Davidson, & Bates, 2008).

Then, we used all sets of features described above to train three different types of classifiers, all implemented in R. We trained a multinomial log-linear neural networks model (MM; multinom function in R's nnet package, Venables & Ripley, 2002); a generalized linear model with penalized maximum likelihood, in which the regularization path is computed for the least angle (LASSO; glmnet in R's glmnet package, Friedman, Hastie, & Tibshirani, 2010); and a support vector machine (SVM; ksvm in R's kernlab package, Chang & Lin, 2011). Note that we used the default setting for all three different models (higher classification accuracy could presumably be achieved by parameters tuning).

The classifiers were trained and tested using tenfold cross-validation, in which the model is trained on 90% of the data and then tested in the remaining 10%; this process is repeated 10 times so that each fold functions as test data exactly once. This is a way to avoid overfitting the training data while still making maximal use of a small data set. We trained the classifiers on the data divided by clutter condition (low, high). This makes sure that we treated the two sets of related images independently.

We measured the accuracy of the classifiers using *F*-score, which is the geometric mean of precision and recall and is defined as $F = 2 \cdot (P \cdot R)/(P + R)$. Precision ($P$) is the number of correctly classified instances over the total number of instances labeled as belonging to the class, defined as $tp/(tp + fp)$. Here, tp is the number of true positives (i.e., instances of the class that were correctly identified as such), and fp is the number of false positives (i.e., instances that were labeled as members of the class even though they belong to a different class). Recall (R) is the number of correctly classified instances over the total number of instances in that class, defined as $tp/(tp + fn)$, where fn is the number of false negatives, i.e., the number of instances that were labeled as nonmembers of the class even though they belong to the class. It is important to consider both precision and recall as a high precision can be achieved simply by underpredicting the class and being right most of the time when a prediction is made. However, in this case, recall will be low. The inverse is true for a classifier that overpredicts a class; in this case, recall will be high, but precision will be low.

Note also that precision and recall (and thus *F*-score) are relative to a given class; as we have three classes, we report a separate *F*-score value for each of them. Furthermore, we report results using three different sets of features: GF alone, OF alone, and all features. We ran this classification on all the data and made sure that data belonging to a scene in a certain clutter condition (high and low) was used either for training or testing and used a t test to determine whether a larger feature set led to a significant improvement in *F*-score.

For comparability with Greene et al.'s (2012) study, we also calculated the accuracy of the classifiers when trained to predict which image was viewed (here, we

collapsed the clutter condition and considered only 24 scenes). Again following Greene et al., we also trained the classifiers to predict which participant generated the data (74 participants).

Moreover, in order to provide a more detailed analysis of the impact of the features on classification performance, we used the classifier that gave us the best performance (SVM) and tested (a) how many features were needed to achieve a classification performance above chance, and (b) which features were important for discriminating between tasks. For the first analysis, we used a stepwise forward model-building procedure, and at each step we added the feature that maximized *F*-score classification performance and tracked changes in *F*-score as more features were included. We repeated this procedure over the 10 folds and then plotted the mean *F*-score obtained when each new feature was added. In the second analysis, we performed the same stepwise model-building procedure but evaluated whether the model with the added feature is significantly better than the one without it. If there was no significant improvement on the *F*-score, we retained the model without that feature. Again, we repeated this procedure over the 10 folds and retained the feature set of the final model obtained for each fold. We report the frequency of observing a certain feature in the final feature set over the 10 folds over 10 iterations (100 final feature sets in total; the folds are randomly regenerated at every iteration to make sure that the data is homogeneously sampled). This measure gives us a rough estimate of how important a feature is for discriminating among tasks.

We conclude the Results section by looking at whether features that are either independent of self-termination, such as initiation time, or purely spatial, such as mean saccade amplitude, are sufficient to obtain accurate classification performance. This test is there to make sure that it is not only the temporal component of the eye-movement data that enables us to accurately classify tasks. The temporal features (e.g., the number of fixations) are affected by self-determination, i.e., by whether the participant is able to terminate the task, or whether the task is of fixed length. (Recall that all our tasks used self-termination.)

Results and discussion

We start by examining the patterns observed for the different eye-movement features in each task. Then, we move on to the results obtained using these features when performing task classification.

Features mediating task variability

In Figure 2, we plot mean and standard error for GF, the set of features proposed by Greene et al. (2012), for our three different tasks.

Figure 2 plots the mean values for the seven features computed by Greene et al. (2012) for our three tasks. Table 1 gives the estimates of coefficients obtained by the linear–mixed effects model. The results show that the three tasks produce distinct patterns for each feature. In particular, we find that the number of fixations is significantly higher for object naming than for description and visual search (lowest). In a naming task, many objects are evaluated as potential naming targets whereas, in search, only objects that are contextually relevant to the search will be inspected. The description task is situated somewhere in between these two extremes: Only objects that will be mentioned or that are contextually related to them are fixated. Importantly, these strategies are also reflected in the mean saccade amplitude: Saccades during object naming are short as objects that are in close proximity are evaluated compared to saccades in visual search, in which large sections of the scene are covered. Interestingly, we find that saccades during scene description are shorter than during naming. Descriptive sentences often have a simple subject-verb-object structure, e.g., *the girl is hugging a teddy*. Visually, this implies fixating the agent, determining the action performed, and then fixating the object of the action. This visual information tends to be spatially proximal.

We also find that mean gaze duration is longer in naming and description compared to search. A communicative task requires the processing of both linguistic and visual information, so gazes are longer when the information from both modalities needs to be evaluated for each object. Furthermore, we also observe a significant difference in the mean gaze duration between naming and description: Fixations are longer in naming than in description. This is presumably due to the fact that a naming task demands a more focused retrieval of lexical material associated with the visual object to be mentioned than in a description task in which dependencies among other objects (semantic relationships, syntactic correspondences) need to be established.

