Psychological Research
https://doi.org/10.1007/s00426-018-0996-5

ORIGINAL ARTICLE

The right look for the job: decoding cognitive processes involved in the task from spatial eye-movement patterns

Magdalena Ewa Król[1] · **Michał Król**[2]

Received: 5 September 2017 / Accepted: 19 February 2018

Abstract
The aim of the study was not only to demonstrate whether eye-movement-based task decoding was possible but also to investigate whether eye-movement patterns can be used to identify cognitive processes behind the tasks. We compared eye-movement patterns elicited under different task conditions, with tasks differing systematically with regard to the types of cognitive processes involved in solving them. We used four tasks, differing along two dimensions: spatial (global vs. local) processing (Navon, Cognit Psychol, 9(3):353–383 1977) and semantic (deep vs. shallow) processing (Craik and Lockhart, J Verbal Learn Verbal Behav, 11(6):671–684 1972). We used eye-movement patterns obtained from two time periods: fixation cross preceding the target stimulus and the target stimulus. We found significant effects of both spatial and semantic processing, but in case of the latter, the effect might be an artefact of insufficient task control. We found above chance task classification accuracy for both time periods: 51.4% for the period of stimulus presentation and 34.8% for the period of fixation cross presentation. Therefore, we show that task can be to some extent decoded from the preparatory eye-movements before the stimulus is displayed. This suggests that anticipatory eye-movements reflect the visual scanning strategy employed for the task at hand. Finally, this study also demonstrates that decoding is possible even from very scant eye-movement data similar to Coco and Keller, J Vis 14(3):11–11 (2014). This means that task decoding is not limited to tasks that naturally take longer to perform and yield multi-second eye-movement recordings.

Introduction

Yarbus (1967) postulated that the patterns of eye-movements depend on the task the observer is performing. In his seminal study, he recorded eye movements of observers looking at i.e., Repin's painting. The Unexpected Visitor (1884) under seven different task conditions, such as assessing the ages or material circumstances of the people depicted in the painting. Each of the seven instructions resulted in characteristic eye-movement patterns, which demonstrated that the way we look does not only depend on the physical qualities of the stimulus but is also shaped by the task we perform. Therefore, Yarbus was the first to show that the eye-movement patterns are governed by the top-down factors such as observer's goals and task (for more details, see Tatler, Wade, Kwan, Findlay, & Velichkovsky, 2010).

The relationship between eye-movement patterns and top-down factors

However, is the relationship between the eye-movement patterns and the observer's task strong enough to reliably identify the task solely from the eye movements? This challenge, called the "inverse Yarbus" problem by Haji-Abolhassani and Clark (2014), has now been undertaken by several research teams. DeAngelus and Pelz (2009) were the first to replicate Yarbus's findings, using the same painting and a self-paced presentation and found that the resulting eye-patterns were task-dependent. Similar results were also obtained by Tatler et al. (2010). Yarbus's findings were also generalized to different stimuli and tasks. In Castelhano, Mack, and Henderson's (2009) study, participant viewed photographs depicting natural scenes under two different

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00426-018-0996-5) contains supplementary material, which is available to authorized users.

✉ Magdalena Ewa Król
mkrol1@swps.edu.pl

[1] Wrocław Faculty of Psychology, SWPS University of Social Sciences and Humanities in Wrocław, Wrocław, Poland

[2] Department of Economics, School of Social Sciences, University of Manchester, Manchester, UK

Published online: 20 February 2018

task conditions: either a visual search or memorisation. The authors reported that some of the eye metrics and the fixated image areas differed between the tasks. Similarly, Mills, Hollingworth, Van der Stigchel, Hoffman, and Dodd (2011) found that both spatial and temporal aspects of fixations are influenced by the observer's task. Going another step further, Betz, Kietzmann, Wilming, and König (2010) recorded eye-movement patterns in response to web pages viewed under three task conditions: free viewing, content awareness and information search. Using computational modelling, they showed that in that instance, top-down influences on the eye-movement patterns were independent of the salience-based bottom-up processes, indicating the predominant role of top-down factors in guiding visual attention. Finally, Kollmorgen, Nortmann, Schröder, and König (2010) quantified the effect of three determinants of overt visual attention: the bottom-up influence of visual salience, the top-down influence of task, and the effect of spatial viewing biases and oculomotor constraints. Their model revealed that all three contribute significantly and independently of one another to gaze position, but the effect of spatial constraints has the largest effect, closely followed by the top-down influence, while the effect of low-level visual salience is relatively lower.

Decoding task from eye-movement patterns

However, the mere presence of statistically significant differences between the eye-movement patterns related to different instructions does not allow to demonstrate that eye-movement data are sufficient to identify the observer's task in a particular instance.

This can be achieved using pattern recognition methods, which allow classifying each observation into predefined classes. In the first study of this kind, Greene, Liu, and Wolfe (2012) recorded eye-movements in response to 64 grayscale photographs depicting social scenes, under four different instructions: memorisation, assessment of the decade in which the picture was taken, assessment of the wealth of depicted people and assessment of the closeness of relationships between the depicted persons. Each observer viewed each stimulus only once and the assignment of tasks to stimuli was randomized between participants. However, they were not able to decode the task from eye-movement patterns above the chance level using both human observers and pattern classifiers, casting doubts on the solvability of the Yarbus's inverse problem.

Henderson, Shinkareva, Wang, Luke, and Olejarczyk (2013) recorded eye-movements in response to two different stimuli types (text and natural scenes), performing two different task with each stimuli type (reading and pseudo-reading with textual stimuli, scene search and scene memorization with the photographic stimuli). They achieved 80% task-decoding accuracy, but given that text and natural scenes have very different spatial distributions, it is possible that this level of accuracy was achieved based on stimuli and not task differences. Moreover, the tasks used in the Henderson et al.'s (2013) and Greene et al.'s (2012) studies are very different as well. In Greene et al.'s study, the differences between the tasks are very subtle and they are likely to require very similar cognitive processes. In contrast, in the Henderson et al. study, the tasks are very different in terms of both the required cognitive processes and strategies typically employed to perform them. This suggests that the ability to decode the task from the eye-movements may depend both on the type of stimuli, the tasks used, and the features selected for the model. Data from the Greene et al. (2012) were re-analysed by Borji and Itti (2014), while Coco and Keller (2014), Haji-Abolhassani and Clark (2014) and Kanan, Ray, Bseiso, Hsiao, and Cottrell (2014) performed similar studies, and all achieved above-chance level accuracy in task decoding by expanding the identification process beyond the summary statistics of eye trajectories and using more powerful computational methods (for a review, see Boisvert & Bruce, 2016).

Decoding cognitive processes from eye-movement patterns

As Borji and Itti (2014) conclude, the answer to the question of whether it is possible to identify the task from the observer's eye movements is simply: it depends. The most important factors that allow or preclude identification are: the tasks (how different they are), the stimuli (what type of information they contain), and finally, the observer (how competent they are in performing the task) (Borji & Itti, 2014). This advances the debate from the general "proof of concept" investigation to the analysis of the specific rules governing the eye-movement-based task identification.

