Vision Research 106 (2015) 47–57

Contents lists available at ScienceDirect

Vision Research

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

Goal-oriented gaze strategies afforded by object interaction

Anna Belardinelli [a,*], Oliver Herbort [b], Martin V. Butz [a]

[a] Computer Science Department, University of Tübingen, Germany
[b] Department of Psychology, University of Würzburg, Germany

ARTICLE INFO

Article history:
Received 9 May 2014
Received in revised form 29 October 2014
Available online 21 November 2014

Keywords:
Task-driven perception
Eye-tracking
Object interaction
Fixation distribution
Eye-hand coordination
Movement preparation

ABSTRACT

Task influence has long been known to play a major role in the way our eyes scan a scene. Yet most studies focus either on visual search or on sequences of active tasks in complex real world scenarios. Few studies have contrasted the distribution of eye fixations during viewing and grasping objects. Here we address how attention is deployed when different actions are planned on objects, in contrast to when the same objects are categorized. In this respect, we are particularly interested in the role every fixation plays in the unfolding dynamics of action control. We conducted an eye-tracking experiment in which participants were shown images of real-world objects. Subjects were either to assign the displayed objects to one of two classes (categorization task), to mimic lifting (lifting task), or to mimic opening the object (opening task). Results suggest that even on simplified, two dimensional displays the eyes reveal the participant's intentions in an anticipatory fashion. For the active tasks, already the second saccade after stimulus onset was directed towards the central region between the two locations where the thumb and the rest of the fingers would be placed. An analysis of saliency at fixation locations showed that fixations in active tasks have higher correspondence with salient features than fixations in the passive task. We suggest that attention flexibly coordinates visual selection for information retrieval and motor planning, working as a gateway between three components, linking the task (action), the object (target), and the effector (hand) in an effective way.

1. Introduction

Since the early works of Buswell (1935) and Yarbus (1967) top-down, task-related guidance has been shown to strongly influence the way people control their gaze. In Yarbus' study different patterns of scanning were observed, depending on the question asked to the subject regarding the presented picture. Such an influence is so dominant that, as soon as a specific task is given, low-level, bottom-up saliency is basically overridden and plays quite a minor role in explaining eye fixations (Einhäuser, Rutishauser and Koch, 2008; Foulsham and Underwood, 2009; Henderson et al., 2007). This is even more the case when an action has to be produced, following visual information collection (Tatler et al., 2011). In the past decades, vision for perceptual tasks has been extensively studied. Still, increasing evidence is suggesting that vision has evolved for the control of action, and hence vision (or perception altogether) should be considered in relation to action. A number of theoretical frameworks has been put forward to reconcile perception and action in a unified framework, considering possibly shared and jointly coded representations (Averbeck, Latham and Pouget, 2006; Gibson, 1979; Hoffmann, 2010; Hommel et al., 2001; Prinz, 1997).

In the context of the theory on the dual nature of vision (Goodale & Milner, 1992), distinct neural pathways subserving the different functional demands of object categorization and object manipulation were suggested. The dissociation between vision-for-action and vision-for-perception has often been investigated by contrasting grasping tasks with visual judgement tasks (Goodale, 2011). However, contrasting evidence has emerged (Franz et al., 2000) and evidence of strong interactions between the two systems has been found (Himmelbach & Karnath, 2005).

What this evidence suggests, still, is that visual cues are weighted and used differently depending on whether the task is a manual action or a perceptual judgement. Thus, when considering action, selective attention assumes a different role. Often thought of as a mechanism to cope with visual processing limitations, attention is evidently most necessary and effective in the context of selection for action (Allport, 1987). Attention can act as a gateway between a parallel, fast system and the slower distal motor system by selecting and prioritizing the information in the scene.

* Corresponding author.
E-mail addresses: belardinelli@informatik.uni-tuebingen.de (A. Belardinelli), oliver.herbort@psychologie.uni-wuerzburg.de (O. Herbort), butz@informatik.uni--tuebingen.de (M.V. Butz).

http://dx.doi.org/10.1016/j.visres.2014.11.003
0042-6989/

Preparation of an action upon an object defines an attentional landscape, (covertly) encoding locations relevant for the subsequent serial motor execution (Baldauf & Deubel, 2010). Indeed, influences of action intentions have been shown even in apparently perceptual tasks such as visual search (Bekkering & Neggers, 2002) or object recognition (Bub, Masson & Lin, 2013; Deubel, Schneider & Paprotta, 1998).

Moving from screen stimuli to real-world scenes and to tasks involving sequences of motor actions, it is even more striking how eye movements are precisely planned to provide information for the execution of the current action. This has been shown in different settings, from tea-making (Land, Mennie & Rusted, 1999) to sandwich-making (Hayhoe et al., 2003) to a wealth of other more or less complex motor tasks (Land & Tatler, 2009). Usually, very few fixations are made on non task-relevant objects and fixations anticipate the next object to manipulate before the hand starts moving towards it. Strategies like 'look-ahead' and 'just-in-time' fixations support the idea that vision is deeply intertwined with the information-related needs to plan and control manual action (Ballard, Mayhoe and Pelz, 1995; Hayhoe et al., 2003). In a seminal paper for eye-hand coordination, Johansson et al. (2001) recorded both eye- and hand movement data during a motor task involving grasping a bar, avoiding an obstacle, touching a goal position, and placing the bar back. Subjects almost exclusively fixated landmark positions on the bar, on the protruding hindrance, and on the target before making contact with them or avoiding them.

However, not only do the objects selected for fixation depend on the task, but also how individual objects are fixated. How differences between perceptual and motor tasks are reflected in eye-movements on single objects has been less investigated. van Doorn et al. (2009) contrasted a visual judgement (by manual estimate) on the length of a stimulus to grasping of the same stimulus using Müller–Lyer illusion. Their hypothesis was that hand aperture during the estimate would be affected by the illusion, but not when the task was to actually grasp the deceptive stimuli. The authors reasoned that, although a complete neural distinction has not been demonstrated, the dorsal and ventral streams might rely on different functional information requirements, which should be reflected in both the illusion effect (as was the case) and in gaze patterns. Relevant to the present study, more central fixations were found in the grasping task (suggesting the acquisition of egocentric information) whereas more fixations on the end points were found when manual estimations had to be made (allocentric information acquired by shifting the gaze between the extremes of the stimulus).

Gaze behavior in pure passive viewing and active grasping was investigated by Brouwer, Franz and Gegenfurtner (2009), who used simple geometric shapes, either to be looked at or to be grasped. First and second fixations were found closer to the center-of-gravity (COG) of the object (in accordance with Foulsham & Underwood, 2009; Nuthmann & Henderson, 2010) in the viewing condition, while the grasping condition was characterized by second fixations closer to the index finger location (or to the more difficult to grasp location). In the study by Desanghere and Marotta (2011), Efron blocks – graspable blocks of different dimensions but equal surface area – were used in a perceptual estimate condition vs. a grasping condition. This time manual estimates were also compared for real and for computer-generated stimuli. In the real stimuli condition, fixations were directed first to the index finger site and second in the direction of the COG, in both tasks, while with pictorial stimuli the pattern was reversed (which the authors attributed to the fact that in the second case the estimate had to be made first when the stimulus had disappeared, so the finger site had to be memorized). These studies used very simple stimuli, devoid of any semantic identity or familiar interaction. They looked into differential functional demands in the extraction of low-level motor parameters, such as the retrieval of finger position, and in the extraction of generic perceptual information, such as size. Most of our daily interaction with objects, however, relies on very specific information regarding both object identity and the associated set of executable actions. Thus, it can be expected that the gaze behavior is aimed at acquiring both object identity features and task-related information, needed to plan and control the motor execution.