When looking at the dwell time in the different regions, we find that the description task is the one in which most attention is allocated to humans (i.e., faces and especially bodies). As mentioned above, a description often entails that an action is verbalized; hence attention is allocated to the animate agent and to the action he or she is performing. This result corroborates the pattern we observed in mean saccade amplitude. In the search task, however, the recognition of an animate agent involves viewing his or her face,

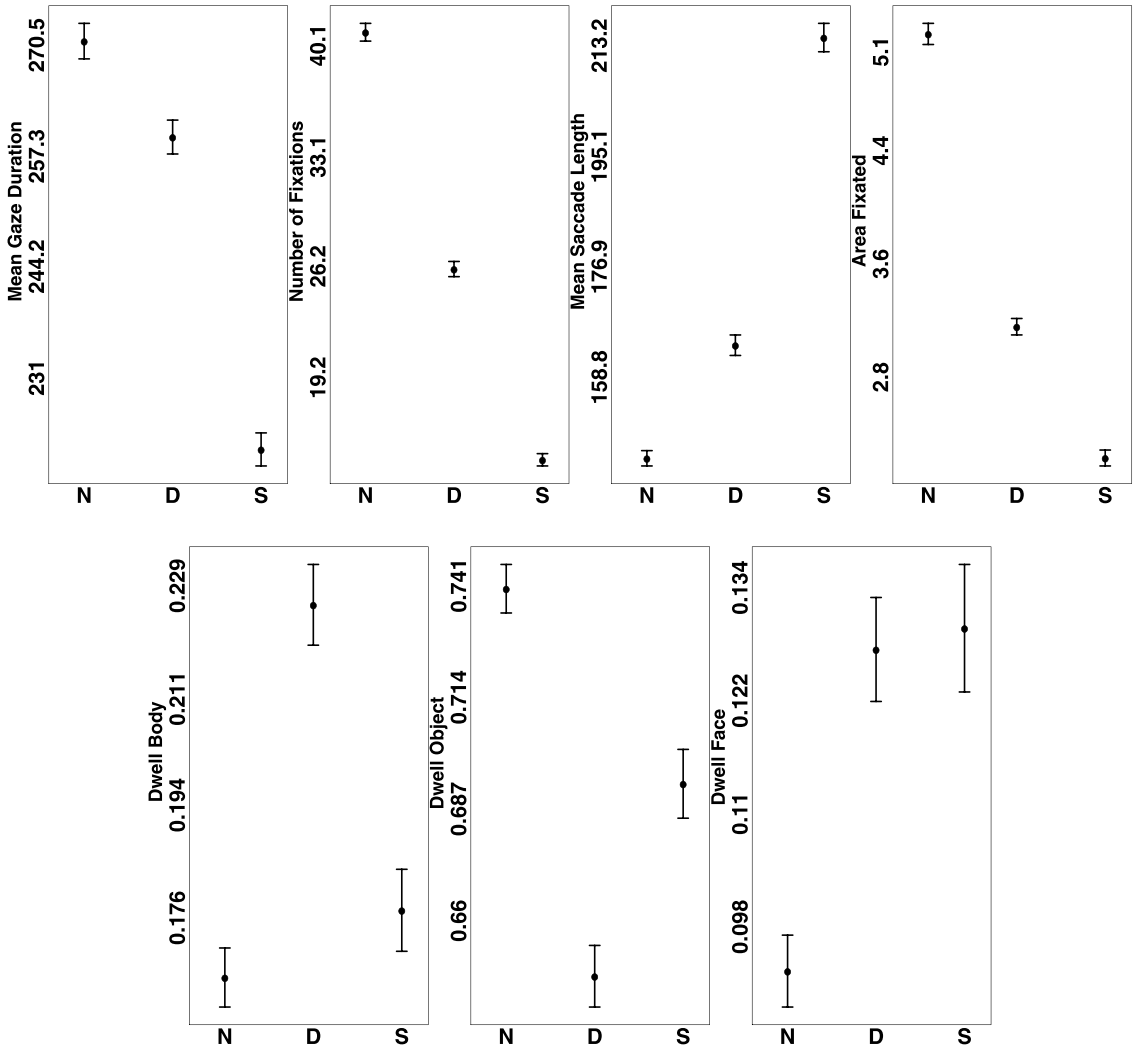

Figure 2. Mean values for the features proposed by Greene et al. (2012). Each feature is plotted in a separate panel for the three different tasks: Naming (N), Description (D), and Search (S). The error bars indicate the standard error. The unit for all eye-movement measures is the millisecond or proportions. The exceptions are mean saccade amplitude, which is in pixels, and area fixated (in percentage).

	Intercept			Description versus naming			Search versus naming		
Features	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*
Number of fixations	26.85	0.99	0.0001	−1.89	2.90	0.1	−26	2.38	0.0001
Area fixated	3.52	0.11	0.0001	−0.75	0.30	0.0001	−2.62	0.28	0.0001
Mean saccade amplitude	175.04	2.58	0.0001	−21.52	6.64	0.0001	80.12	7.23	0.0001
Mean gaze duration	250.65	4.06	0.0001	−17.51	10.9	0.003	−57.81	10.65	0.0001
Dwell body	0.19	0.004	0.0001	0.07	0.01	0.0001	−0.02	0.01	0.02
Dwell face	0.11	0.003	0.0001	0.02	0.01	0.04	0.02	0.01	0.01
Dwell object	0.69	0.005	0.0001	−0.09	0.01	0.0001	0	0.0	0.9
Initiation time	312.01	8.05	0.0001	97.39	21.33	0.0001	63.79	21.51	0.0001
Saliency	247.41	0.75	0.0001	5.70	2.16	0.005	1.18	2.15	0.5
Entropy	11.47	0.02	0.0001	−0.19	0.06	0.0001	−0.38	0.06	0.0001

Table 1. Coefficients of linear–mixed effects models with maximal random structure (intercept and slopes on Participants and Scenes). *Notes*: Each feature is modeled as a function of Task, which is contrast coded with Naming as a reference level for Description and Search.