Focusing on the task factor, Coco and Keller (2014) postulated that task decoding is possible when tasks differ in terms of the underlying cognitive processes. Cognitively similar tasks may require extraction of similar type of information from the stimulus, and thus may result in similar eye-movement patterns. Additionally, Kardan, Berman, Yourganov, Schmidt, and Henderson (2015) demonstrated that the task-related eye-movement patterns generalise across observers, which means that they reflect universal information-extraction strategies stemming from the same underlying cognitive mechanisms. Therefore, the next step would be to systematically identify characteristics of eye-movement patterns emerging from various classes of cognitive processes, thereby providing a new tool to study cognition via the eye-movement patterns. For example, Kardan, Henderson, Yourganov, and Berman (2016) demonstrated that salience-based bottom-up processing had more impact

on eye-movement patterns in the visual search task, compared to aesthetic judgment or scene-memorisation tasks.

Design

The purpose of this study was to compare eye-movement patterns elicited under different task conditions, with task differing systematically with regard to the types of cognitive processes involved in solving the tasks. We used four tasks, differing along two dimensions: global vs. local processing (Navon, 1977) and deep vs. shallow processing (Craik & Lockhart, 1972). The dimension of global/local processing concerns the hierarchy of processing of the spatial features of visual stimuli. The global features of a stimulus consist of its overall form and large elements, conveyed by low-frequency spatial information, whereas the local features consist of the details conveyed by the high-frequency spatial information. Local processing is prevalent in tasks where scrutiny of details of the stimulus is required, such as in visual search. Global processing will be dominant in tasks where a holistic judgment is required, such as the assessment of the aesthetic value of the image.

The dimension of deep/shallow processing concerns the depth of semantic processing involved in encoding the stimulus. Shallow processing involves the identification of only superficial aspects of the stimulus, such as its physical features. Deep processing involves encoding more detailed aspects of stimulus meaning, such as its identity and characteristics. The local/global processing dimension concerns the spatial features of the stimulus, whereas the deep/shallow processing concerns the semantic dimension of the stimulus. For this reason, we hypothesize that the global vs. local processing dimension will be reflected in the patterns of eye-movements to a higher extent than the deep/shallow processing dimension.

To this end, we chose four tasks, differing along both dimensions systematically. In the first task (dot task), participants were asked to determine whether the stimulus contained a small black dot, superimposed on a black and white photograph. This task required local and shallow processing, as it involved a visual search of a small detail and did not require any semantic processing of the image. In the second task (social task), participants were asked to determine whether the image contained any people. This task required local and deep processing because it involved a visual search of image detail but also semantic identification of its contents. In the third task (black or white task), participants were requested to assess whether the displayed image contained more white or more black pixels. This task required global and shallow processing because it involved making a general impression of the image, without any semantic analysis of its contents. Finally, the fourth task (indoor or outdoor task) required the participants to judge whether the image depicted an indoor or outdoor scene. It required global and deep processing, as the judgment pertained to the holistic impression of the image and semantic processing of its contents. We also added a memory test at the end of the experiment, which required participants to decide whether the displayed stimulus was shown during the experiment or not. As demonstrated by Craik and Lockhart (1972), deep processing leads to longer lasting memory traces. Thus, the purpose of the memory test was to verify that the deep processing tasks would lead to stronger memories than shallow processing tasks, to confirm that the tasks reflected the deep/shallow processing dimension.

Stimuli set contained 120 black and white photographs depicting natural scenes that were additionally degraded to increase task difficulty. Each participant saw each photograph only once, under one of four task instructions that were randomly assigned to each participant (note, however, that each task was performed in a block to allow adaptation to the task). This ensured that the observed differences in the eye-movement patterns could not be caused by either stimulus repetition-induced learning or differences between the stimuli. Additionally, we used a large set of stimuli and a large sample of participants to decrease the risk of non-generalizable patterns of results caused by idiosyncratic features of the stimulus set or participant sample.

Tasks used in our study were easier than tasks used in most of the above-mentioned studies, such as scene memorsation or assessing subtle details of the scenes such as the wealth or relationships between the depicted people. However, such complex tasks are likely to require many intertwined cognitive processes which affect eye-movement patterns in ways that cannot be easily separated from one another. Using multiple simple tasks, differing systematically with respect to cognitive processing dimensions, allows us to delineate the specific effect of underlying cognitive processes. However, such simple tasks are usually performed very quickly. If the task is resolved early, acquisition of eye data after the decision is made only adds noise to the results. For this reason, we degraded the stimuli to make them more difficult to process and we used short stimulus presentation times of 800 ms, which was also necessary to prevent participant's fatigue, given the large number of stimuli we used.

Finally, given that the purpose of the experiment was to study the influence of the task on the movement patterns, we decided not to limit the eye-movement analysis to the period of stimulus presentation. Given that tasks were performed in blocks, we would expect an adaptation of looking patterns to the task at hand. We hypothesized that participant would learn to prepare for optimal task-specific processing of the upcoming stimulus by adjusting their eye-movement patterns even before the stimulus appears. For this reason, we performed all analyses on two time windows—the period of stimulus presentation and the earlier period of presentation

of the fixation cross. There is evidence that the eye-movements displayed while imagining an object are similar to the eye-movement elicited by that object when it was originally presented (Altmann, 2004; Laeng, Bloem, D'Ascenzo, & Tommasi, 2014). For this reason, we postulated that the eye-movement patterns obtained during the presentation of the fixation cross will allow to decode the observer's task above the chance level.

Methods

Participants

Participants were 148 (105 females) volunteers, aged between 18 and 46 ($M = 23.7$; SD = 6). All participants had normal or corrected to normal eyesight. Participants took part in exchange for credits in the faculty credit system and/or 30 PLN (around 7$) per h. The study was approved by the SWPS University of Social Sciences and Humanities, Faculty of Psychology II in Wrocław Research Ethics Committee, in accordance with the 2008 version of the Declaration of Helsinki. However, the study falls short of the 2013 version of the declaration in terms of the preregistration requirement for all research studies involving human subjects. Participants provided their written informed consent to take part in the study.