In contrast to the reviewed experiments, we were interested in assessing at which point the eyes reveal the final goal of an object interaction. Thus, we designed an experiment with three blocks of trials. In the passive block, participants had to judge an object property, while in the active blocks participants had to either lift or interact with the displayed object pantomimically. On the one hand, we aimed at extending previous results, by enriching the motor task palette with a rather generic object-related task (open), which is somewhat more variable and complicated than grasping. On the other hand, we wanted the perceptual task to be a pure recognition task, hence not dependent on any object size estimation. Another novelty in our approach is the usage of familiar, real world objects. Due to the different nature of the objects, the same task must be adapted to the specific size, orientation and affordances (such as different openings) of the presented object. We expected the gaze behavior to reflect – specifically in the active conditions – the task-dependent and object-specific nature of each trial. We analyzed the proximity of fixations to Regions Of Interest (ROI) pertaining to lifting actions and opening actions, and expected that the position of mean fixations would be closer to the ROI of the corresponding task and that the distance would decrease with the temporal evolution of the scanpath. Moreover, we investigated the relationship of the fixation distribution to the distribution of the behaviorally relevant locations (in the lift and open condition) and to the distribution of visually salient features in every object. The results show that the participants indeed looked at the object in anticipation of the current task demands, which is reflected in properties of the eye fixation positions and reaction times. The role of saliency appears to be relevant to the extent that it enhances affording locations.

2. Experiments

We conducted a main eye-tracking experiment and a parallel experiment aimed at extracting Regions Of Interest (ROI) from every stimulus in every condition. The latter experiment was conducted to be able to deduce an objective measure of the contact point regions that would be chosen for an actual grasp or opening action. Both experiments are detailed in the following subsections. In both experiments participants gave informed consent and were compensated with study credits or money. Both experiments were carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

2.1. Eye-tracking experiment

2.1.1. Participants

Twenty-two participants (12 women, aged 19–41, $M = 23.1$) carried out the eye-tracking experiment in all 3 conditions (task). One female participant's data was discarded because of very bad quality (high composite RMS error during calibration). All subjects were right-handed with corrected to normal vision.

2.1.2. Stimulus material

Stimuli were chosen from the ALOI dataset (Geusebroek, Burghouts & Smeulders, 2005), containing pictures of 1000 daily-use objects in different light/view conditions. 14 objects (plus 2

Fig. 1. Stimulus material used in the experiment.

test objects for practice trials) were chosen such that all of them could be easily lifted with one hand and had an opening part. They are all portrayed in a frontal view against a black background. Six objects are displayed upright, six lie horizontally with the opening part on the right. Two objects present a handle on the right and the opening on top. All 14 stimuli are showed in Fig. 1. Each picture has a resolution of 768 × 576 pixel. In each condition they were presented at mid-height on the right side of the screen. The right-sided presentation was necessary since the eye-tracker remote cameras reside centrally below the screen and grasping to the left of the screen would let the hand occlude the cameras. Therefore, the high predictability of the stimulus position was hence compensated by the unpredictability of the grasping or opening affording points and by the different orientation of the stimuli. In this way, we wanted to avoid any bias in the first saccades, for example moving toward the superior part of the image in the opening condition.

2.1.3. Apparatus and procedure

Participants sat in front of the screen on which the object stimuli were presented. Their head was resting on a chin rest with forehead support, about 70 cm away from the monitor (1680 × 1050 pixel, subtending 45.3° × 28.3° of field of view). Stimulus pictures subtended 20.7°, with the center of the picture lying at 12.3° from the center of the monitor. Eye movements were recorded via a binocular remote eye-tracker (EyeFollower, LC Technologies) working at 120 Hz, with an accuracy <0.4° even under head movement. A keyboard was placed between the chin rest and the monitor to record reaction times and manual responses. The participants had to execute three different tasks, each of which was presented in a single block of trials. The task order was randomized across participants, as was the stimulus order within each block. For each task, every object was presented five times, resulting in 210 trials per participant. To familiarize subjects with the procedure, 30 training trials were conducted on 2 other objects before the main experiment.

In the *classify* task, participants were asked to look at the presented object and to decide whether it could contain liquid or not. The response was given by a left/right (yes/no) arrow key press. This served the purpose of both having participants looking at the objects each time and making a manual response as in the other conditions. In the *lift* condition, participants had to reach to the screen and to mimic lifting the presented object in front of the screen. Analogously, in the *open* condition, they reached to the screen and mimicked opening the object. They were instructed to use only the right hand. To lift objects, they were asked to always perform a grasp frontally, either with the thumb leftwards (vertical objects) or downwards (horizontal objects) or by the handle, where present. As to the opening, they were told to imagine that the objects were "glued" to the shelf so that they could open them with just one hand. They were asked to execute the movement as naturally as possible and to act on the object according to the perceived size.[1] In each trial, participants were asked to keep the spacebar pressed until they were ready to execute the proper response. Each trial proceeded as follows: (1) the task (classify/lift/open) is displayed as a reminder at the center of the screen for 1.5 s; (2) the fixation cross is presented for a random time between 1 and 2 s (if by that time the spacebar is not pressed, a beep prompts the participant to press the spacebar and the cross stays until .5 s after pressing); (3) the stimulus appears on the right side of the screen; (4) Eye-tracking data and reaction times are collected up to the release of the spacebar (pre-movement data) and during the execution of the motor response; (5) the hand returns to the spacebar and the next trial starts once the spacebar is pressed.

2.1.4. Data processing and analysis

Trials in which no fixations were available (in which case the eye-tracker lost the eye or eye samples were just outside the stimulus) or the reaction time was longer than 2.5 s or shorter than 100 ms were eliminated from the analysis, for an amount of 119 out of 4410 total trials.

Fixations were extracted for each trial via the dispersion algorithm (Salvucci & Goldberg, 2000) with a temporal threshold of

[1] The displayed images were all of the same size, so that objects were presented larger or smaller than they typically are in reality. However, this scaling was not excessively pronounced so that the action to perform was still plausibly and naturally performable.

100 ms and a spatial dispersion threshold of 1.5°. We first looked at data collected prior to spacebar release, since this time period is most informative about the information extraction and motor planning processes preceding movement initiation. Moreover, to also have a closer look at the scanpath evolution, quantitative evaluations were done also on the first 3 fixations (or up to the third fixation). This choice was motivated by the consideration that 3 fixations amount to about 1 s of stimulus presentation, sufficient to retrieve necessary visual information and start the movement (according to reaction times), while later fixations could be more arbitrary and dependent on the subjects' preference and interest for the object.

For the analyses detailed in the next section repeated-measures ANOVAs were used. Sphericity violations (Mauchly's test) were assessed and corrected using Greenhouse-Geisser estimates of sphericity. Bonferroni's corrections (α level of $p < .05$) were applied for the means of interest (post hoc comparisons).

2.1.5. Heatmaps

For qualitative evaluation and informative visualization, heatmaps were computed from fixation data. These were obtained by placing a Gaussian with $\sigma = 0.5°$, centered on each fixation and height proportional to the duration of the fixation, so that longer fixations would be weighted more in the heatmap surface. Each map was scaled between 0 (not fixated) and 1 (longest fixated) to make maps comparable.

Fig. 2 shows the heatmaps of pre-movement fixations for one of the up-right objects and one of the horizontal objects. Already before movement initiation, task-dependent differences in eye fixations are evident.

An evolution in time across the first 3 fixations/conditions for one object is presented in Fig. 3. While the first fixation is usually left of the (visual) COG (i.e. with undershoot) for all conditions, already by the second fixation it seems possible to infer the current intention of the participants.

2.2. Touch-screen experiment

2.2.1. Participants

Ten different participants (6 women, aged 18–41, $M = 25.2$) carried out the ROI extraction experiment. All of them were confirmed right-handed.