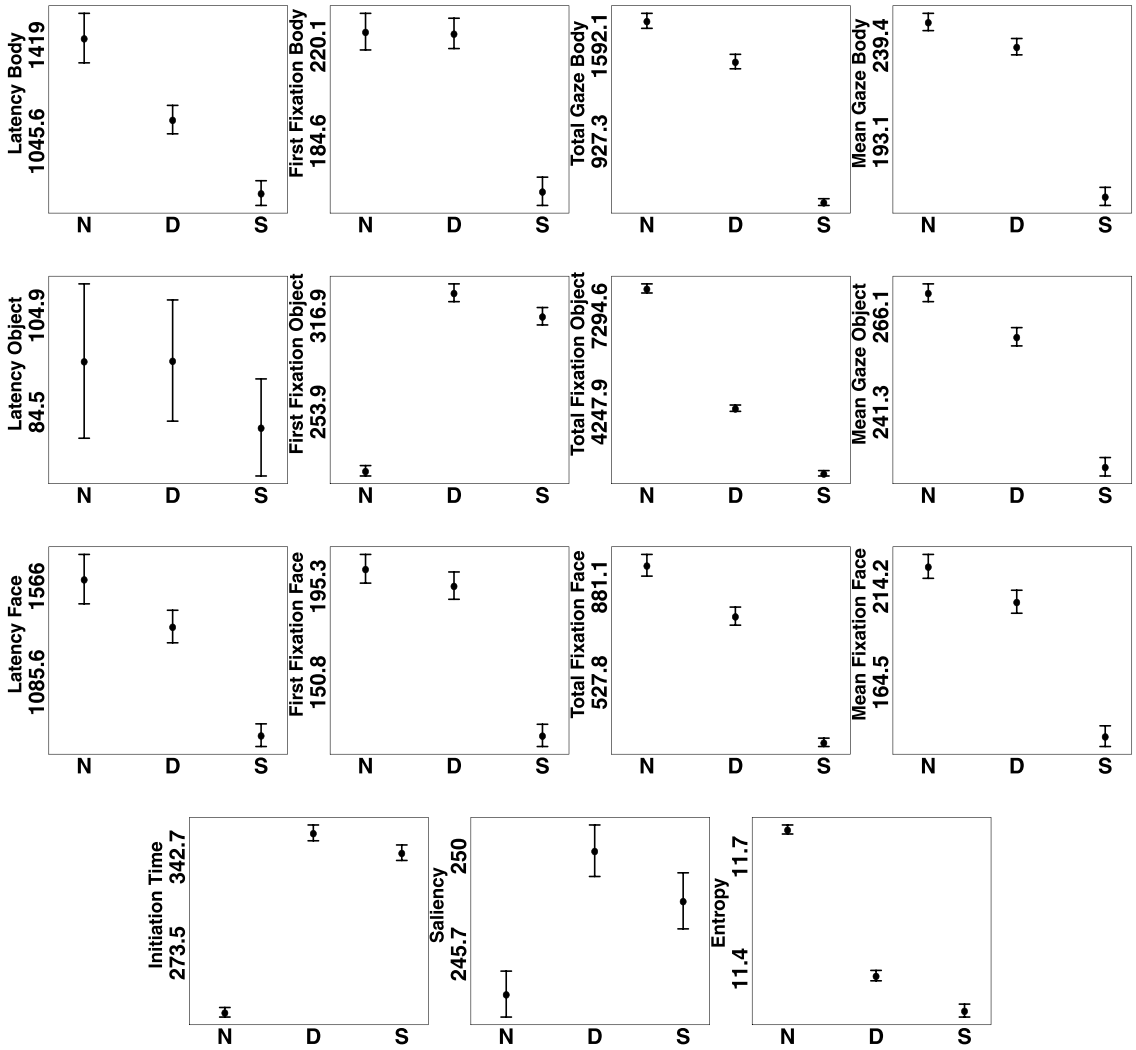

Figure 3. Mean values for the additional features that were computed. Each feature is plotted in a separate panel for the three different tasks: Naming (N), Description (D), and Search (S). The error bars indicate the standard error. The unit for all eye-movement measures is the millisecond with the exceptions of saliency, the mean value of saliency of the image at the fixation location as returned by the model of Torralba et al. (2006), and entropy.

but not much attention needs to be spent on inspecting the body (see Table 1). This contrasts with naming, the task in which most of the attention is allocated to objects in the background that can be recognized and named: An animate agent is easy to name and hence does not require much attention. This difference is, however, significant only with respect to a description task. Search and naming do not significantly differ in their dwell time on objects.

We will discuss the remaining three features more thoroughly (initiation time, saliency, and entropy) as they are not region-specific.[5] These three features can be visualized in Figure 3, in which we plot the mean and standard errors for the set of additional features considered in this study (OF).

We find that initiation time is longer in description and search, compared to naming. This is an interesting result as it shows that the processing of the cue, taking place in description and search but not in naming, increases the time required to launch the first eye movement. Presumably, the cue needs to be integrated with the scene gist to inform the first eye movement. When looking at saliency, we find that naming is the task that relies least on fixations in high-saliency areas, especially compared to the description task. As the task is to name as many objects as possible, visual attention is presumably allocated to objects that are easy to recognize rather than to objects that are salient in the scene. Finally, when looking at entropy of the attentional landscapes, we observe a pattern similar to both area of scene fixated and number of fixations (refer to Figure 2). Naming results in larger entropy than search and description as the scene needs to be explored more widely in order to identify as many objects as possible.

		LASSO			MM			SVM		
Scene clutter	Task	GF	OF	All	GF	OF	All	GF	OF	All
High	Naming	.77	.81	.82	.8	.81	.84	.81	.85	**.86**
	Description	.61	.65	.66	.65	.67	.71	.68	.71	**.74**
	Search	.8	.79	.82	.81	.81	**.83**	.82	.8	**.83**
Low	Naming	.75	.8	.8	.77	.82	.83	.79	**.86**	**.86**
	Description	.64	.66	.67	.66	.69	.74	.67	.76	**.77**
	Search	.86	.85	.86	.86	.86	**.88**	.87	.85	**.88**

Table 2. Mean *F*-score classification performance for each task (object naming, scene description, and visual search) computed over 10 folds of the eye movement using different sets of eye-movement features: Greene et al.'s (2012) features (GF), other features (OF), and all features (All). *Notes*: *F*-scores are reported for three different classifiers: least-angle regression (LASSO), multinomial logistic regression (MM), and support-vector machine (SVM). Boldface indicates the best *F*-scores achieved for each task and classifier.

Overall, we find that each task is defined by a characteristic pattern of eye-movement responses. A visual search task is characterized by long exploratory saccades, relatively short fixations to verify the object fixated against the cue, and a focused distribution of fixation on areas contextually relevant to it. In contrast, an object-naming task is characterized by shorter saccades but longer fixations as the linguistic information associated with the object fixated is also evaluated and retrieved if the object is mentioned. A naming task also triggers a spread out distribution of attention over the scene as different objects could be candidates for naming.