Stimuli

The stimuli were 120 high-quality photographs in landscape orientation, selected from our database of 440 photographs purchased from Dreamstime (https://www.dreamstime.com/), and then processed using Adobe Photoshop CS2. All images in the database were converted to grayscale mode and their size was adjusted to fit the screen of 1280 × 720, subtending 15.9° × 27.7° of visual angle. Finally, to create degraded stimuli, photographs were treated with a "stamp" filter, that converted grays into black and white and removed high-frequency spatial information from the image. In a pilot study, we obtained recognisability ratings for each picture from 63 (25 male) participants, mean age = 25.3 (SD = 3.92), who reported whether they recognized what the image represented after a short presentation. Next, for each photograph, we calculated the proportion of black and white pixels, by dividing the number of black pixels by the total number of pixels in the image. Finally, we selected 120 stimuli from the set that fitted our criteria. Half of the images had a higher proportion of black pixels than white pixels (on average 41% of pixels were white), while the other half were predominantly white (on average 68% pixels were white). Half of the selected images contained a person (social images) and half of them did not, representing instead landscapes, objects or animals (non-social images). Half of the pictures represented indoor scenes and the other half represented outdoor scenes. There were 15 images representing each combination of the three variables (social vs. non-social, predominantly white vs. predominantly black, indoor/outdoor-each of those was additionally prepared in two versions: with and without dots). There were no statistically significant differences between the groups of images in terms of recognisability, $F(7,98) = 0.05$, $p = 1.0$. Mean recognisability was equal to 0.78 (SD = 0.11) and it was defined as the proportion of participants who reported recognizing the image after a short presentation. There was no significant difference in the proportion of white pixels between the groups for the predominantly white pictures [$F(3,42) = 2.17$, $p = .11$], where the mean proportion of white pixels was equal to 0.70 (SD = 0.08). There was no significant difference in the proportion of white pixels between the groups for the predominantly black pictures [$F(3,42) = 0.58$, $p = .63$], where the mean proportion of white pixels was equal to 0.41 (SD = 0.07). See Fig. 1 for stimuli examples.

Finally, for the dot task, we created an additional stimuli set by adding a small black dot to a random location on each image, subtending 1° visual angle. We have also selected another 16 images, with similar characteristics to the main experimental set that served as foils in the memory test. There were four images for every combination of two variables—social vs. non-social and outdoor vs. indoor.

Procedure

Participants' eye movements were recorded using a remote eye-tracking device SMI RED250Mobile, with a sampling rate of 60 Hz and gaze position accuracy of 0.4°. Participants were seated 70 cm from the computer screen. The experiment was programmed in C# and displayed on a 15" Dell Precision M4800 workstation. Participants completed a 5-point calibration and 4-point validation in-house procedure.

The experiment consisted of four tasks with order randomized for every participant, followed by a memory test. Each task consisted of 30 trials, so there were 120 trials altogether in the experimental sessions, while the memory test consisted of 32 trials. The aim of all tasks was to assess the displayed image in terms of certain property, specified at the beginning of each task. Tasks differed along two dimensions: semantic processing (shallow vs. deep) and spatial processing (global vs. local) (Fig. 2). The first task was to decide whether the displayed image contained a person, with a yes/no response choices (social task). This task was deep and local because it required processing the meaning of the image and searching for a detail of the image. The second task was to decide whether the displayed image represented an outdoor or an indoor scene, with an outdoor/

Fig. 1 Stimuli examples

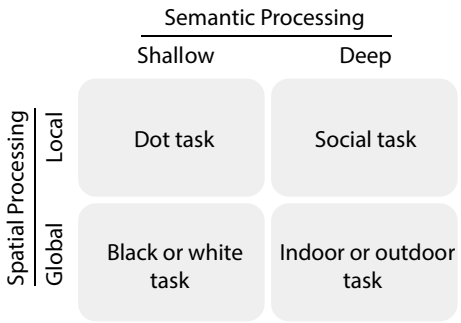

Fig. 2 Study design

indoor response choices (indoor or outdoor task). This task was deep and global because it also required recognition, but it did not involve a local search—only a global impression. The third task was to decide whether the displayed picture was predominantly white or black, with white/black as response choices (black or white task). This task was shallow and global because it required making a global judgment based on the image's visual characteristics and did not require recognition of what the image represented. Finally, the fourth task required deciding whether there was a small, black dot added somewhere to the image, with a yes/no response choices (dot task). This was local and shallow, as it involved a visual search but it did not involve processing meaning of the image.

Each participant saw all 120 images from the experimental set, but each picture was randomly assigned to one of the tasks for each participant-with the following constraints. Each picture was displayed only once. Additionally, in each task for half of the stimuli, one type of response was correct and for the other half of stimuli, the alternative response was correct. For example, in the social task, half of the displayed stimuli contained a person, and half of them did not.

In all tasks, each trial consisted of a fixation cross displayed for 300 ms, following that the experimental stimulus was displayed for 800 ms, followed by a task-specific question with two alternative choices displayed on the screen. Participants made their choice by pressing the "A" key for the response displayed on the left side of the screen, and the "L" key to choose the response on the right. The side of the screen displaying each of the alternative responses was randomized between participants. The trial ended with a blank screen displayed for 500 ms.

After all four tasks were completed, participants took part in the memory test. Half of the stimuli in the test were randomly selected from the experimental set, and as such the participants were familiar with them (targets). The other half of the stimuli were novel (foils). The order of display of stimuli was randomized. A single trial consisted of a fixation cross displayed for 500 ms, followed by the stimulus displayed for 1500 ms, and then the question: "Have you seen this pictures before?" was displayed, along with yes/no response alternatives. The trial ended with a blank screen displayed for 500 ms.

Data analysis

Behavioural data

Reaction times were measured from the onset of the question to the key press. To mitigate the effect of outliers, the data were winsorised, i.e., all values higher or lower than mean and two standard deviations were replaced with the sum of the mean RT and two standard deviations. Overall, 4.5% of reaction times were replaced this way. Both correct and incorrect responses were included in the analysis.

Due to nonparametric nature of our behavioural data, we used Friedman test to compare the accuracy and the reaction times of responses. The main test was followed by post hoc analysis using Bonferroni-corrected Wilcoxon signed-rank tests. We performed four comparisons, and therefore, we adopted alpha level of 0.0125.

Eye-tracking data

For each trial, we collected the eye samples, separately for the 300 ms period during which the fixation cross was displayed, and for the 800 ms period during which the picture was shown. We eliminated trials with more than 20% of 'bad' samples, which we defined as samples where the eye-tracker could not determine the eye-position (e.g., due to blinks). This occurred in 8.2% of trials. We excluded data from two participants because of extremely poor quality. Both eye-tracking measures were calculated for two time periods first, the duration of the target stimulus and second, for the fixation cross preceding the target stimulus.

The eye-movement dispersion was calculated as follows: for each trial and each of the two periods, we computed the average Euclidean distance between the corresponding eye samples and the centre of the screen (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka, & van de Weijer, 2011). Screen coverage was calculated using the coverage method. We calculated the proportional screen coverage based on a grid of $16 \times 9 = 144$ rectangular cells of 80×80 pixels each (Cowen, Ball, & Delin, 2002).

To mitigate the effect of outliers, the data were winsorised above the values of mean ± 2 standard deviations. Overall, between 3.3–6% of values were replaced this way, depending on the variable and the time period.

For all eye-tracking analyses, we performed a 2 (spatial processing: global vs. local) $\times$ 2 (semantic processing: shallow vs. deep) repeated-measures ANOVA, followed by post hoc analysis using Bonferroni-corrected paired samples *t*

tests. We performed four comparisons, and therefore, we adopted alpha level of 0.0125.