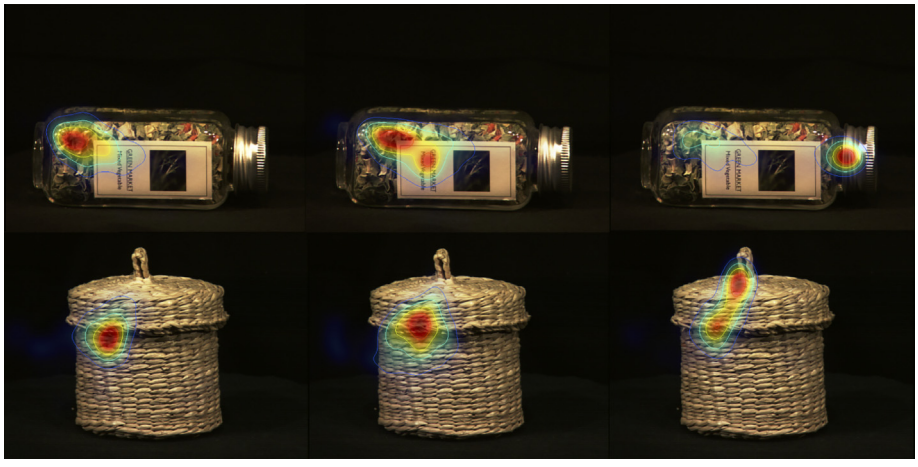

Fig. 2. Two heatmaps of pre-movement fixations superimposed on respective stimuli. From left to right: 'classify', 'lift' and 'open' condition.

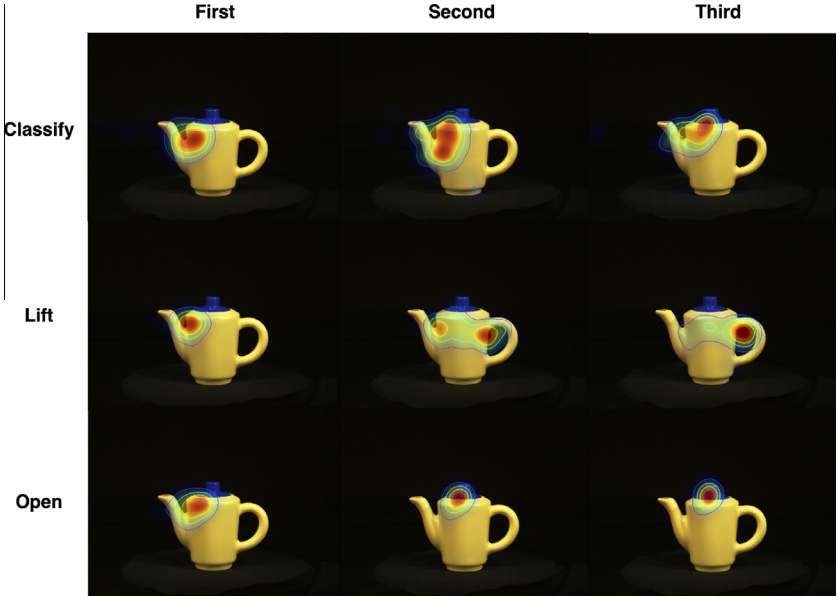

Fig. 3. Heatmaps of the first, second and third fixation (left to right).

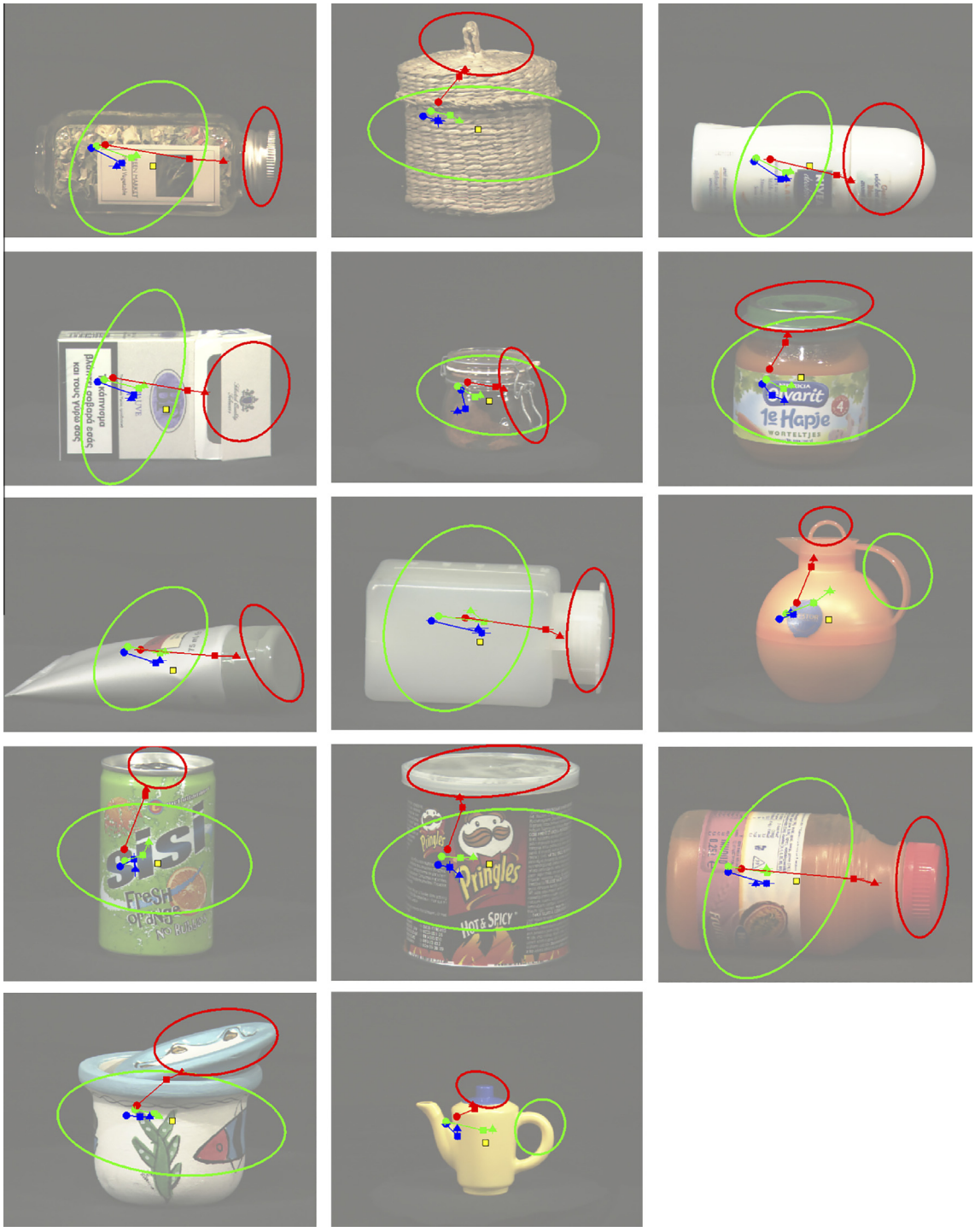

Fig. 4. Mean scanpaths for the first (circle), second (square) and third (triangle) fixation in the classify (blue), lift (green) and open (red) task. The green and red ellipses show the lifting and opening ROIs of every object. The yellow square indicates the COG position. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2.2. Stimulus material

The same stimuli were used as in the eye-tracking experiment. In this case too, stimuli were presented on the right side of the screen, for the sake of comparability with the first experiment.

2.2.3. Apparatus and procedure

In just 2 blocks (lifting and opening), participants were asked to place the tips of their fingers on the displayed object, as they would do to execute the lifting or the opening of the real object. These points were recorded via a touch screen. After each trial, the participant was shown the selected points and, if not satisfied, the trial could be repeated. Every object was presented 3 times per block, resulting in a total of 84 trials per participant.