Even if both object naming and scene description are communicative tasks, they generate distinct eye-movement patterns. During scene description, saccades are shorter than in naming, and fixations are longer and more focused on animate objects, their bodies, and the objects with which they interact. A sentence requires not only that the linguistic labels of the visual objects are retrieved, but also that dependencies between objects are evaluated, selected, and structured into a message. The deeper involvement of language processing mechanisms in scene description could imply a poorer classification performance compared to visual search and object naming. By using only features of visual attention to train the models, we fail to capture features that relate to ongoing linguistic processing.

In summary, the large and statistically significant differences observed across task on the associated eye-movement features strongly suggest that it should be possible to classify tasks accurately based on these features.

Predicting the task using eye-movement features

In Table 2, we report the performance in terms of *F*-score for the three classifiers using different sets of features trained on two distinct data sets, which are separated by the clutter of the scene. We obtain classification performance above chance (i.e., 0.33 given that there are three tasks) using any of the classifiers in both high- and low-clutter scenes. We achieve the highest accuracy of 0.88 on the visual search task with an SVM classifier in low-clutter scenes and the lowest accuracy of 0.61 on scene description with the LASSO classifier in high-clutter scenes. In terms of the impact of the features set,[6] we find that there is no significant improvement using GF ($F = 0.76$ averaged over all three tasks) compared to OF (average $F = 0.78$), $t(177) = -1.43$, $p = 0.1$. However, using all features (average $F = 0.81$) gives a significant improvement over both GF, $t(177) = -4.18$, $p = 0.0001$, and over OF, $t(177) = -2.79$, $p = 0.005$.

When evaluating the performance of the different classifiers when all features are used, we find no significant difference between LASSO (average $F = 0.78$) and MM (average $F = 0.81$), $t(57) = -1.42$, $p = 0.1$. However, SVM (average $F = 0.84$) outperforms LASSO, $t(53) = 3.07$, $p = 0.003$, but not MM, $t(56) = 1.68$, $p = 0.09$. When comparing classification performance across the tasks using classifiers trained on all features, we find that scene description (average $F = 0.73$) is classified significantly less accurately than both visual search (average $F = 0.86$), $t(47) = -11.53$, $p < 0.0001$, and naming task (average $F = 0.84$), $t(55) = -9.01$, $p < 0.0001$. The difference between the classification accuracy for search and naming is only marginal, $t(52) = -1.75$, $p < 0.08$.

The results observed in our classification analysis are consistent with our mixed-model analysis of the eye-movement features in the previous section. Each task generates a distinctive eye-movement signature, which makes it possible to detect the task from eye-movement information. Nevertheless, Greene et al. (2012) found the opposite result, viz., they were not able to classify tasks based on eye-movement features. The crucial difference between their study and our study is presumably the nature of the tasks given to participants: Greene et al. used four purely visual tasks (memorize the picture, determine the decade the picture

	Search-correct	Description-correct	Naming-correct
Search-predicted	0	12	5
Description-predicted	8	0	8
Naming-predicted	2	13	0

Table 3. Percentage of misclassified trials using an SVM classifier trained on all features. *Notes*: The columns indicate the correct class in the test set; the rows indicate the class predicted by the classifier. For instance, Search-Correct, Description-Predicted gives the percentage of visual search instances misclassified as scene description instances.

was taken in, determine people's wealth and social context). We, on the other hand, used one purely visual task (visual search) and two communicative tasks (scene description and object naming). In communicative tasks, additional cognitive modalities are needed to achieve the tasks goals. Communicative tasks situated in a visual context demand an inspection strategy that is also informed by language processing. Thus, searching for targets happens in the service of planning, encoding, and articulating linguistic output. The necessary guidance is simpler in naming than in description as the target does not need to be contextualized; this is reflected in the fact that scene description is the most difficult task to classify (see Table 2).

Interestingly, when we look at the percentages of misclassified trials and their distribution across different tasks, we find that visual search is typically misclassified as scene description (see Table 3). It seems plausible that the fact that they are both cued paradigms and both focus on a specific target object triggers relatively similar eye-movement patterns. Object naming, in contrast, is typically misclassified as scene description, which suggests that communicative tasks tend to share more features, and hence they are more often misclassified as each other.

In summary, we find differences in classification performance by task, by feature set, and by type of classifier, but overall, our results convincingly show that tasks can be accurately discriminated by using the associated eye-movement information.

For comparability with Greene et al. (2012), we also conducted two analyses in which we trained the same three classifiers to predict the participants and the images viewed. Using the GF feature set with the multinomial log-linear neural networks model (MM) classifier, we obtained an above-chance classification performance with both participants, viz., 0.12 (when chance is $1/74 = 0.013$), and images, viz., 0.14 (when chance is $1/24 = 0.04$).[7] This result indicates that eye movements carry detailed information about both viewers and scenes, but also that classification is substantially worse than on tasks, on which we achieve a classification accuracy of almost 90% for visual search.

In the next section, we conclude the study by presenting three more analyses, which provide answers for three questions about classifying tasks given eye-movement information: (a) How many features do we need to achieve a classification performance above chance? (b) Which features are most predictive of the task being performed? (c) Can we accurately classify tasks using features that are independent of the fact that our tasks used self-termination?

Feature analysis

In Figure 4, we plot how *F*-score changes as a function of the features used. We find that, already with just one feature, the classifier is able to detect which task is performed with an accuracy above chance. Classification performance does not monotonically increase with the number of features added. Rather, we observe that accuracy peaks at eight features and then

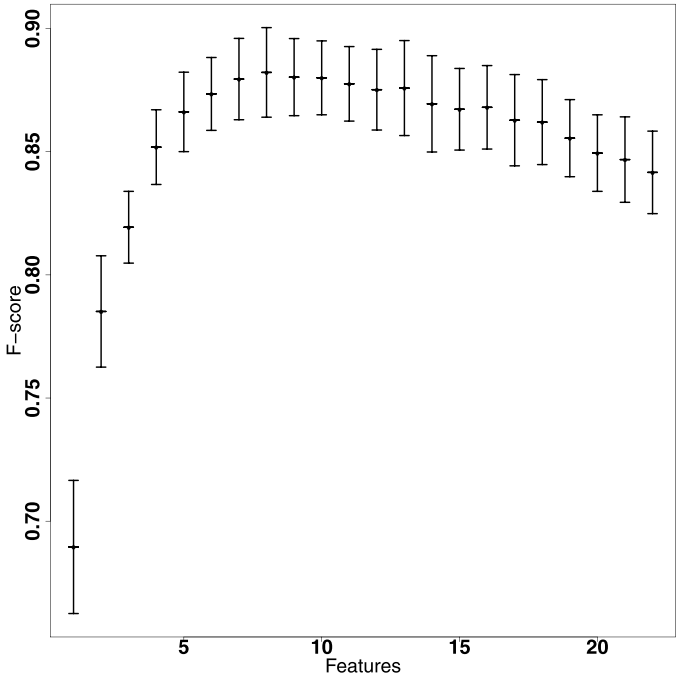

Figure 4. Mean *F*-score classification performance (y-axis) as a function of the features used (x-axis). Features are selected one at a time using a forward stepwise procedure, which includes features based on their impact on classification performance (best first). Error bars show the standard error for the *F*-score obtained from tenfold cross-validation.