Results

Behavioural data

Performance in the experimental tasks

The main effect of task was significant both in case of accuracy, $\chi^2(3) = 157.28$, $p < .001$, and reaction times, $\chi^2(3) = 84.76$, $p < .001$. In general, local tasks (social and dot tasks) were easier than global tasks (black or white and indoor or outdoor), a similar pattern was also presented in the reaction times (Fig. 3a, b).

In case of reaction times, there was no significant difference between the shallow processing tasks differing in terms of spatial processing—the dot task and the black or white task, $Z = -0.28$, $p = .78$, $r = .02$, but there was a significant difference in accuracy, i.e., participants were more accurate in the dot task than in the black or white task, $Z = -6.50$, $p < .001$, $r = .54$. There was also a statistically significant difference between the deep processing tasks differing in terms of spatial processing, i.e., the social task was related to shorter reaction times than the indoor or outdoor task, $Z(3) = -8.23$, $p < .001$, $r = .68$, and also to higher accuracy, $Z = -8.93$, $p < .001$, $r = 74$. There was also a statistically significant difference between the global tasks differing in terms of semantic processing, i.e., the black or white task was related to short reaction times than the indoor or outdoor task, $Z = -2.99$, $p < .001$, $r = .25$, and also to higher accuracy, $Z = -4.12$, $p < .001$, $r = .34$. Finally, there was a statistically significant difference between the local tasks differing in terms of semantic processing, i.e., social task was related to significantly shorter reaction times than the dot task, $Z = -6.19$, $p < .001$, $r = .51$, but accuracy was not significantly different, $Z = -0.85$, $p = .39$, $r = .07$.

Performance in the memory test

There was a significant main effect of task in case of accuracy in the memory test, $\chi^2(3) = 202.26$, $p < .001$ (Fig. 3c). The difference between the deep processing tasks differing in terms of spatial processing, i.e., social and indoor or outdoor tasks, was insignificant, $Z = -0.31$, $p = .75$, $r = .03$. The performance in the memory test was better in the black or white task (shallow and global) than in the dot task (shallow and local), $Z = -4.54$, $p < .001$, $r = .38$. Conversely, participants performed better in the indoor or outdoor task (deep and global) than in the black or white task (shallow and global), $Z = -7.76$, $p < .001$, $r = .64$. They also performed

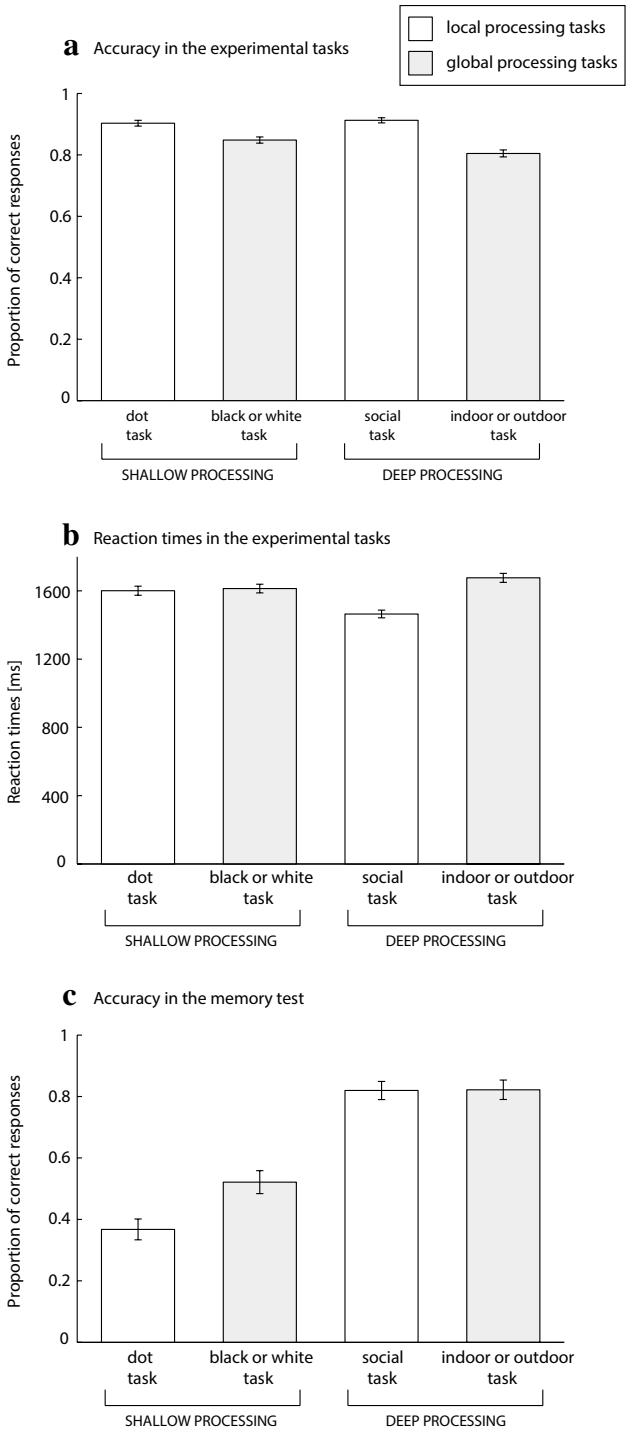

Fig. 3 Behavioural data in the study. **a** Accuracy in the experimental tasks. **b** Reaction times in the experimental tasks. **c** Accuracy in the memory test

better in the social task (deep and local) than in the dot task (shallow and local), $Z = -9.74$, $p < .001$, $r = .81$.

Eye-tracking data

Fixation cross

Eye-movement dispersion Global tasks were related to significantly lower dispersion than local tasks, $F(1,145) = 70.21$, $p < .001$, $\eta_p^2 = 0.33$ (Fig. 4a). Deep processing tasks were related to significantly lower dispersion than shallow tasks, $F(1,145) = 6.07$, $p = .02$, $\eta_p^2 = 0.04$. The interaction between spatial and semantic processing was not significant, $F(1,145) = 3.72$, $p = .06$, $\eta_p^2 = 0.03$. Post hoc tests revealed significant differences between the two deep processing tasks differing along the spatial dimension (i.e., the social task was related to higher dispersion than the social task) [$t(145) = 8.08$, $p < .001$, $d = 0.67$], and between the two shallow processing tasks differing along the spatial dimension (i.e., the dot task was related to higher dispersion than the black or white task) [$t(145) = 4.54$, $p < .001$, $d = 0.38$]. The differences between the two local tasks (differing in terms of semantic processing) [$t(145) = 0.52$, $p = .60$, $d = 0.04$], and two global tasks (differing in terms of semantic processing) [$t(145) = 3.21$, $p = .002$, $d = 0.27$], were insignificant.

Screen coverage There was no significant effect of spatial processing on screen coverage, $F(1,145) = 2.28$, $p = .13$, $\eta_p^2 = 0.02$ (Fig. 4c). Deep processing tasks were related to significantly higher coverage than shallow tasks, $F(1,145) = 10.51$, $p = .001$, $\eta_p^2 = 0.07$. However, the interaction between spatial and semantic processing was significant, $F(1,145) = 5.23$, $p < .02$, $\eta_p^2 = 0.04$.