2.2.4. Data processing and analysis

Regions of interest were extracted considering the distribution of the finger points in each condition. In the 'open' condition, points were compactly concentrated around the opening region,

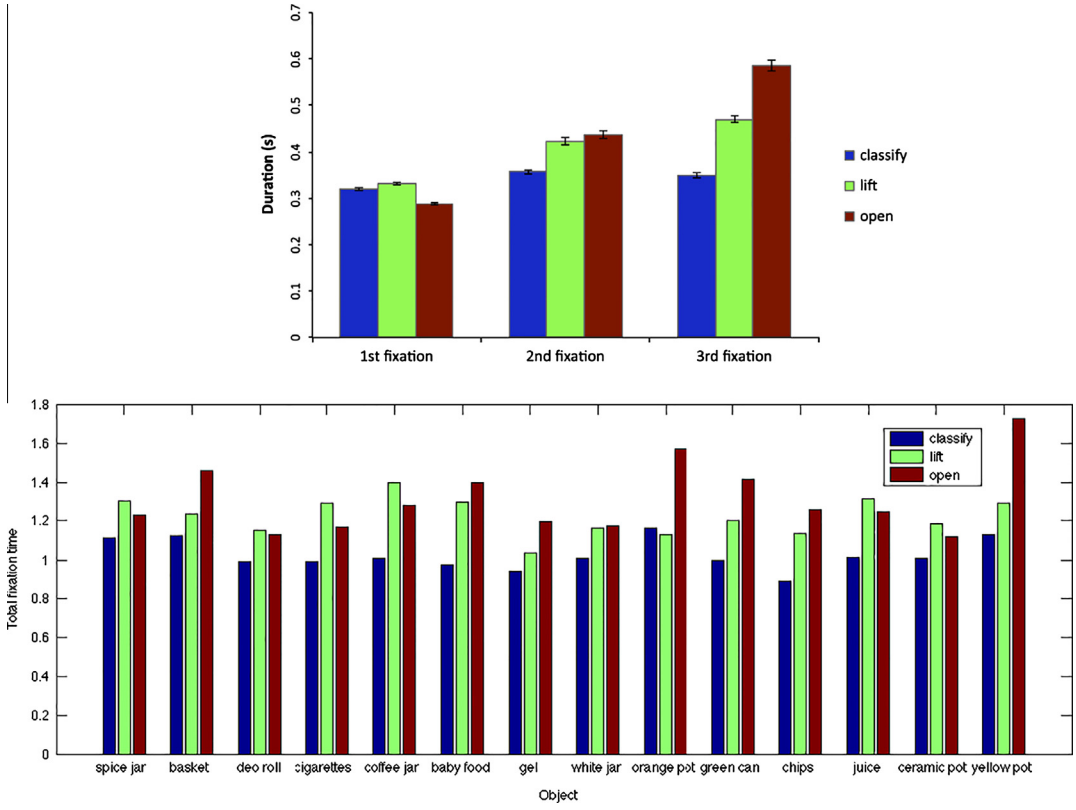

Fig. 5. Top, mean fixation duration for the first three fixations, according to task. Bottom, total fixation time for the first three fixations according to task and object.

hence mean and covariance matrix of the point coordinates sufficed to identify an ellipsoid containing the underlying region. In the case of 'lift', points were more evidently multi-modal, resulting in two major clusters, one typically smaller cluster for the thumb and one for the rest of the fingers. To include both clusters in the ROI, points were grouped via k-means, and an ellipsoid containing the whole lifting region was identified by sampling an equal number of points for the two clusters. In most objects the two ROIs were well-separated. In a few cases, they were slightly overlapping. Only in one case was there a major overlap, which, nevertheless, did not hamper our analyses. The 'lift' and 'open' ROIs for every object (with border 1.5 standard deviation along each axis) are displayed in Fig. 4.

Moreover, the COG was estimated for every object. This was obtained segmenting the objects by thresholding the gray-level picture of every stimulus and refining its outline by means of morphological operations. The COG was computed as the centroid of the points belonging to the object mask. This is also shown in Fig. 4.

3. Results

3.1. Fixation and reaction times

We considered (up to) the first three fixations for each trial (for a total of 12,276). As a measure of the amount of information needed before making a response, we looked at what portion of fixations was made prior to movement for each task. The first fixation occurred before hand movement in 93.7% of cases,[2] the second fixation in 58.4%, and the third in 27.1%. Of all examined fixations, 7473 occurred before starting the movement. While for the first fixation these fixations were equally distributed across tasks (4020 fixations, 32.9% classify, 33.2% lift, 33.9% open), in the second the proportion is in favor of lifting and opening (2404 fixations, 23.2% classify, 33.2% lift, 43.6% open). By the third fixation, it was mostly for the 'active' tasks that motion had not yet been initiated (1049 fixations, 16.7% classify, 37.9% lift, 45.4% open).

Independently of when fixations occurred with respect to movement onset, we considered whether the fixation duration of the first three fixations was affected by the task. Mean fixation duration provide information on the sequential processing effort required by features and semantic properties of the stimulus, within each task (Rayner, 2009). Mean durations and Standard Error of the Mean (SEM) according to task and number of fixation are shown in Fig. 5, top graph. A repeated measures ANOVA with factors fixation, task and object showed a main effect of task ($F(2,40) = 6.68, p = .003$), fixation ($F(1.28, 25.70) = 26.54, p < .001$) and object ($F(6.22, 124.42) = 5.83, p < .001$). Yet, differences across tasks were significant only between classify and the two active conditions (classify vs. lift, $p = .043$, classify vs. open, $p = .019$). Fixation duration increased significantly from the first to the third fixations (1 vs. 2, $p = .001$, 1 vs. 3, $p < .001$, 2 vs. 3, $p < .001$). Interaction effects showed that fixation durations increased from first to third and more so from classify to open ($F(2.44, 48.71) = 15.57, p < .001$), and marginally depending on the object ($F(26, 520) = 1.54, p = .045$). There was also an interaction of object and task ($F(26, 520) = 3.21, p < .001$), and an interaction of all three factors ($F(52, 1040) = 1.48, p = .016$).

Total fixation time (across the first three fixation) according to task and object is shown in Fig. 5, bottom graph.

Considering reaction times, we were expecting them to increase from the passive to the most refined active condition. Mean reaction times in releasing the spacebar indeed increase from 'classify' to 'lift' to the 'open' condition (classify: 0.583 s, lift: 0.715 s, open:

[2] In few cases, subjects were able to start executing the response before landing with the eyes on the picture.

0.779 s), but the difference again is significant only between 'passive' and 'active' conditions (classify-lift $p = .001$, classify-open $p < .001$). A repeated measures ANOVA on the reaction times with task and object as within-subject factors showed a main effect of task, $F(2, 40) = 20.81, p < .001$, a main effect of object, $F(6.39, 127.84) = 3.84, p = .001$, and an interaction effect, $F(26, 520) = 2.27, p < .001$.

3.2. Fixation locations and distribution relative to the ROIs

We first considered how the peaks of the heat maps of the first three fixations was located w.r.t. the COG, in order to assess whether these showed some undershoot effects or finger preference across tasks.