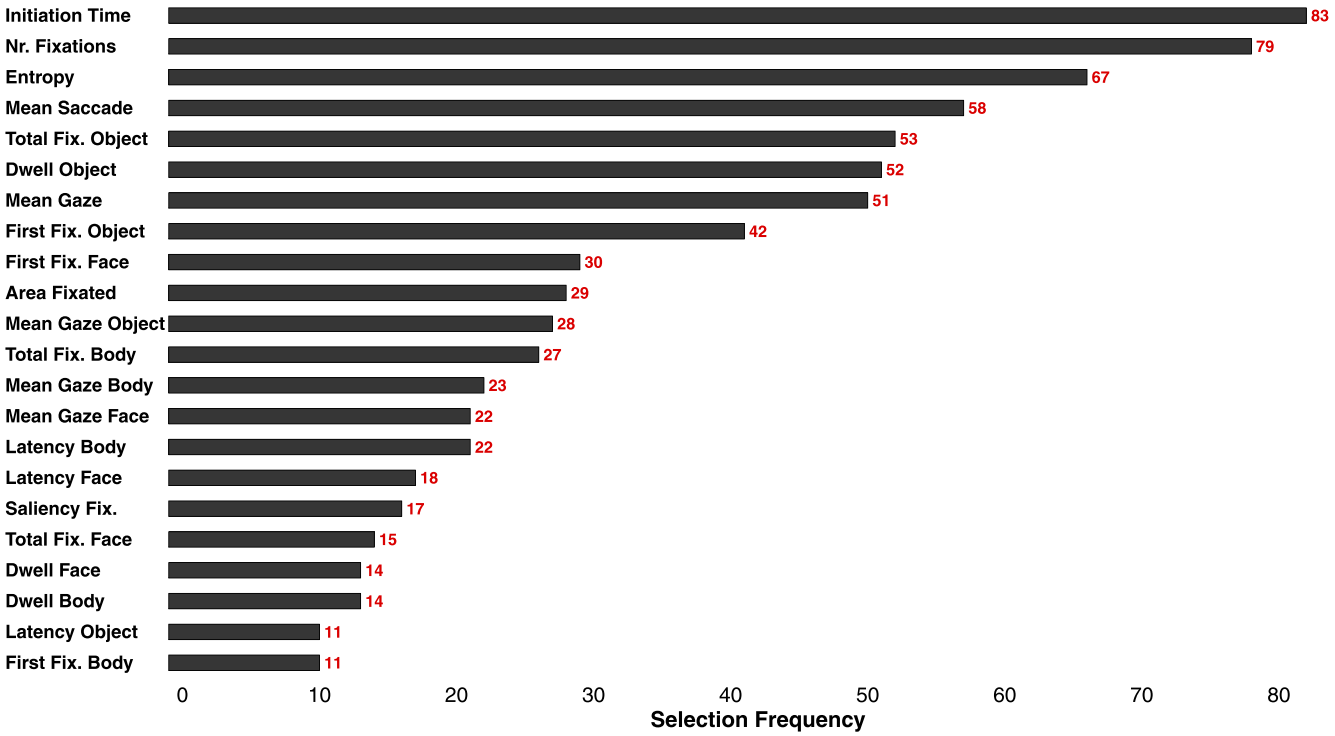

Figure 5. Bar plot showing the frequency of inclusion of each feature during forward stepwise model selection over 10 iterations and 10 folds (100 feature sets in total).

reduces slightly as more features are included. This is likely to indicate that the classifier suffers from data sparseness when it has to use a larger feature set.

When we perform stepwise selection, we evaluate whether the inclusion of a new feature improves the classification performance; if it does not, then we retain the model without the additional feature (refer to the section Methods for analysis and classification for details). Over 100 runs (10 iterations for 10 folds), the final feature sets we obtain contained an average of 7.66 $\pm$ 1.93 features, which confirms the trend observed in Figure 4. The mean *F*-score performance over the 100 runs is 0.87 $\pm$ 0.02. In order to test whether more features implies a higher classification performance, we run a linear model in which the dependent measure is the *F*-score obtained and the predictor is the size of the associated feature set. We find that *F*-scores slightly improve if the feature set is larger ($\beta_{size} = 0.001$), but this improvement is not statistically significant in a one-way ANOVA, $F(1, 98) = 2.37$, $p = 0.1$. This clearly indicates that not all features have the same importance for discriminating between tasks.

In Figure 5, we plot how frequently a feature is selected to be in the final feature set. We find that the most discriminative feature is the initiation time, i.e., the time to program the first saccade after scene onset. As discussed in the section Features Mediating Task Variability above, both visual search and scene description are cued. Initiation time indicates the time needed to integrate the cue with the gist of the scene and plan the first saccade based on this. Thus, initiation is longer for search and especially scene description compared to naming (see Figure 3). Integration time therefore constitutes a major discriminant factor for distinguishing between noncued tasks, such as naming, and cued tasks, such as description and search.

The second and third best features are the number of fixations and the entropy of the attentional landscape, i.e., features capturing the spatial distribution of visual attention. Just behind this, we find mean saccade amplitude, which reflects the exploration strategies employed to perform the task. Temporal features, such as the total number of fixations on objects, object dwell time, and mean gaze duration, also perform well, each being selected more than 50% of the time. This indicates that tasks can be differentiated in both the temporal and the spatial allocation of visual attention. It is interesting to note, moreover, that the mean visual saliency of fixations is selected only 17% percent of the time. This suggests that bottom-up scene information is accessed in similar ways for our three different tasks and hence might not be a key discriminant feature in the classification.