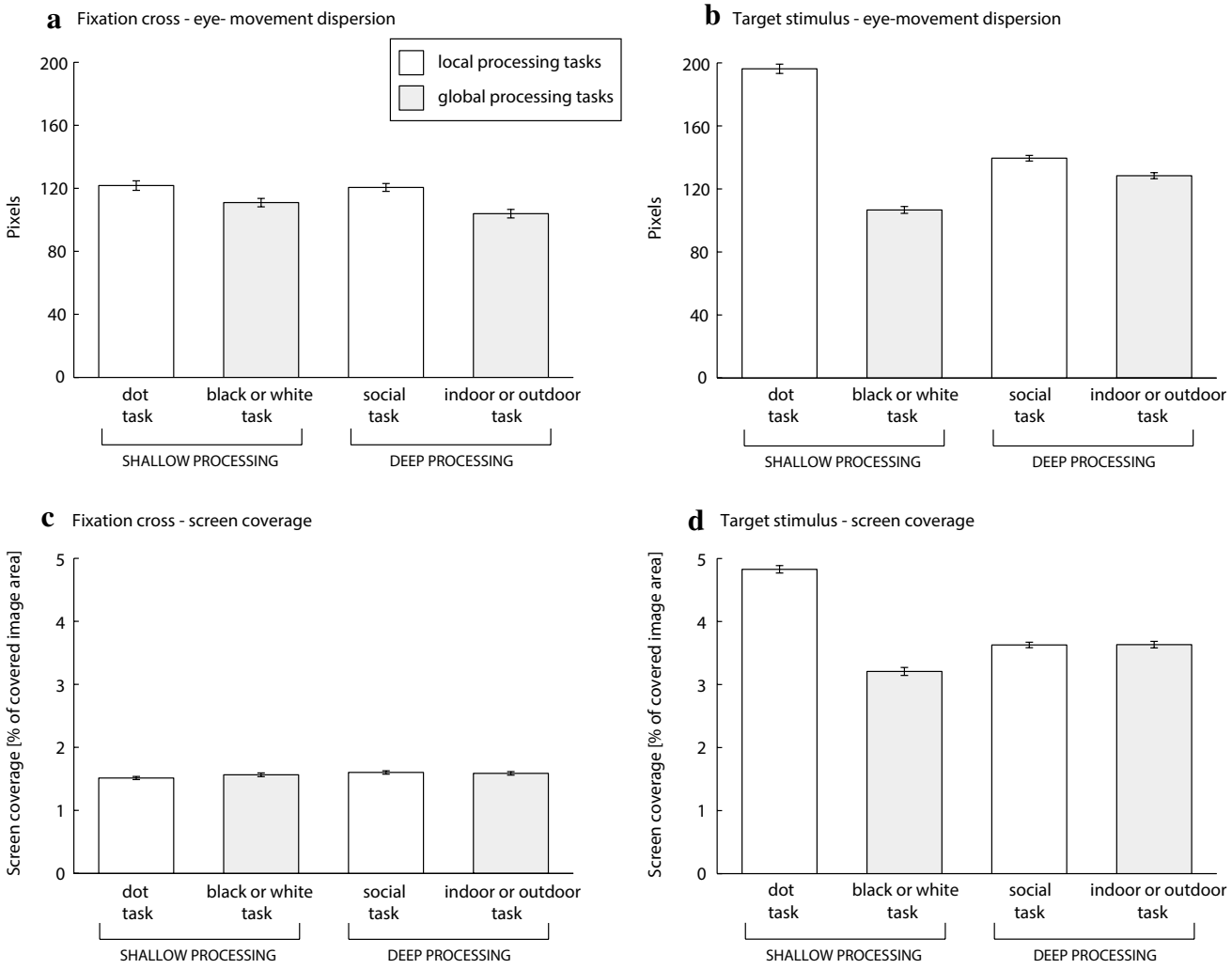

Fig. 4 **a** Eye-movement dispersion for the fixation cross period. **b** Eye-movement dispersion for the target stimulus presentation period. **c** Screen coverage for the fixation cross period. **d** Screen coverage for the target stimulus presentation period

Post hoc comparisons revealed no significant differences between the shallow processing tasks (differing in the spatial aspect) [$t(145) = 2.29$, $p = .02$, $d = 0.19$], the deep processing tasks (differing in the spatial aspect) [$t(145) = 0.64$, $p = .53$, $d = 0.05$], the global tasks (differing in the semantic dimension) [$t(145) = 0.99$, $p = .32$, $d = 0.08$]. The differences between the two local tasks (differing in the semantic dimension) was significant (i.e., the social task was related to higher coverage), $t(145) = 4.28$, $p < .001$, $d = 0.35$.

Target stimulus

Eye-movement dispersion Global tasks were related to significantly lower dispersion than local tasks, $F(1,145) = 1473.54$, $p < .001$, $\eta_p^2 = 0.91$ (Fig. 4b). Deep processing tasks were related to significantly lower dispersion than shallow tasks, $F(1,145) = 168.66$, $\eta_p^2 = 0.54$. The interaction between spatial and semantic processing was also significant, $F(1,145) = 872.71$, $p < .001$, $\eta_p^2 = 0.86$. All post hoc comparisons revealed significant differences, between the two local tasks (differing in the semantic aspect, i.e., the dot task was related to higher dispersion than the social task) [$t(145) = 26.86$, $p < .001$, $d = 2.22$], between the two global tasks (differing in the semantic aspect, i.e., the indoor or outdoor task was related to higher dispersion than the black or white task) [$t(145) = 13.28$, $p < .001$, $d = 1.10$], between the two deep processing tasks (differing in the spatial aspect, i.e., the social task was related to higher dispersion than the indoor or outdoor task) [$t(145) = 8.46$, $p < .001$, $d = 0.70$], and between the two shallow processing tasks (differing in the spatial aspect, i.e., the dot task was related to higher dispersion than the black or white task) [$t(145) = 39.14$, $p < .001$, $d = 3.24$].

Screen coverage Global tasks were related to significantly lower coverage than local tasks, $F(1,145) = 598.71$, $p < .001$, $\eta_p^2 = 0.81$ (Fig. 4d). Deep processing tasks were related to significantly lower coverage than shallow tasks, $F(1,145) = 139.66$, $p < .001$, $\eta_p^2 = 0.49$. The interaction between spatial and semantic processing was also significant, $F(1,145) = 663.06$, $p < .001$, $\eta_p^2 = 0.82$. Post hoc comparisons revealed no significant difference between the two deep processing tasks (differing in the spatial aspect), $t(145) = 0.14$, $p = .89$, $d = 0.01$. All other post hoc comparisons revealed significant differences: between the two local tasks (differing in the semantic aspect, i.e., the dot task was related to higher coverage than the social task) [$t(145) = 28.84$, $p < .001$, $d = 2.39$], between the two global tasks (differing in the semantic aspect. i.e., the indoor or outdoor task was related to higher screen coverage than the black or white task) [$t(145) = 8.64$, $p < .001$, $d = 0.72$], and between the two shallow processing tasks (differing in the spatial aspect, i.e., the dot task was related to higher coverage than the black or white task) [$t(145) = 30.20$, $p < .001$, $d = 2.50$].