The distance between the fixation map maximum and the COG along the x and the y axes was analysed in a one-way ANOVA with task as independent variable. The task has indeed a main effect ($F(1.14, 14.76) = 16.34, p < .001$), with the 'open' peak to the right of the COG ($M = 75$ pixel, SD $= 30.1$) while 'lift' and 'classify' peaks concentrated to the left of the COG ($M = -73.6, SD = 12.0$ and $M = -91.2, SD = 16.8$, respectively). The difference was significant just between 'open' and the other two conditions ($p = .002$ and $p = .005$, lift and classify, respectively). Since the difference for the open condition is mostly driven by the horizontally lying objects, we repeated the same ANOVA for the 6 vertical (excluding the two with handle) and the 6 horizontal objects. The effects basically did not change. Again the open case differed from the other two conditions, $F_{horizontal}(2, 10) = 36.02, p < .001; F_{vertical}(2, 10) = 15.19, p = .001$. When considering the distance to the COG along the vertical dimension, task had again a main effect, $F(1.07, 13.09) = 6.62, p = .021$, but no significant difference among conditions, with 'open' still farther above the COG than the other two conditions ($M_c = -22.6, SD_c = 9.5; M_l = -23.7, SD_l = 9.0$; $M_o = -92.4, SD_o = 22.4$). A further analysis with only horizontal objects showed no effect, while for the vertical objects again the task had an effect, $F_{vertical}(2, 10) = 24.18, p < .001$, showing that this is just due to the vertically presented objects.

To gain a more specific insight regarding to what extent the fixation map matches with and can predict the region on which the motor action is performed, we compared the ROIs extracted for the two 'active' conditions with the peak of the corresponding heatmaps considering the distribution of the first three fixations (see Fig. 6). The peak of the fixation map (where the map has value 1) consistently falls within the corresponding ROI. The mean distance between the peak and the center of the ROI for the 'lift' condition was 89.8 ± 65.8 pixel, while for the 'open' condition was 53.4 ± 22.4. In both conditions the distance between the peak and the center of the corresponding ROI was always smaller than that to the center of the other ROI (one-tailed t-test, $p < .001$).

We extended this analysis considering the first, second and third fixation mean position for every object and their respective distances to the two ROIs. The hypothesis was that the first fixation's distance would not differ across tasks, while subsequent fixations should move progressively closer to the ROI center, and hence to the affordance corresponding to the task. Fig. 7 shows the distance of the fixations to the lift-ROI and open-ROI for the first 3 fixations. The distances were subjected to a repeated measures ANOVA with within-subject factors task, fixation number, and ROI. Overall, fixations were closer to the lift-ROI, $F(1, 13) = 17.63, p = .001$, fixations in the open condition were closer to both ROIs than those in other tasks, $F(2, 26) = 76.47$,

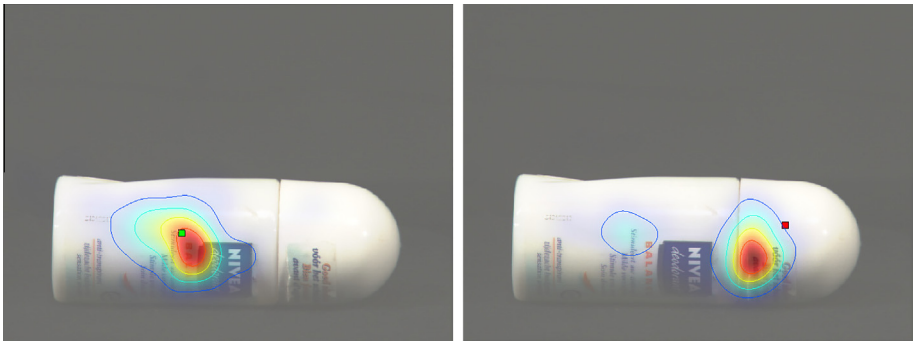

Fig. 6. Left: heatmap of the first 3 fixations in the 'lift' condition (green dot: center of corresponding ROI). Right: heatmap of the first 3 fixations in the 'open' condition (red dot: center of corresponding ROI). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

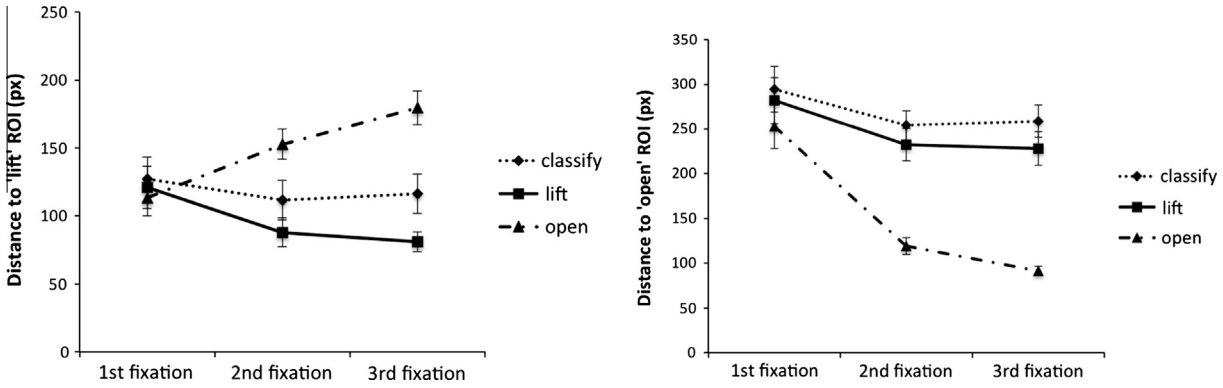

Fig. 7. Left: Mean distances of the first, second and third fixation to the center of the 'lift' ROI. Right: Mean distances of the first, second and third fixation to the center of the 'open' ROI.

$p < .001$, and later fixations were closer to both ROIs than first fixations, $F(1.11, 14.44) = 57.26, p < .001$. Distances to ROIs were shortest for the corresponding task, $F(1.05, 13.72) = 65.64$, $p < .001$. Distances to the open-ROI tended to decrease with each fixation, whereas distances to lift ROI were minimal by the second fixation, $F(1.07, 13.92) = 14.19, p = .002$. Distances in every task decreased from the first to the third fixation and more so for the active conditions, while for classify stayed relatively constant, $F(2.18, 28.33) = 12.03, p < .001$. Finally, distances were closer to the ROI corresponding to the task, more so by the third fixation, $F(1.33, 17.30) = 51.57, p < .001$. A separate analysis for the 6 horizontal and the 6 vertical objects, showed similar effects but the interaction between fixation and task was not significant for the former ones.

Multiple comparisons showed that the measured distances were significantly different between first and second and first and third fixation ($p < .001$) but not between second and third fixation. Significant differences were also present between each task ($p < .001$), with fixations in the classify condition being generally the furthest from the ROIs.

To better interpret these results a $2 \times 2 \times 2$ ANOVA was conducted (task levels: lift and open; fixation levels: first and third; ROI levels: lift and open). This confirmed the effects above: shorter distances for the more easily reachable lift-ROI, $F(1, 13) = 14.53, p = .002$, shorter distances for the open task, $F(1, 13) = 55.93$, $p < .001$, and, of course, shorter distances for the third fixation, $F(1, 13) = 67.08, p < .001$. Interaction effects could be observed between ROI and task, $F(1, 13) = 70.70, p < .001$, ROI and fixation, $F(1, 13) = 19.70, p = .001$, and between all three factors, $F(1, 13) = 69.10, p < .001$. How the mean scan paths (first, second, third fixation) were positioned on each object and relative to each ROI is detailed in Fig. 4.