The final test regards the question of self-termination and the possibility that high classification performance could be driven by features related to the amount of viewing time, such as the total number of fixations. We train an SVM classifier with only two features:

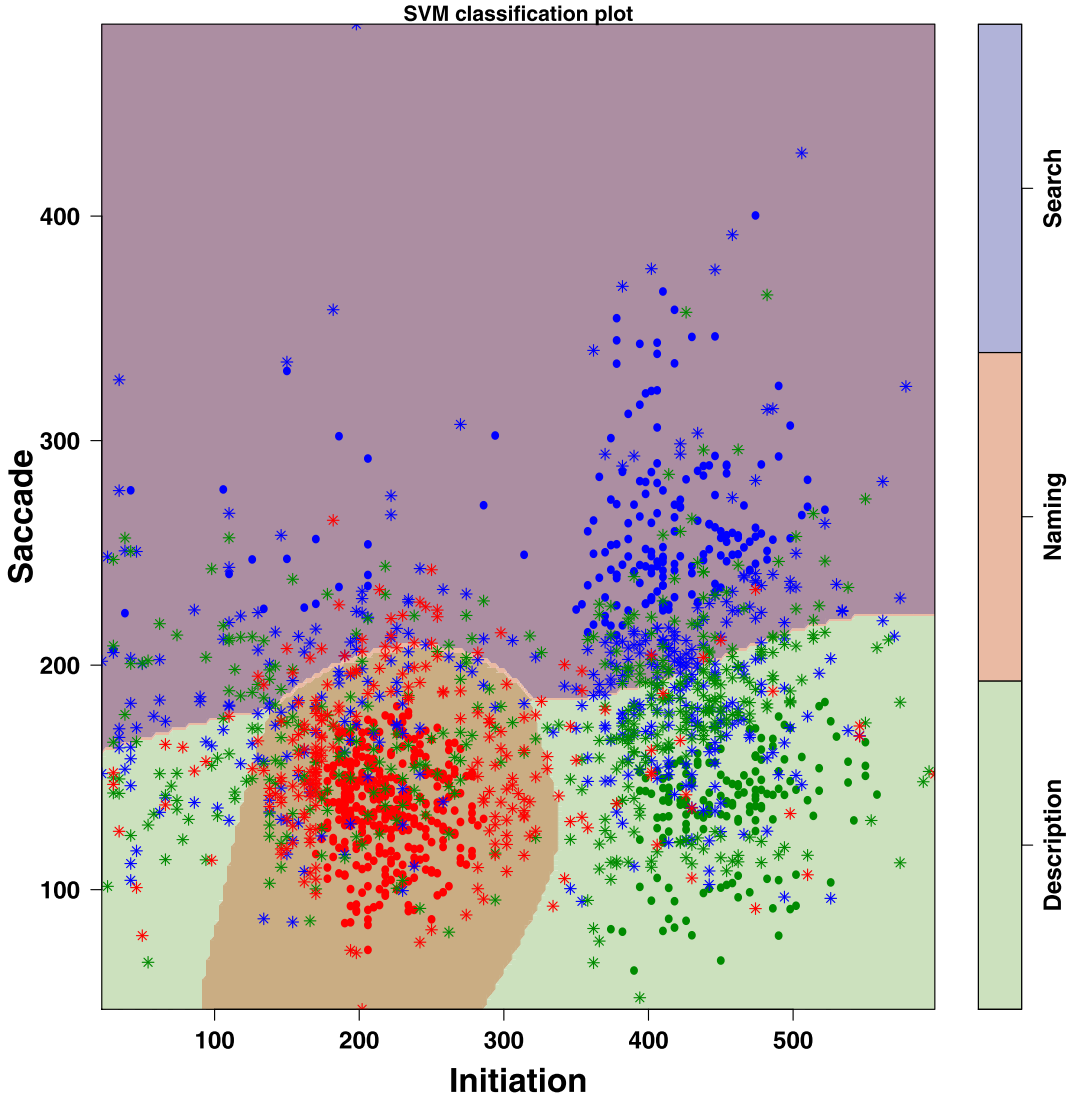

Figure 6. Scatter plot of Tasks as a function of Mean Saccade Amplitude (pixels), and Initiation Time (ms). The observations are represented as full dots, and the predicted values are represented as asterisks. The three different tasks are color coded as Description (green), Naming (red), Search (blue). The colored contours represent the SVM classification distributions for the three tasks.

initiation time and mean saccade amplitude. The first feature reflects the cuing aspect of the task, the second one the spatial span covered as the scene is explored.

We achieve a classification accuracy of 0.79 for object naming, 0.65 for visual search, and 0.63 for scene description. This result is visualized in Figure 6, in which we show a scatter-plot of tasks as a function of mean saccade amplitude and initiation time; the full dots are observations, and the asterisks are the predicted values from the SVM. We find that the naming cluster is more clearly defined compared to visual search and scene description, and this explains why we find the highest classification accuracy here. Moreover, if we drop one feature and train the classifier with either mean saccade amplitude or initiation time, we still have an accuracy above chance, i.e., 0.33. With initiation time only, we achieve an accuracy of 0.78 for object naming, 0.51 for scene description, and 0.45 for visual search.

Theoretically, this result demonstrates that tasks do not differ only in terms of their implicit timing as set by self-termination, but also in terms of other features germane to the task, such as whether prior information has to be integrated (cuing or not), or in terms of strategies used to sample the visual percept, such as the distance covered during a saccade. These task-driven routines contribute to the optimal completion of the task, e.g., long saccades during search, or serve other cognitive processes that are concurrently active, e.g., smaller saccades during language processing.

General discussion

Since very early research in visual cognition, task has played a pivotal role in formulating causal explanations for the different eye-movement responses observed.

First Buswell (1935) and then, a few decades later, Yarbus (1967), in his influential chapter "Eye-Movements during Complex Object Perception," have discussed the role of expertise, task instructions, and object knowledge in the allocation of visual attention. The qualitative analyses of scan-path trajectories during different visual tasks presented in these studies have suggested that eye-movement information includes evidence of the task performed. Hence, scan paths observed in different tasks should differ. If a task entails different goals, then the underlying cognitive processes should differ, and the scan paths should reflect this difference.