Task classification

For each trial and each of the two periods, we further split that period into three subsequent sub-periods/time-bins of equal length. We then calculated the spatial median of horizontal–vertical eye-positions obtained from samples corresponding to that subperiod. Thus, for each trial and each of the two periods, we obtained three spatial medians, each being the point minimizing the distance between itself and the other samples of eye-position in the same time-bin. In addition, for each trial and each of the two periods, we included the two variables described in the statistical analysis, i.e., gaze dispersion and screen coverage.

Thus, for each trial/period, this gives a total of 3×2 (three spatial median horizontal/vertical position) $+ 2 = 8$ numbers. Each vector of nine numbers, together with the task number (1–4) attempted in the trial, constituted a single data point, were the data points corresponding to the two time periods were analysed separately, as detailed below. Additionally, the eye movement data for each participant and each feature were standardised, so between-participants variance decreased by reducing idiosyncrasies of individual participants.

We sought to predict the task number based on the 9-number input vector encoding the concurrent eye data. To this end, we used a feed forward neural network classifier with one input layer of nine nodes, one for each input variable, one hidden layer of fifteen nodes and one output layer of four nodes, one for each type of task.

We used a rectified linear hidden layer activation function, L2 regularization, and cross-validate the classification algorithm in the following manner:

First, we collected all data points corresponding to the given period, i.e., either the 300 ms or the 800 ms period. Next, we separated the data points corresponding to one of the subjects to form a 'testing set', with the rest of the data comprising a 'training set'. The latter was used to train the neural network. During training, each case in the training data was presented to the model, and the weights of the neural network were adjusted to fit the (known) true classes of the training cases (i.e., the actual task numbers). Afterwards, the accuracy of the trained neural network was evaluated on the previously unseen 'testing set'. In other words, we predicted the task number from the data of one of the subjects using a model trained on data from all other subjects. We repeated this process for each of our 148 subjects, and reported the overall cross-validated classification accuracy, separately for the 300 and the 800 ms period.

For the fixation cross that preceded the target stimulus, we achieved accuracy of 34.8% (nonsignificant vs. chance;

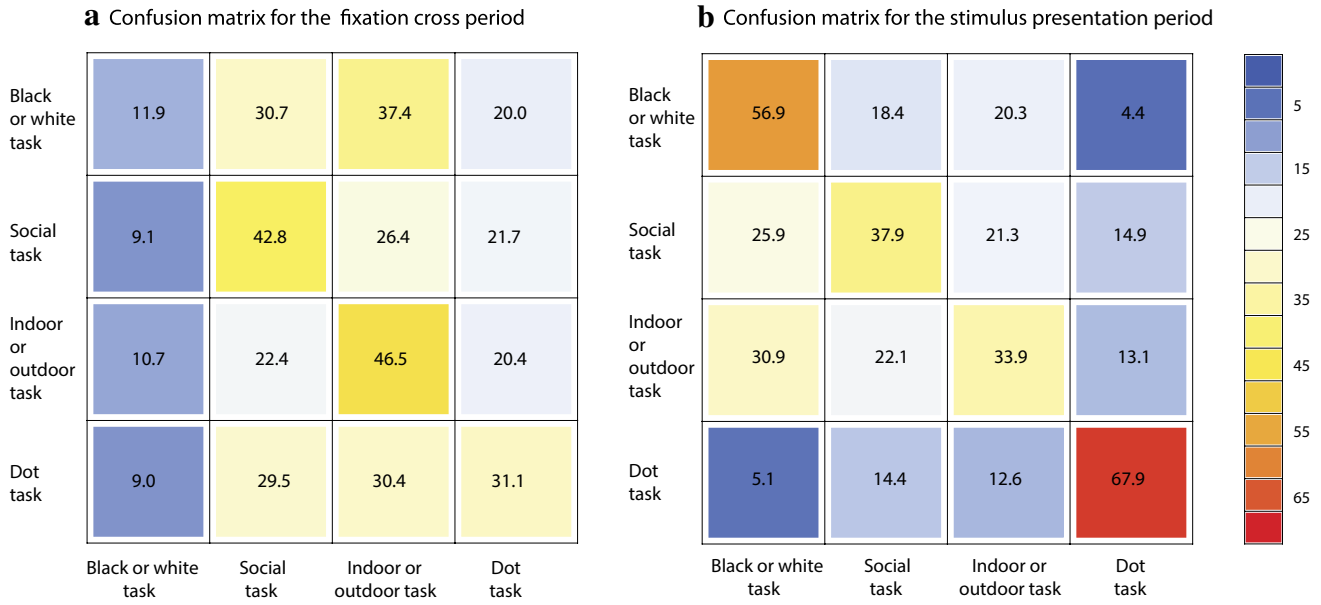

Fig. 5 Confusion matrices for **a** the fixation cross period and **b** the stimulus presentation period

binomial test, $p < .001$). For the target stimulus, we achieved accuracy of 51.4%, binomial test, $p < .001$ (for the confusion matrices see Fig. 5). Finally, combining both time periods resulted in 51.5% classification accuracy (binomial test, $p > .001$). Additionally, we compared classification accuracy using three different models: neural network, support vector machine and gradient boosting trees (Online Resource 2) and report both overall classification accuracy and *F*-scores for each task. We also tried different combinations of features, to check which contribute the most to the classification accuracy. Finally, to understand better the result regarding decoding the task using eye-movement data from the fixation cross period, we performed correlation analyses of eye-movement data pre- and post-stimulus presentation. We found significant correlations between the fixation cross period and the target stimulus period (at $p < .005$) for 81 (56%) participants for eye-movement dispersion, and for 8 (5%) participants for screen coverage.

Discussion

We compared spatial eye-movement patterns for four tasks, differing along two dimensions: semantic processing (deep vs. shallow) and spatial processing (global vs. local). We used eye-movement patterns obtained from two time periods: fixation cross preceding the target stimulus and the target stimulus. We found above chance task classification accuracy for both time periods. The aim of the study was not only to demonstrate whether eye-movement-based task decoding was possible but also to investigate whether eye-movement patterns can be used to identify cognitive processes behind the tasks.

The choice of tasks

Of course, there is the question of how well the tasks represented the two factors: spatial and semantic processing. Behavioural results allowed us to compare the tasks in terms of their level of difficulty and the subsequent memory strength.