3.3. Saliency analysis

Although saliency is believed to influence vision dominantly in a bottom-up fashion, we asked the question if saliency nonetheless may co-determine or correlate with fixation positions. Our reasoning was that if salience of some specific feature was consistently underlying fixations in some condition, as e.g. in Rothkopf and Ballard (2009), this feature can also be used to predict fixations in that condition and to possibly characterize affordances in terms of low-level features. To this end we made use of the simplified version of the saliency model by Itti, Koch and Niebur (1998), in the Matlab implementation by Harel. This model relies on three basic feature channels (color, intensity and orientation), which are separately processed in a center-surround fashion at multiple scales so that locations strongly contrasting to the surrounding can emerge. The feature maps are finally combined in a single saliency map. Saliency maps were produced for each of the stimuli. To measure the amount of saliency underlying each subject's first three fixations according to task and object we used the Normalized Scanpath Saliency (NSS, cf. Peters et al., 2005). To this aim, saliency maps were normalized to have zero mean and standard deviation of one. Saliency was then collected at fixation locations and averaged over the total number of fixations (3). This means that a NSS value of zero corresponds to a chance level of saliency underlying fixations, while a positive value of 1, for example, means that fixations collect on average saliency at an amount of one standard deviation above the mean salience present in the stimulus. Saliency is not absolute but always relative to the contrast between one point and its surrounding, hence it must be noted that the saliency maps are influenced by incidental light reflections or contrasted oriented edges. However, still these cues might be exploited by the visual system in a correlational fashion to efficiently extract opening regions or affording grasping points such as handles.

Table 1
Comparison with the saliency map: mean NSS values and similarity score for fixations and touch points data, in the lifting and opening conditions.

Metric	Eye Lift	Hand Lift	Eye Open	Hand Open
NSS	1.47 ± 0.80	0.63 ± 0.88	1.60 ± 0.76	1.07 ± 0.88
SimScore	0.45 ± 0.07	0.39 ± 0.08	0.40 ± 0.07	0.32 ± 0.08

NSS values across objects and tasks are shown in Fig. 8. A two-way ANOVA confirmed again a main effect of task, $F(2, 40) = 40.99, p < .001$, and object, $F(4.03, 80.61) = 69.12, p < .001$, and an interaction effect, $F(26, 520) = 18.76, p < .001$. In general, the classify condition correlates with salience the least, while the open condition the most. Pairwise comparisons show significant differences among all conditions (classify-lift and classify-open: $p < .001$; lift-open: $p = .005$). We looked also into saliency maps produced considering each feature separately. The major amount of saliency, in any task condition, was collected by fixations in the case of orientation saliency. Contrasted oriented edges, not surprisingly, characterize salient opening parts, handles or written text/pictures in some objects.

This relationship between fixation position and saliency, even if not causal, is still a stronger prerogative of the visual rather than of the manual behavior. Hand data, indeed, although predicted by eye movements on a larger scale, were less coherent with salient loca-

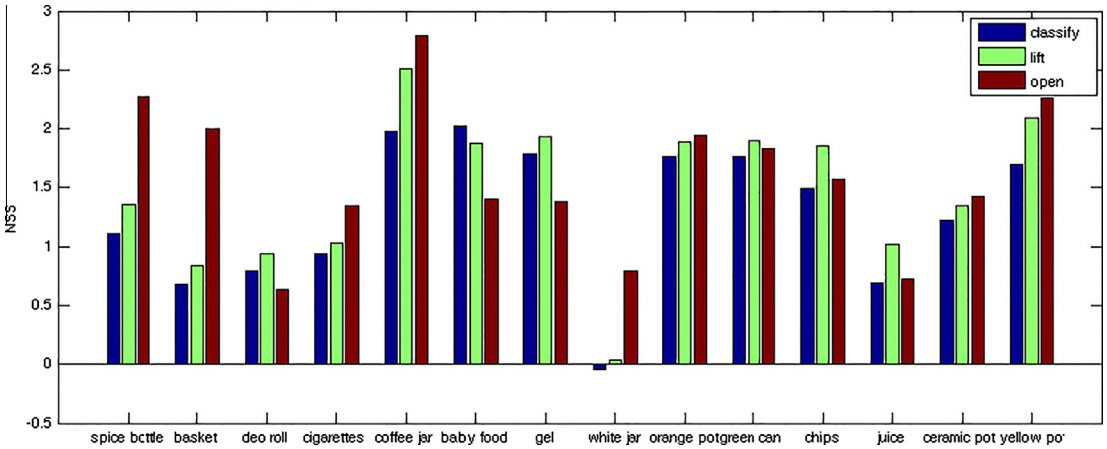

Fig. 8. Mean normalized scanpath saliency across task and object.

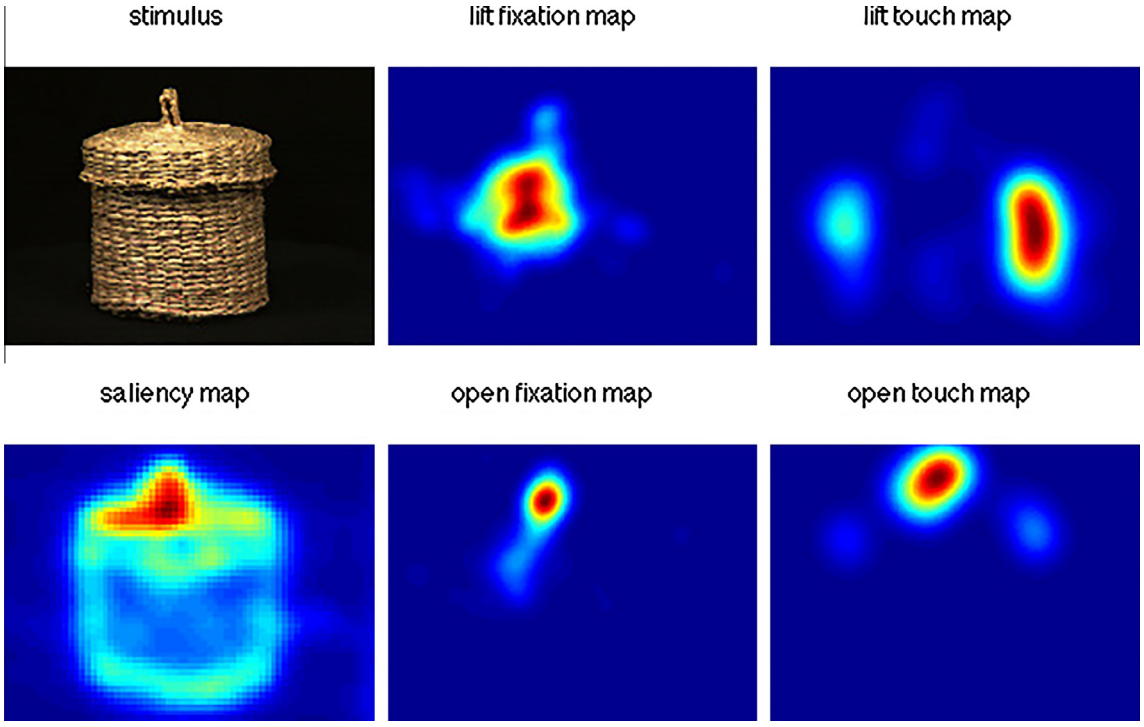

Fig. 9. Touch, fixation, and saliency maps for the stimulus 'basket'.

tions. Computing the NSS scores on the finger touch points and comparing the mean values for each subject and object for the active conditions with those from the fixation data, in both task conditions the NSSs obtained from the touch points resulted inferior to the NSS scores from eye fixations (one-tailed t-test: lift $t(432) = -9.95, p < .001$, open, $t(432) = -6.67, p < .001$). Mean values and standard deviations across tasks and modality are reported in Table 1.

Still, it can be the case that finger tip points for the most part did not land on the object but very close to it, significantly reducing the amount of collected saliency. To rule out this possibility, touch heatmaps were created in the same way as for the fixation heatmaps. In this way, the finger distribution creates a slightly spread and smoother distribution on the contact points. An example is presented in Fig. 9, along with the fixation and the saliency map for the two active tasks.[3]

As can be seen, the peak in the fixation map in both conditions is roughly in the middle of the thumb peak and the other fingers peak. We have now three maps with values ranging between 0 and 1. To assess which of the behavioral maps had greater overlap with the saliency map, a similarity score for every object and task was computed. We used the similarity score proposed by Judd, Durand and Torralba (2012) to compare how similar the fixation and touch heatmaps are relatively to the saliency map. This metric considers the intersection of histograms of two distributions A and B and sums across bins the minimum value of the two distributions:

$$S(A,B) = \sum_{i,j} min(A(i,j), B(i,j))$$

where the two distributions sum to 1. The score ranges between 0 (completely different distributions) and 1 (coincident distributions). Normalizing every lifting and opening map to obtain a probability distribution (with a bin for every pixel), it can be shown that the eye fixation heatmaps in both active tasks are more similar to the saliency maps than the touch heatmaps (one-tailed t-test, lift $t(26) = -2.09, p = .023$, open $t(26) = -2.77, p = .005$). Mean similarity scores with standard deviation for every condition are also shown in Table 1. Even if the fixation map is generally more similar to the saliency map than the touch map, the strong goal-orientedness of the gaze behavior is demonstrated by a maximal similarity score of just 0.56 (in the lifting condition of the teapot in Fig. 3).

In analogy to the comparison with the ROIs above, we considered the similarity score as a distance measure and analysed if the evolution of the eye heatmaps across the first 3 fixations on every object and every task was more overlapping with the touchmaps in the active tasks or with the saliency map. Since the heatmaps integrate the duration of each fixation, they represent the actual visual behavior more faithfully than averaged fixation positions. We ran a 3-way ANOVA with factors fixation (1st, 2nd, 3rd) × task (classify, lift, open) × map (lift touchmap, open touchmap, saliency map) on the similarity scores. This showed a main effect of fixation, $F(2,26) = 60.71, p < .001$, task, $F(2,26) = 8.32, p = .002$, and map, $F(2,26) = 73.17, p < .001$, and also interaction effects for map x fixation, $F(2.37, 30.81) = 11.77, p < .001$, map × task, $F(2.27, 29.56) = 68.51, p < .001$, fixation × task, $F(4,52) = 5.01, p = .002$, and map × fixation × task, $F(3.77, 49.02) = 60.71, p < .001$. Pairwise comparisons showed that the mean similarity of the eye heatmaps with the saliency map ($M = 0.361$) scored significantly higher than similarity with the touch maps (lift-saliency and open-saliency, $p < .001$), further supporting the complementarity of eye- and touch- heatmaps, but also indicating that the gaze behavior is influenced by bottom-up salient features in addition to the top-down task influences. Moreover, in general the overlap with any map increased across fixations (first-second and first-third, $p < .001$; second-third, $p = .002$). Differences across tasks were less significant (classify-lift, $p = .042$, classify-open, $p = .008$), or not at all (lift-open, $p = .449$). The mean similarity across task and map is shown in Table 2. These data are averaged across the three fixations.

[3] The maps for all objects can be seen at http://www.wsi.uni-tuebingen.de/lehrstuehle/cognitive-modeling/staff/staff/anna-belardinelli/object-interaction.html.

Table 2
Mean similarity values of the eye heatmaps across task (columns) to the other maps (touch and saliency maps, rows). In this case the heatmaps were computed separately for the first three fixations (averaged values are shown).

Map	Classify	Lift	Open
Lift touchmap	0.165 ± 0.022	0.200 ± 0.019	0.122 ± 0.017
Open touchmap	0.076 ± 0.015	0.084 ± 0.013	0.275 ± 0.020
Saliency map	0.368 ± 0.014	0.392 ± 0.015	0.323 ± 0.013

4. General discussion

The present study was aimed at assessing different eye movement strategies employed when recognizing a functional property of an object in contrast to tasks in which actual interactions had to be performed on the object. Even though the interaction with real objects in our daily life heavily relies on depth perception, Westwood et al. (2002) showed how subjects can effectively program actions to 2D pictures, suggesting that the dorsal stream does not critically rely on binocular information for prehension movements (see also Kwok & Braddick, 2003). This turned out to be the case in this study, where indeed familiar objects were used and the scanning patterns were similar to those described for real objects. Both task and object-specific affordance points were expected to strongly influence the distribution of eye fixations. Indeed, significant differences in the scanpath behavior across the 3 conditions were found, suggesting for each object and in each task the construction of a specific attentional landscape around the behaviorally most relevant points, with targeting of those locations starting already before movement initiation.

When considering the weighted distribution of the first three fixations in the classification task, these concentrated in the direction of the COG of the object, but lingering left of it, since landing just on the stimulus was probably enough to confirm the peripheral impression, formed before saccading, and to make the required decision. When grasping an object to lift it, fixations landed on the object and moved toward the center of the lift ROI. It seems reasonable that instead of fixating both contact points in an alternate fashion, fixating near the center of the object allows both contact points to be in the fovea and para-fovea, as suggested by Desanghere and Marotta (2011). The most distinctive case was the opening task, where already the second fixation clearly targets the opening region, which requires the most processing for the planning of the finer motor operations (usually performed with a precision grip). Comparison with the center of the respective ROI in each condition showed that the peak of the fixation map (which accounted also for fixation duration) was closely located in the middle of the distribution of the thumb points and of the rest of the fingers.

Even if the overall distribution of fixations is already indicative, the different patterns in the unfolding of the scanpath are best appreciable when looking at the temporal evolution of the first three fixations. The distributions of the first fixation is hardly distinguishable across tasks, but already by the second fixation (at which point the reaching movement often had not been initiated, yet) the task 'signature' became evident.

These results confirm the general predictive nature of eye movements, especially in an active task, where successful, refined motor action relies on task-specific and timely information. Moreover, observed results confirm the tight interaction of attention and action. Objects do not only capture attention by offering a potential for action and priming related motor programs (Handy et al., 2003), but, as discussed by Humphreys et al. (2010), current task-specific interaction goals determine which information about the object get focused over time. For example, Humphreys and Riddoch (2001) showed that specifying a target for a visual search task by its function instead of by its name helped a neglect patient find it more efficiently in the contralesional hemifield. This suggests that even if vision-for-action and vision-for-perception may rely on functionally separate neural pathways, attention is a general-purpose mechanism that can flexibly implement selection-for-perception and selection-for-spatial-motor-action (see also Hesse, Schenk & Deubel, 2012). Schneider and Deubel (2002) have shown how these two selective modalities are coupled and mediated by a common attentional mechanism. In our case, it is not the action nor the target that needs to be selected among other competing items, but in a similar way, the task-relevant and object-specific affording points. Hence, the motor preparation needs to be bound to the perceptual spatial representation, a coordination step not needed in the judgement task. That is, the attentive integration mechanism flexibly weighs perceptual factors and high-level top-down guidance according to the information-seeking and the motor programming processes currently in focus.

In Land, Mennie and Rusted (1999), fixations in complex sequential tasks were classified in 4 categories: 'locating' the target, 'directing' the hand approaching the target, 'guiding' the action between two interacting locations or 'checking' the result/state of an action. In our constrained set-up, the considered fixations should be probably best ascribed to the first two classes. It seems plausible that the general flow of processing is first concerned with locating and recognizing the object of interest (first fixation, in the direction of the COG). This is a first step common to all three tasks and indeed, first fixations are almost indistinguishable in position and duration across tasks. Next, the gaze moves towards the most relevant points – either for decision making in the case of the classification task, or for the purpose of initiating anticipatory behavior control (Butz et al., 2007; Hoffmann, 2003) towards functional points (for lifting/opening) with proper interaction routines. In the former case, mostly the ventral system would be involved, pooling resources for recognition and decision-making. In the latter, 'active' conditions, also the dorsal pathway and premotor cortical regions would be substantially involved. After object localization and recognition, object-relative behavior needs to be planned, which involves reference-frame transformations of position, size, and shape and planning of reaching and grasping movements with properly aligned hand shapes (Cisek, 2007; Herbort and Butz, 2011; Jeannerod et al., 1995). In this situation, the gaze, particularly by the third fixation, remains anchored at a location in between the touching locations (mostly in the center of the object for lifting and on the lid for opening). The consequentially more elaborate motion planning is also confirmed by significantly longer reaction times and fixation times when an active motor task, different for every object, has to be planned anew.

Beyond that, however, our data indicate that tracking eye movements, combined to some extent with analysis of low-level saliency features, may be exploited in even more subtle ways, inferring the exact intention of how a user may interact with an object. Such discriminability of eye scanpaths according to the intended interaction goal may substantially help in devising machine learning algorithms and object-based attention systems (e.g., Wischnewski et al., 2010) to quickly infer the intention of impaired patients and possibly inform assistive interfaces to control prosthetic devices in accordance with the inferred intention. The reliability with which the fixation mode consistently fell within the specific ROI supports these considerations. Thus, we believe that our results also have important implications on how the brain gathers information for behavior control.

In conclusion, as for more complex behavior, such as those investigated by Land, Mennie and Rusted (1999) and Hayhoe et al. (2003), even for single actions to be performed within the same object, the eyes extract visual information in a goal-oriented, anticipatory fashion, incrementally revealing the interaction inten-

tions. As a last remark, similar behavior has been observed when subjects are extemporaneously constructing a verbal sentence describing the observed scene Griffin and Bock (2000), possibly suggesting that ultimately behaviorally grounded processes are at work also in this case.

Acknowledgments

The authors would like to thank Garrison Cottrell and an anonymous reviewer for valuable comments on previous versions of this manuscript.

A. B. is supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63).

References

Allport, D. (1987). Selection for action: Some behavioural and neurophysiological considerations of attention and action. In H. Heuer & A. F. Sanders (Eds.), *Perspectives on perception and action* (pp. 395–419). Hillsdale, NJ: Lawrence Erlbaum Associates.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. Nature reviews. *Neuroscience, 7*, 358–366.

Baldauf, D., & Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision Research, 50*, 999–1013.

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

Bekkering, H., & Neggers, S. F. (2002). Visual search is modulated by action intentions. *Psychological Science, 13*, 370–374.

Brouwer, A. M., Franz, V. H., & Gegenfurtner, K. R. (2009). Differences in fixations between grasping and viewing objects. *Journal of Vision*, 9.

Bub, D. N., Masson, M. E. J., & Lin, T. (2013). Features of planned hand actions influence identification of graspable objects. *Psychological Science*.

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

Butz, M. V., Sigaud, O., Pezzulo, G., & Baldassarre, G. (Eds.). (2007). *Anticipatory behavior in adaptive learning systems: From brains to individual and social behavior (LNAI 4520)*. Springer-Verlag.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*, 1585–1599.

Desanghere, L., & Marotta, J. (2011). "Graspability" of objects affects gaze patterns during perception and action tasks. *Experimental Brain Research*, 1–11.

Deubel, H., Schneider, W. X., & Paprotta, I. (1998). Selective dorsal and ventral processing: Evidence for a common attentional mechanism in reaching and perception. *Visual Cognition, 5*, 81–107.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8.

Foulsham, T., & Underwood, G. (2009). Does conspicuity enhance distraction? Saliency and eye landing position when searching for objects. *Quarterly Journal of Experimental Psychology, 62*, 1088–1098.

Franz, V. H., Gegenfurtner, K. R., Bülthoff, H. H., & Fahle, M. (2000). Grasping visual illusions: No evidence for a dissociation between perception and action. *Psychological Sciences, 11*, 20–25.

Geusebroek, J. M., Burghouts, G. J., & Smeulders, A. W. M. (2005). The amsterdam library of object images. *International Journal of Computer Vision, 61*, 103–112.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Goodale, M. A. (2011). Transforming vision into action. *Vision Research, 51*, 1567–1587.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*, 20–25.

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

Handy, T. C., Grafton, S. T., Shroff, N. M., Ketay, S., & Gazzaniga, M. S. (2003). Graspable objects grab attention when the potential for action is recognized. *Nature Neurosciences, 6*, 421–427.

Harel, J. A saliency implementation in MATLAB <http://www.klab.caltech.edu/harel/share/gbvs.php>.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*, 49–63.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 537–562). Elsevier.

Herbort, O., & Butz, M. V. (2011). Habitual and goal-directed factors in (everyday) object handling. *Experimental Brain Research, 213*, 371–382.

Hesse, C., Schenk, T., & Deubel, H. (2012). Attention is needed for action control: Further evidence from grasping. *Vision Research, 71*, 37–43.

Himmelbach, M., & Karnath, H. O. (2005). Dorsal and ventral stream interaction: Contributions from optic ataxia. *Journal of Cognitive Neuroscience*, 632–640.

Hoffmann, J. (2003). Anticipatory behavioral control. In M. V. Butz, O. Sigaud, & P. Gérard (Eds.), *Anticipatory behavior in adaptive learning systems: Foundations, theories, and systems* (pp. 44–65). Springer-Verlag.

Hoffmann, J. (2010). Speculations on the origin of STM. *Psychologica Belgica, 50*(3,4), 175–191.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849–878.

Humphreys, G. W., & Riddoch, J. M. (2001). Detection by action: Neuropsychological evidence for action-defined templates in search. *Nature Neuroscience, 4*, 84–88.

Humphreys, G. W., Yoon, E. Y. Y., Kumar, S., Lestou, V., Kitadono, K., Roberts, K. L., et al. (2010). The interaction of attention and action: From seeing action to acting on perception. *British Journal of Psychology, 101*, 185–206.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: The cortical mechanisms of visuomotor transformation. *Trends in Neurosciences, 18*, 314–320.

Johansson, R. S., Westling, G., Backstrom, A., & Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. *Journal of Neurosciences, 21*, 6917–6932.

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. Technical Report Tech. Rep. No. MIT- CSAIL-TR-2012-001. Cambridge, MA, USA: MIT Computer Science and Artificial Intelligence Laboratory.

Kwok, R., & Braddick, O. (2003). When does the Titchener Circles illusion exert an effect on grasping? Two- and three-dimensional targets. *Neuropsychologia, 41*, 932–940.

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.

Land, M., & Tatler, B. (2009). *Looking and acting vision and eye movements in natural behaviour*. Oxford University Press.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, 10.

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*, 2397–2416.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*, 129–154.

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

Rothkopf, C. A., & Ballard, D. H. (2009). Image statistics at the point of gaze during human navigation. *Visual Neuroscience, 26*, 81–92.

Salvucci, D., & Goldberg, J. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research and applications* (pp. 71–78).

Schneider, W., & Deubel, H. (2002). Selection-for-perception and selection-for-spatial-motor-action are coupled by visual attention: A review of recent findings and new evidence from stimulus-driven saccade control. In W. Prinz & B. Hommel (Eds.), *Attention and performance XIX: Common mechanisms in perception and action* (pp. 609–627). Oxford: Oxford University Press.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11.

van Doorn, H., van der Kamp, J., de Wit, M., & Savelsbergh, G. J. P. (2009). Another look at the Müller–Lyer illusion: Different gaze patterns in vision for action and perception. *Neuropsychologia, 47*, 804–812.

Westwood, D. A., Danckert, J., Servos, P., & Goodale, M. (2002). Grasping two-dimensional images and three-dimensional objects in visual-form agnosia. *Experimental Brain Research, 144*, 262–267.

Wischnewski, M., Belardinelli, A., Schneider, W., & Steil, J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation, 2*, 326–343.

Yarbus, A. L. (1967). *Eye movements and vision* (1st ed.). Plenum Press.