This idea has been extensively explored in subsequent research, in which visual attention was mainly investigated in the context of real-world tasks (Ballard & Hayhoe, 2009; Ballard et al., 1995; Land & Furneaux, 1997; Land & Hayhoe, 2001; Pelz & Canosa, 2001). This research demonstrates that task indeed drives eye-movement responses. In particular, eye movements are proactively employed to anticipate useful task information, e.g., looking at the spout and the kettle before pouring (Land & Furneaux, 1997). Moreover, performing a sequence of tasks, e.g., "wash hands" versus "fill a cup" and then "wash hands," can modulate the pattern of fixation observed for the same task ("wash hands" in this case, see Pelz & Canosa, 2001). The memorability of information attended to is time-locked to precise phases of the task. Changes that occurred on attended objects, for example, are detected only if the object was important for the specific action being concurrently performed (Triesch et al., 2003).

The effect of tasks extends also to other visual tasks, such as search for a target or memorization of a scene. The amount of area inspected, for example, was found to be wider in memorization than in search (Castelhano et al., 2009). Memorization benefits from a wider sampling of the scene compared to visual search, which instead focuses on precise segments of the scene, especially when animate objects are the search targets (Coco & Keller, 2009; Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Torralba et al., 2006). Task effects on visual attention are also manifested during other cognitive activities, such as reading (aloud vs. silent), and, in turn, eye movements in reading significantly differ from those observed in scene perception (Rayner, 2009).

Taken together, prior research strongly suggests a mapping between eye movements and tasks, and hence it should be possible to determine which task was performed by an observer given his or her eye-movement pattern.

Greene et al. (2012) tested precisely this hypothesis and found, contrary to expectations, that it was not possible to use eye-movement features to classify the task performed with an accuracy above chance. This study followed from DeAngelus and Pelz (2009), in which the original Yarbus (1967) study was successfully replicated, and significant differences across tasks were found. In DeAngelus and Pelz, however, an important aspect that contributed to the differences observed might have been the spontaneous self-termination of the participants. Tasks were completed at different times, and this influenced the associated eye-movement responses, e.g., total number of fixations. Thus, the fixed viewing time in Greene et al. (2012) could have flattened the variability observed in the eye-movement responses, hence making classification impossible.

The alternative hypothesis we propose is that the differences in eye-movement patterns across tasks were not large enough in Greene et al.'s (2012) study to be detected accurately. In particular, as there was no involvement of other cognitive processes beside visual attention, the different visual tasks might have had a common underlying pattern, hence making it difficult to separate them during classification. Task differences can be expected to be particularly prominent when visual attention is concurrently deployed with motor actions as discussed in the Introduction. Thus, we hypothesized that tasks need to be more distinct in their underlying goals and cognitive processes for differences in eye-movement responses to emerge.

In our study, we tested this hypothesis and assumed that relevant differences would be observed when comparing purely visual tasks (e.g., visual search) with communicative tasks (e.g., object naming and scene description) in which visual and linguistic information are processed concurrently (Cooper, 1974; Tanenhaus et al., 1995). Prior work on language processing situated in a visual context has convincingly demonstrated that visual responses are time-locked with linguistic processes during language comprehension (e.g., Altmann & Kamide, 1999; Spivey-Knowlton, Tanenhaus, Eberhard, & Sedivy, 2002) and production (e.g., Gleitman et al., 2007; Griffin & Bock, 2000). Moreover, visual and linguistic responses are so closely intertwined that it is possible to retrieve the correct sentence based on the associated scan pattern with an accuracy above chance (Coco & Keller, 2012). If eye movements carry detailed information about a sentence being heard or spoken, then they should also carry discriminant information about the task performed.

In the present study, we tested this claim using eye-movement data collected in three different tasks: visual search, object naming, and scene description. From the eye movements of each trial, we extracted the seven

features used by Greene et al. (2012) and an additional 15 new features, which we included in the analysis to widen the range of eye-movement properties investigated.

A linear–mixed effect modeling analysis showed that tasks significantly differ in a wide range of eye-movement responses, including ones that depend on the time to complete the task (e.g., total number of fixations) and ones that do not (e.g., initiation time). In particular, each task is characterized by a distinctive eye-movement pattern: Visual search requires long exploratory saccades to quickly sample the visual scene along with short fixation durations to verify object identity against a cued target. During object naming, saccades are much shorter to quickly attend as many objects as possible, and fixations are longer to retrieve and activate lexical information associated with the fixated object. Scene description also has its characteristic pattern, in which saccades are shorter than search but longer than naming. During scene description, objects mostly related to the sentence being produced are attended whereas, in naming, all objects are possible naming candidates. Fixations are longer than during search, which highlights the involvement of language processing, but they are shorter than during naming. Structuring a sentence requires a deeper involvement of language-processing mechanisms, which presumably reduces the amount of attentional resources available for processing specific objects.

The large differences observed suggest that an automatic classification of tasks using eye-movement features should be possible. Therefore, we trained three different classifiers, a least-angle regression classifier, a multinomial logistic model, and a support vector machine, with three different sets of features, Greene et al.'s (2012) features, our additional features, and both, and demonstrated that we can classify tasks using each of these feature sets with an accuracy well above chance (a maximum *F*-score of 0.88 was achieved). In a nutshell, we show that tasks are associated with distinctive eye-movement patterns and that these patterns can be successfully used to perform task classification.

This evidence for task-specific eye-movement patterns provides broad support for the active vision hypothesis, according to which task goals play a fundamental role in the allocation of visual attention. However, not all tasks can be easily classified by relying on eye-movement information. In our analysis, we found that the most difficult task to classify was scene description and speculated that, in order to achieve higher accuracy in this task, we might need to include linguistic features of the sentences produced in the training data. If the eye-movement information is dependent on concurrently processed information (such as linguistic information), then it would be insufficient to correctly characterize the observed task.

We also conducted three additional analyses to identify (a) how many features are needed to achieve maximal classification performance; (b) which of the 22 features we considered were most useful for classification; and (c) whether features independent of self-termination, such as initiation time and mean saccade amplitude, are sufficient to discriminate tasks above chance.