The pattern of behavioural results suggests that social task was the easiest (both accuracy and reaction times), while the indoor–outdoor task was the most difficult, with the longest reaction times and lowest accuracy. The differences between the dot and black and white task were insignificant in terms of reaction times, but the dot task had significantly higher accuracy. This pattern of results suggests that task difficulty was not determined by either the global/local factor or the deep/shallow processing factor, but other variables unaccounted for in the study (such as task difficulty, temporal characteristics of task completion, the average size of the elements of the image required to complete specific task). Ideally, the level of difficulty in the tasks should be similar because differences in task difficulty could lead to differences in eye-movement patterns, and as such be a confounding factor. However, previous similar studies did not control for this factor either. Nonetheless, controlling task difficulty in studies with similar design and purpose might be advisable in the future, at least in some cases, even though such fine-tuning of the difficulty level in very different tasks would not always be easy and sometimes not even

possible. Some tasks require a different type of responses—for example, performance in a memorisation task cannot be quantified using reaction times straight after the stimulus presentation. In many other cases, either the processing time or the level of difficulty is naturally different.

However, the differences in performance in the memory test are clearly related to the deep/shallow processing factor. Both deep processing tasks are related to significantly higher accuracy than the shallow processing tasks. Interestingly, there is a significant difference between the dot and the black or white tasks. The images displayed within the dot task were remembered least accurately of all four tasks. This task was the only one, where the image itself did not matter for performing the task, participants were requested to find a dot that was added to the image. In contrast, in the black or white task, participants had to judge the amount of black and white in the image. These results thus show that even superficial processing of the image leads to better memory than finding an element superimposed on the image, where the image itself does not have to be processed. Overall, this pattern of results confirms that level of processing influences the memory for pictures.

Another question is how well did the spatial processing tasks reflect the global/local dimension? First, visual search in the 'local processing' tasks (the dot task and the social task) certainly did incorporate elements of both local and global processing. Second, the tasks differed from the classic Navon task to the extent that they may have led to different eye-movement patterns that we would expect from a classic global/local letter task. For example, it could be argued that focusing on the global aspect of such stimulus (the big letter) requires more dispersed-looking patterns than focusing on one of the local letters. However, what is important is that, in both cases, we would expect that differences in spatial processing lead to different patterns of looking.

To summarize, the tasks differed in terms of difficulty, but these differences most likely stemmed from some variables other than the variation within the experimental factors. However, the results of the memory test are consistent with the research on the depth of processing, given that the deep processing tasks were related to significantly better performance in the memory task than the shallow processing tasks. We also point out the importance of controlling for task difficulty.

Are spatial and semantic processing reflected in the eye-movement patterns?

Our hypothesis that spatial processing dimension would have a larger impact on the eye-movement patterns than semantic dimension was confirmed. The effect of spatial processing was larger than the effect of semantic processing dimension both in case of eye-movement dispersion and screen coverage. Given that both of these are spatial measures, the result is not surprising. Naturally, local processing task will elicit wider gaze spread than global processing task, where only a general impression is usually needed. Both eye-movement dispersion and screen coverage were the highest in the dot task, where the dot was placed in a random location of the image. This is especially interesting with regard to the other local processing task, i.e., the social task because it shows that in that task, the eye-movement patterns were restricted by the knowledge of natural image statistics. The dot in the dot task could have been hidden in any location in the image, but given the natural image regularities, the position of a person in the image was more probable in certain areas, which may be the reason for lower dispersion and coverage. Similarly, in the Ehinger, Hidalgo-Sotelo, Torralba, and Oliva (2009) study, participants searching for pedestrians in natural images consistently fixated similar areas of the image, where pedestrians' presence was more probable.

Alternatively, this result could be simply caused by the relative difference in search targets in the two tasks—the dot was smaller than people appearing in the images. This is also reflected in the behavioural measures of task difficulty—the social task was related to higher accuracy and lower reaction times than the dot task. On the other hand, objects that do not belong to a scene naturally attract attention (Friedman, 1979; Loftus & Mackworth, 1978), so it could be argued that the task with a dot superimposed on the image would be easier in that respect. Thus, we return to the issue of task control. For this reason, even though there was a statistically significant effect of semantic processing, we think that such direct interpretation ought to be treated with caution. The dot task was related to much higher dispersion and coverage than the other three tasks, which led to elevated mean dispersion and coverage for the shallow processing tasks, compared to the deep processing tasks. To summarise, the results of the study demonstrate that spatial processing is reflected in the eye-movement patterns. However, even though we obtained a statistically significant effect of semantic processing on the eye-movement data, this result may be an artefact of insufficient task control.

Task decoding

For the four tasks used in the study, we obtained classification accuracy of 51.4% using eye-movement data from the period of stimulus presentation and 34.8% using the data from the period of fixation cross presentation, i.e., before the target stimulus was presented and 51.5% suing both time windows. In all cases, accuracy is above the chance level. The confusion matrices (Fig. 5) present the percentage of trials in each task classified correctly (the diagonal line) and incorrectly, as one of the other three tasks. For example, the maximum accuracy in the study was obtained for the dot

task, using the data from the period of stimulus presentation, where 69.5% of trials were classified correctly. The confusion matrices also allow us to identify which tasks were often confused with one another—for example, we can see that the two global tasks were often mistaken for one another, which suggests that the semantic processing level may not be enough to differentiate between the tasks.

Of course, given the differences between tasks used in this study and previous studies (such as Borji & Itti, 2014), direct comparison of accuracy would not be meaningful. However, the results we obtained provide additional evidence in support of the solvability of the inverse Yarbus problem, at least in some cases. Additionally, we show that task can be decoded from the eye-movement patterns recorded over a very short period of time (800 ms), and even to some extent, before the stimulus is presented. Borji and Itti (2014) reported that in their study, task decoding accuracy was higher for early fixations. Moreover, Coco and Keller (2014) have also achieved remarkable above-chance classification accuracy with very scant eye-movement data—i.e., the initiation time, which is the time spent to launch the initial saccade after stimulus presentation.

This suggests the particular importance and information richness of the early period of stimulus presentation in revealing the task. However, given that the task is known to the observer even before the stimulus is displayed, we expected task-specific eye-movement patterns reflecting the task-solving strategy adopted by the observer in the period preceding stimulus presentation. Even though the overall accuracy for the fixation cross period was naturally lower than for the period of stimulus presentation, it was still significantly above chance level. Moreover, for the shallow processing tasks, classification was more accurate for the target stimulus presentation period than the fixation cross period. However, for the deep processing tasks, the opposite was the case. Task decoding accuracy was higher for the extremely short (300 ms) period of fixation cross than for the period of target stimulus presentation. This suggests that anticipatory eye-movements reflect the visual scanning strategy employed for the task at hand. The fact that accurate classification based on data recorded before the stimulus presentation is interesting enough on its own, but classification accuracy for the deep processing tasks was actually higher for the period when the stimulus was absent than for the period when it was present. This suggests that at least in some circumstances, the presence of visual input actually occludes the pattern of the top-down factors imprinted on the eye-movement patterns. When the stimulus is present, the impact of task-solving strategy on the eye-movement patterns is distorted by the spatial characteristics of the stimulus. The question is, why combining the data from both time-periods resulted only in a very small improvement in classification power? Our speculation would be that the reason is that there is a lot of similarity between eye-movement data pre-stimulus and at the very beginning of stimulus presentation. Perhaps then, the very beginning of stimulus presentation may capture the anticipatory EM patterns from the pre-stimulus period.

However, it is also possible that this result is an artefact of eye-repositioning after the presentation of the target stimulus. If the previous stimulus required more dispersed looking patterns, then it is possible that repositioning the eyes to the centre of the image (i.e., to the fixation cross) resulted in more eye movement dispersion during the subsequent fixation cross presentation, in a "trickle-down" effect. Even though there was a blank screen (as a rest period) displayed for 500 ms between stimulus presentation and the fixation cross, it is still possible that, even given this additional time to re-position the eyes after the stimulus disappeared, repositioning might have continued even during the fixation cross presentation. The design of the current experiment does not allow exclusion of this possibility. However, the differences between the patterns of dispersion and screen coverage between the fixation cross and target presentation speak at least to some extent against this possibility. For example, the dot task stands apart among the other tasks in terms of both the highest dispersion and screen coverage in the target presentation period. However, this pattern is not present in the fixation cross period. The dot task is related to only slightly higher dispersion and actually has the lowest screen coverage. Moreover, for the fixation cross period, deep processing tasks are related to significantly higher screen coverage (compared to shallow processing tasks), while for the target presentation period, deep processing are related to significantly lower coverage. If this was purely a "trickle-down" effect caused by eye-repositioning after the target stimulus presentation, we would expect eye-movement patterns in the fixation cross to closely mimic the pattern for the target stimulus presentation period. Given that this was not the case, we may cautiously conclude that the observed effect is not entirely an artefact.

Conclusion

To summarise, this study and previous studies (Borji & Itti, 2014; Coco & Keller, 2014; Haji-Abolhassani & Clark, 2014; Henderson & Hollingworth, 1999; Kanan et al., 2014; Kardan et al., 2015, 2016) provide evidence that at least for some tasks, it is possible to decode task from eye-movement patterns. Thus, research in this area can move beyond the "proof of concept" stage, to the next task of establishing the conditions that make decoding possible. Specifically, the question is what kinds of tasks are decodable and which eye-movement measures are best suited to revealing a specific type of task. Ultimately, the most important issue is which

cognitive processes are reflected in the eye-movement patterns and which do not reveal themselves in the way the eyes move. However, this can be achieved only with a very strict control of any potential confounds in the experimental tasks. So far, tasks selected for comparison in similar studies were not specifically controlled, because the aim of these studies was to investigate whether eye-movement-based task decoding was at all possible.

Additionally, we show that decoding is possible even for very short stimulus presentation. This is very important because it means that task decoding is not limited to tasks that naturally take longer to perform and yield multi-second eye-movement recordings. Finally, we also show that task can be to some extent decoded from the preparatory eye-movements before the stimulus is displayed.

Funding This work was supported by the National Science Centre in Poland under Grant 2013/11/D/HS6/04683.

References

Altmann, G.T.M. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition, 93*(2), B79–B87. https://doi.org/10.1016/j.cognition.2004.02.005.

Bernstein, L.J., Beig, S., Siegenthaler, A.L., & Grady, C.L. (2002). The effect of encoding strategy on the neural correlates of memory for faces. *Neuropsychologia, 40*(1), 86–98. https://doi.org/10.1016/S0028-3932(01)00070-7.

Betz, T., Kietzmann, T.C., Wilming, N., & König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision, 10*(3), 1–14. https://doi.org/10.1167/10.3.15.

Boisvert, J.F.G., & Bruce, N.D.B. (2016). Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing, 207*, 653–668. https://doi.org/10.1016/j.neucom.2016.05.047.

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision, 14*(3:29), 1–22. https://doi.org/10.1167/14.3.29.

Bower, G.H., & Karlin, M.B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology, 103*(4), 751–757. https://doi.org/10.1037/h0037190.

Castelhano, M.S., Mack, M.L., & Henderson, J.M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*(3), 6–6. https://doi.org/10.1167/9.3.6.

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

Cowen, L., Ball, L.J., & Delin, J. (2002). An eye movement analysis of web page usability. In *People and computers XVI—memorable yet invisible* (pp. 317–335). London: Springer. https://doi.org/10.1007/978-1-4471-0105-5_19.

Craik, F.I.M., & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X.

DeAngelus, M., & Pelz, J.B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition, 17*(6–7), 790–811. https://doi.org/10.1080/13506280902793843.

Ehinger, K.A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17*(6–7), 945–978. https://doi.org/10.1080/13506280902834720.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology General, 108*, 316–355.

Grady, C.L., McIntosh, A.R., Rajah, M.N., & Craik, F.I.M. (1998). Neural correlates of the episodic encoding of pictures and words. *Proceedings of the National Academy of Science USA, 95*, 2703–8.

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. https://doi.org/10.1016/j.visres.2012.03.019.

Haji-Abolhassani, A., & Clark, J.J. (2014). An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision Research, 103*, 127–142. https://doi.org/10.1016/j.visres.2014.08.014.

Henderson, J.M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243–271. https://doi.org/10.1146/annurev.psych.50.1.243.

Henderson, J.M., Shinkareva, S.V., Wang, J., Luke, S.G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PLoS ONE, 8*(5), e64937. https://doi.org/10.1371/journal.pone.0064937.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Intraub, H., & Nicklos, S. (1985). Levels of processing and picture memory: The physical superiority effect. *Journal of Experimental Psychology Learning, Memory, and Cognition, 11*(2), 284–298. https://doi.org/10.1037/0278-7393.11.2.284.

Kanan, C., Ray, N.A., Bseiso, D.N.F., Hsiao, J.H., & Cottrell, G.W. (2014). Predicting an observer's task using multi-fixation pattern analysis. In proceedings of the symposium on eye tracking research and applications—ETRA'14 (pp. 287–290). New York, USA: ACM Press. https://doi.org/10.1145/2578153.2578208.

Kardan, O., Berman, M.G., Yourganov, G., Schmidt, J., & Henderson, J.M. (2015). Classifying mental states from eye movements during scene viewing. *Journal of Experimental Psychology Human Perception and Performance, 41*(6), 1502–1514. https://doi.org/10.1037/a0039673.

Kardan, O., Henderson, J.M., Yourganov, G., & Berman, M.G. (2016). Observers' cognitive states modulate how visual inputs relate to gaze control. *Journal of Experimental Psychology Human Perception and Performance, 42*(9), 1429–1442. https://doi.org/10.1037/xhp0000224.

Kollmorgen, S., Nortmann, N., Schröder, S., & König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology, 6*(5), e1000791. https://doi.org/10.1371/journal.pcbi.1000791.

Laeng, B., Bloem, I.M., D'Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition, 131*(2), 263–283. https://doi.org/10.1016/j.cognition.2014.01.003.

Lockhart, R.S., & Craik, F.I. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 44*(1), 87–112. https://doi.org/10.1037/h0084237.

Loftus, G.R., & Mackworth, N.H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology Human Perception and Performance, 4*, 565–572.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M.D. (2011). Examining the influence of task set on eye

movements and fixations. *Journal of Vision, 11*(8), 17–17. https://doi.org/10.1167/11.8.17.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3.

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

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