We found that already just one feature is able to classify the tasks with an accuracy above chance and that maximum performance was achieved with seven to eight features. When looking at which features were most important, we found that initiation time, i.e., the time to launch the first eye movement; the number of fixations; the mean saccade amplitude; the total amount of fixation on objects; and the entropy of the attentional landscape were selected as the best performing more than 50% of the time. Interestingly, these are measures covering both spatial aspects (i.e., the amount of the scene inspected) and temporal aspects (i.e., the time spent looking) of visual attention. The best performing feature overall was initiation time as it allows the classifier to distinguish between cued and noncued tasks. We concluded that, when the task is cued, the cue needs to be integrated with the visual percept of the scene (such as the gist), delaying the programming of the first saccade. This effect is more prominent for description than for search; in description, the cued target does not just need to be located in the scene, but also verbally contextualized within the scene. Finally, we showed that features that are independent of self-termination, such as initiation time and mean saccade amplitude, are also able to accurately classify the tasks. This result casts doubts on the hypothesis that the null result of Greene et al. (2012) was due to their use of fixed viewing times.

Our results open new intriguing questions regarding when different tasks show distinct patterns of eye movements on different tasks as some tasks clearly do not as per Greene et al. (2012). Future research should investigate, more specifically, how the involvement of nonvisual cognitive processes, language in our study, can influence the pattern of eye movements observed and the accuracy of task classification.

Keywords: task classification, eye-movement features, active vision, visual attention, communicative tasks

Acknowledgments

We thank Monica Castelhano, the editor, and two anonymous reviewers for their insightful feedback on earlier versions of this manuscript. European Research

Council under award number 203427 "Synchronous Linguistic and Visual Processing" to FK and Fundação para a Ciência e Tecnologia under award number SFRH/BDP/88374/2012 to MIC are gratefully acknowledged.

Commercial relationships: none.
Corresponding author: Moreno I. Coco.
Email: micoco@fp.ul.pt.
Address: Faculdade de Psicologia, Universidade de Lisboa, Lisboa, Portugal.

Footnotes

[1]We only use the two-targets condition to allow full comparability between the different experimental data sets.

[2]The images are all original, created with PhotoshopC2 using components that are in public domain sources (e.g., Flickr).

[3]In Greene et al. (2012), objects are defined as any "discrete artifact" not making up the boundary of the scene.

[4]We used only the saliency part of the model, not the context part.

[5]We do not present analyses for all features as most of them would not contribute substantially to a theoretical understanding of differences between tasks.

[6]As the classification results are very similar for the high- and low-clutter conditions, we measured the impact of the features on the models trained using a single data set.

[7]Choosing a LASSO classifier results in image classification performance close to chance with an accuracy of 0.05.

References

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73,* 247–264.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Ballard, D., & Hayhoe, M. (2009). Modeling the role of task in the control of gaze. *Visual Cognition, 17,* 1185–1204.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68(3),* 255–278.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using s4 classes*. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html.

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology and perception of art*. Chicago: University of Chicago Press.

Castelhano, M., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in real-world scenes. *Attention, Perception and Psychophysics, 72,* 1283–1297.

Castelhano, M., Mack, M., & Henderson, J. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*(3):6, 1–15, http://www.journalofvision.org/content/9/3/6, doi:10.1167/9.3.6. [PubMed] [Article]

Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2, 27,* 1–27.

Coco, M., & Keller, F. (2009). The impact of visual information on referent assignment in sentence production. In N. A. Taatgen, & H. Van Rijn (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 274–279). Amsterdam: Cognitive Science Society.

Coco, M., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, *36*(7), 1204–1223.

Coco, M., Malcolm, G., & Keller, F. (2013). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Quarterly Journal of Experimental Psychology,* (ahead-of-print), 1-25, 10.1080/17470218.2013.844843.

Cooper, R. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

DeAngelus, M., & Pelz, J. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, *17*(6–7), 790–811.

Dziemianko, M., Keller, F., & Coco, M. (2009). Incremental learning of target locations in visual search. In I. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 1729–1734). Boston: Cognitive Science Society.

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

Fletcher-Watson, S., Findlay, J., Leekam, S., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*(4), 571–583.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Glaholt, M., & Reingold, E. (2012). Direct control of fixation times in scene viewing: Evidence from analysis of the distribution of first fixation duration. *Visual Cognition*, *20*(6), 605–626.

Gleitman, L., January, D., Nappa, R., & Trueswell, J. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language, 57,* 544–569.

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

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11,* 274–279.

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.

Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7,* 498–504.

Henderson, J., Shinkareva, S., Wang, S., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PloS one, 8*(5), e64937.

Inhoff, A., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, *40*(6), 431–139.

Land, M., & Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1358), 1231–1239.

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

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

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

Malcolm, G., & Henderson, J. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision, 10*(2):4, 1–11, http://www.journalofvision.org/content/10/2/4, doi:10.1167/10.2.4. [PubMed] [Article]

Malcolm, G., & Henderson, J. (2009). The effects of target template specificity on visual search in real-world scenes. *Journal of Vision, 9*(11):8, 1–13, http://www.journalofvision.org/content/9/11/8, doi:10.1167/9.11.8. [PubMed] [Article]

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision, 11*(8):17, 1–15, http://www.journalofvision.org/content/11/8/17, doi:10.1167/11.8.17. [PubMed] [Article]

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research, 11,* 919–942.

Pelz, J., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research, 41*(25), 3587–3596.

Pomplun, M., Ritter, H., & Velichkvosky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception, 25,* 931–948.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision, 7*(2):17, 1–22, http://www.journalofvision.org/content/7/2/17, doi:10.1167/7.2.17. [PubMed] [Article]

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

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1–3), 151–173.

Salverda, A., & Altmann, G. (2011). Attentional capture of objects referred to by spoken language. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1122.

Spivey-Knowlton, M., Tanenhaus, M., Eberhard, K., & Sedivy, J. (2002). Eye movements and spoken language comprehension: Effects of syntactic con-

text on syntactic ambiguity resolution. *Cognitive Psychology,* (45), 447–181.

Tanenhaus, M. K., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 632–634.

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.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 4*(113), 766–786.

Triesch, J., Ballard, D., Hayhoe, M., & Sullivan, B. (2003). What you see is what you need. *Journal of Vision, 3*(1):9, 86–94, http://www.journalofvision.org/content/3/1/9, doi:10.1167/3.1.9. [PubMed] [Article]

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (4th ed.). Berlin, Springer.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum.