Gaze strategies in object identification and manipulation

Anna Belardinelli (belardinelli@informatik.uni-tuebingen.de)
Department of Computer Science, University of Tübingen,

Martin V. Butz (martin.butz@uni-tuebingen.de)
Department of Computer Science, Department of Psychology, University of Tübingen,
Sand 14, Tübingen, 72076 Germany

Abstract

Task influence has long been known to play a major role in the way our eyes scan a scene. Interestingly, how the task modulates attention when interacting with objects has been less investigated. Only few studies have contrasted the distribution of eye fixations during viewing and grasping objects. How is attention differently deployed when different actions have to be planned on objects in contrast to a purely perceptual viewing condition? To investigate these issues, we conducted an eye-tracking experiment showing participants 2D images of real-world objects. In blocks of trials, participants were asked either to assign the displayed objects to one of two classes (classification task), to mimic lifting the object (lifting task), or to mimic opening the object (opening task). Mean fixation locations and attention heatmaps show different modes in gaze distribution around task-relevant locations, in accordance with previous literature. Reaction times, measured by button release in the manual response, suggest that the more demanding the task in terms of motor planning the longer the latency in movement initiation. Results show that even on simplified, two dimensional displays the eyes reveal the current intentions of the participants. Moreover, the results suggest elaborate cognitive processes at work and confirm anticipatory behavioral control. We conclude with suggesting that the strongly predictive information contained in eye movements data may be used for advanced, highly intuitive, user-friendly brain-computer interfaces.

Keywords: Eye-tracking, object interaction, fixation distribution, eye-hand coordination, movement preparation

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 move their gaze upon pictures. In the second study, depending on the question asked, different patterns of scanning were observed. Such an influence is so critical that, as soon as a specific task is given, low-level, bottom-up visual saliency is basically overridden and plays quite a minor role in explaining eye fixations w.r.t. higher-level cognitive factors (Henderson, Brockmole, Castelhano, & Mack, 2007; Einhäuser, Rutishauser, & Koch, 2008). Similarly, moving from pictures to real-world scenes and to tasks involving motor actions, it is even more striking how eye movements are precisely planned to provide information for the execution of the current piece of action. This has been shown in different settings, from tea-making (Land, Mennie, & Rusted, 1999) to sandwich-making (Hayhoe, Shrivastava, Mruczek, & Pelz, 2003) to a wealth of other more or less complex motor tasks (Land & Tatler, 2009). In this case, anyway, the nature of attention deployment is quite different. The purpose of vision is here indeed less to get sense of the scene and more to direct effectors and coordinate a much slower and more complex behaviour than scanning. Strategies like 'look-ahead' and 'just-in-time' fixations (Hayhoe et al., 2003; Ballard, Hayhoe, & Pelz, 1995) support the idea that vision is deeply intertwined with the needs of motion planning and supervising.

Further, in the context of the theory on the duplex nature of vision (Goodale & Milner, 1992), distinct neural pathways subserving the different functional demands of object categorization and object manipulation were suggested. This dissociation between vision-for-action and vision-for-perception has often been investigated by means of grasping tasks contrasted to perceptual judgement tasks, with visual illusions or in covert attention settings (Goodale, 2011), but contrasting evidence has emerged and it seems reasonable to assume a strict interaction between the two systems.

How the differences between perceptual and motor task are reflected in eye-movements has been less investigated. In a seminal paper for eye-hand coordination, Johansson, Westling, Backstrom, and Flanagan (2001) recorded both eye- and hand movements 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 or in the experimental set-up, before making contact to them. The preparation of an action upon an object defines an attentional landscape (Baldauf & Deubel, 2010), (covertly) encoding in parallel locations relevant for the subsequent serial motor execution.

This evidence suggests that visual cues are sought and weighted differently depending if the task is a skilled movement or a perceptual judgement. Gaze behaviour in viewing and grasping was investigated by (Brouwer, Franz, & Gegenfurtner, 2009) and (Desanghere & Marotta, 2011). The first ones used simple geometric shapes to be simply viewed or grasped, while in the latter study Efron blocks were used and in the viewing condition a perceptual judgement had to be made. In both cases, the viewing condition produced first fixations closer to the center-of-gravity (COG) of the object (in accordance with (Foulsham & Underwood, 2009), among others), while the grasping condition was characterized by first fixations closer to the index finger location (or to the more difficult to grasp location).

In this paper, we present an experiment building on that of Brouwer et al. (2009). The main novelty of our approach is the use of real object stimuli (displayed on a monitor) and the comparison of three simple but realistic tasks, one 'passive' (classification) and two 'active' (lifting and opening).

We were interested in investigating to what extent eye movements subserve and anticipate the task demands, in the form of information collection for movement planning, and the relation to affordances (Gibson, 1979). This relation was expected to show in different scanning strategies determined by the different landmarks associated to each task. Even though the interaction with real objects in our daily life heavily relies on depth perception, Westwood, Danckert, Servos, and Goodale (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.

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. This was done to have an objective measure of the contact point regions that would be chosen for an actual grasp instead of arbitrarily choosing some expected ROIs. Both experiments are detailed in the following subsections.

Participants

Eleven participants (6 women, 5 men, aged 22-41) carried out the eye-tracking experiment in all 3 conditions (task). One female participant's data was discarded because of very bad quality. All subjects were right-handed with corrected to normal vision. Ten different (4 men, 6 women, aged 18-41) participants carried out the ROI extraction experiment. All of them were confirmed right-handed. In both experiments participants were compensated with study credits or money.

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 test objects) were chosen such that all of them could be easily lifted 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 the top. All 14 stimuli are showed in Fig.1. Each picture is 768 × 576 pixels. In each condition they were presented at mid-height on the right of the screen.

Apparatus and Procedure

Participants sat in front of the screen, where the object stimuli were presented. In the eye-tracking experiment their head was resting on a chin rest, about 70 cm away from the monitor, 1680 × 1050 pixels, 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.

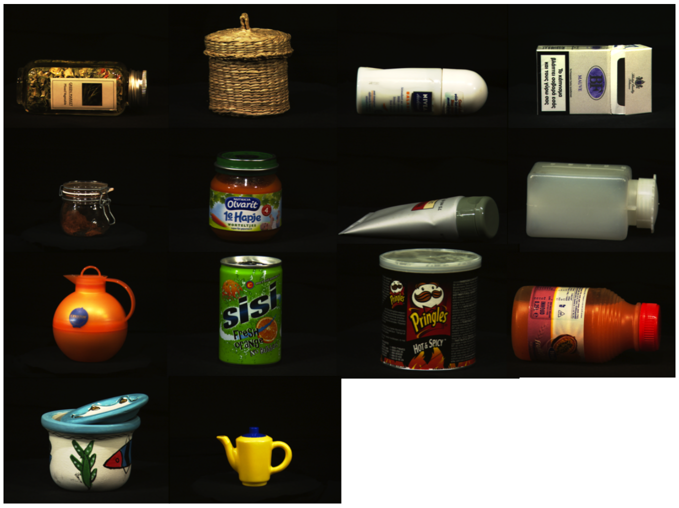

Figure 1: Stimuli pictures used in the experiment.

Eye movements were recorded via a binocular remote eye-tracker (EyeFollower, LC Technologies) working at 120 Hz. A keyboard was placed between the chin rest and the monitor to record reaction times. Participants had to look at the same stimuli with three different tasks in mind – each in one block. The task order was randomized across participants, so was the stimulus order within each block. For each task, every object was presented five times, resulting in 210 trials per participant. For training purposes, 30 more test 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 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 grasp objects, they were asked to always perform a grasp frontally, either with the thumb rightwards or downwards or by the handle, where present. As to the opening, they were told to imagine that the objects were glued to the shelf so 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 press the spacebar 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 at least 1 s (or as long as the space bar is not pressed); 3)

[1] The displayed object stimuli 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.

the stimulus appears on the right side of the screen; 4) Phase A: eye data and reaction times are collected up to the release of the space bar; 5) Phase B: eye data collected during the execution of the motor response; 6) the hand goes back to the spacebar and the next trial starts.

In the ROI extraction experiment, the same objects were presented to different participants. In just 2 blocks (lifting and opening), they were asked to place the tips of their fingers on the object, picturing the requested action. These points were recorded via a touch screen. After each trial, the participant was shown the selected points and, if not satisfied, she could repeat the trial. Every object was presented 3 times per block, resulting in 84 trials total per participant.

Data Processing and Analysis

Fixations for the phase A and B were extracted for each trial via the dispersion algorithm (Salvucci & Goldberg, 2000) with a temporal threshold of 100 ms and a spatial dispersion threshold of 1.5°. Data collected during phase A are supposed to be indicative of the information extraction and motor planning preceding movement initiation. Still, since in many cases there was just one or even no phase A fixation on the stimulus, quantitative evaluations were done on the first 3 fixations (or up to the third fixation) and on the mean of these first three fixations. 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 qualitative evaluation and informative visualization, heatmaps were computed from fixation data. These were obtained by placing a Gaussian with $\sigma = 1°$, 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. 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, hence mean and variance of the point coordinates sufficed to identify a rectangle containing the underlying region. In the case of 'lift', points were more evidently multi-modal, resulting in two major clusters one, smaller, for the thumb and one for the rest of the fingers. To include both clusters in the ROI, points were clustered via k-means, and a rectangle containing the region underlying both clusters was identified (see Fig.6, left, for an example of extracted ROIs). In most objects the two ROIs were well-separated. In a few cases, they were slightly overlapping and just in one case there was a major overlap. This, nevertheless, did not hamper the comparison with the heatmaps.

Results

Heatmaps

As a first qualitative impression of the general patterns of behavior observed in the three examined conditions, we compared heatmaps obtained from fixations of phase A, from first 3 fixations (in total and separated) and for the mean of the first 3 fixations. The same pattern was shown at different extents across all maps and objects, namely a maximum left of the object center in the 'classify' condition, a slightly higher-left of the center maximum in the 'lift' condition and a clear maximum on the opening region in the 'open' condition. Fig. 2 shows the phase A maps for one of the up-right objects and one of the horizontal objects. Already in phase A, task-dependent differences in eye fixations are evident.

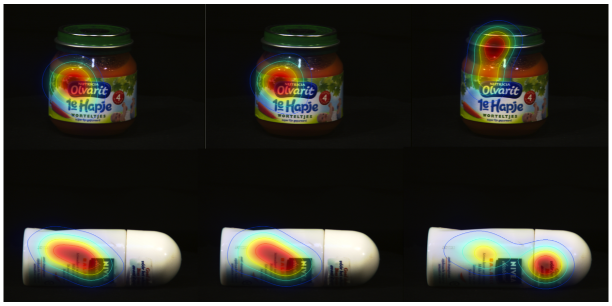

Figure 2: Heatmaps of the phase A fixations superimposed on corresponding stimuli. From left to right: 'classify','lift' and 'open' condition.

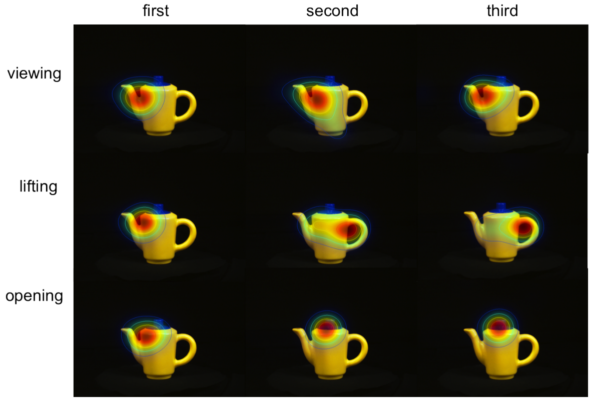

Figure 3: Heatmaps of the first, second and third fixation (left to right). From top to bottom: 'classify','lift' and 'open' condition.

An evolution in time across the first 3 fixations/conditions for one object is presented in Fig.3. If the first fixation is usually close to the COG (with some undershoot) for all conditions, already by the second fixation is possible to infer where the scanpath will lead. The first fixation was a 'phase A' fixation in 90% of cases, the second fixation in 53% , while the third just in 28%. Of 5733 examined fixations, 3359 were phase A. While for the first fixation phase A fixations are equally distributed across tasks (1832 A fixations, 34% classify, 32% lift, 34% open), in the second the proportion is in favor of lifting and opening (1037 A fixations, 24% classify, 33% lift, 43% open), by the few third A fixations mostly for the 'active' tasks motion had not yet initiated (490 A fixations, 17% classify, 38% lift, 45% open).

Average Fixations

The mean of the first three fixations (or up to 3) on each stimulus image was extracted for each trial. Often the first fixation was in the direction of the COG of the object but landed either on the black background or on the edge of the object, hence showing some undershoot along the x-axis (we use image coordinates since the objects are not shown in a completely frontal view but in perspective, hence the center of the object outline would not correspond to the COG). A repeated measures ANOVA on the x coordinate of the average fixation with task and object as within-subject factors showed a main effect of task ($F(2,18) = 36.9$, $p < .001$), a main effect of object ($F(13,117) = 19.87$, $p < .001$), and and interaction effect of object and task ($F(26,234) = 13.73$, $p < .001$). The mean X coordinates according to object and task are presented in Fig. 4. For most objects, the 'classify' mean position was the most left and the 'opening' the most right. This is of course more extreme for horizontal objects, e.g, the gel tube, the white jar, the juice bottle, while for three up-right objects (yellow tea pot, orange tea pot, and chips tube) the lifting mean position is to the right of the opening position either because the handle was on the right or the plastic lid was best opened by exerting force with the right thumb.

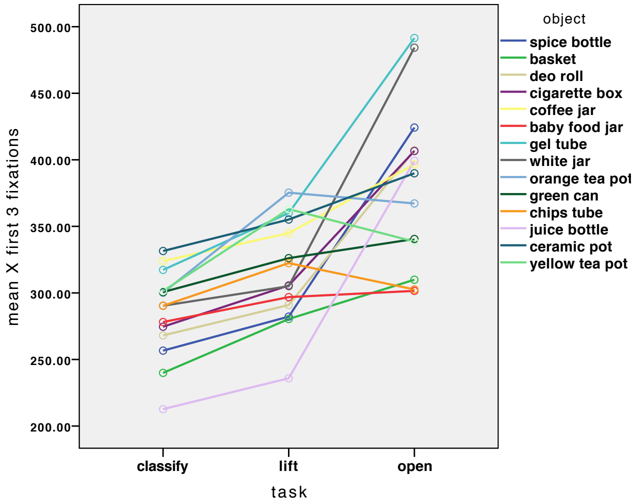

Figure 4: Mean X coordinate of each object across task.

An analogous analysis was performed on the vertical mean location. Again the effect of task was significant ($F(2,18) = 51.58$, $p < .001$) as that of object ($F(13,117) = 134.02$, $p < .001$) and interaction ($F(26,234) = 28.13$, $p < .001$). The mean Y coordinates according to object and task are presented in Fig. 5. In this case the ordinate is expressed in image coordinates, with origin in the top left corner. Up-right objects (such as the green can or the chips tube) present of course the most extreme mean vertical location for the 'open' task, while for horizontal objects the mean y location is always at the same height with a slight tendency upwards in the 'lift' condition.

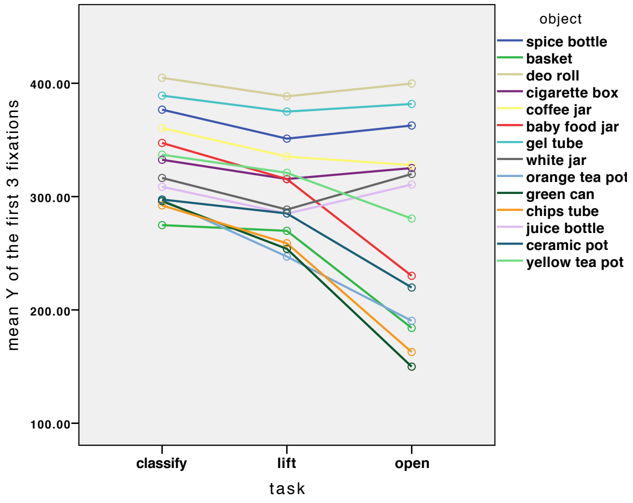

Figure 5: Mean Y coordinate of each object across task. Note that the y axis is in picture coordinates, hence the lower the value the higher the location in the picture.

Comparison Heatmaps-ROI

To gain a more specific insight regarding to what extent the fixation map 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 achieved considering 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 91.1 ± 59.52 pixel, while for the 'open' condition was 63.2172 ± 35.53. 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, $p < .001$).

Reaction Times

Mean reaction times in releasing the spacebar significantly increase from 'classify' to 'lift' to the 'open' condition. The difference is most pronounced between 'passive' and 'active' conditions (classify: 0.596 ± 0.052s, lift: 0.805 ± 0.126s, open: 0.826 ± 0.110s). A repeated measures ANOVA on the average reaction time with task and object as within-subject factors showed a main effect of task ($F(2,18) = 7.04$, $p = 0.006$) and a main effect of object ($F(13,117) = 2.14$,

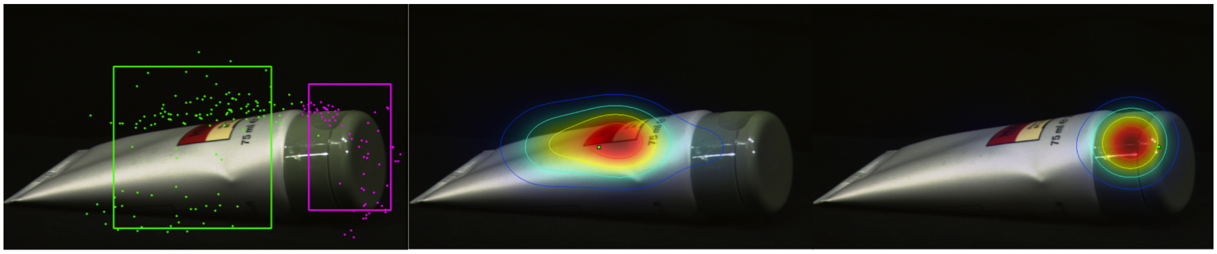

Figure 6: Left: touched points and Regions Of Interest extracted for one of the stimuli (green: 'lift' condition; magenta: the 'open condition'). Center: heatmap of the first 3 fixations in the 'lift' condition (in green the center of the corresponding ROI). Right: heatmap of the first 3 fixations in the 'open' condition (in green the center of the corresponding ROI).

$p = 0.016$). The mean reaction times for object and task are presented in Fig.7. Three objects (spice bottle, basket, and yellow tea pot) obtained shorter reaction times for opening than for lifting, in contrast to the general pattern – possibly because of the size difference compared to the real object, which made the decision on how to lift the object more difficult, and because of the particularly obvious opening action for all three objects. It must be noted that longer reaction times in the active tasks may be due not only to motion planning and affording points selection but also to the extraction of 3D information in absence of disparity cues.

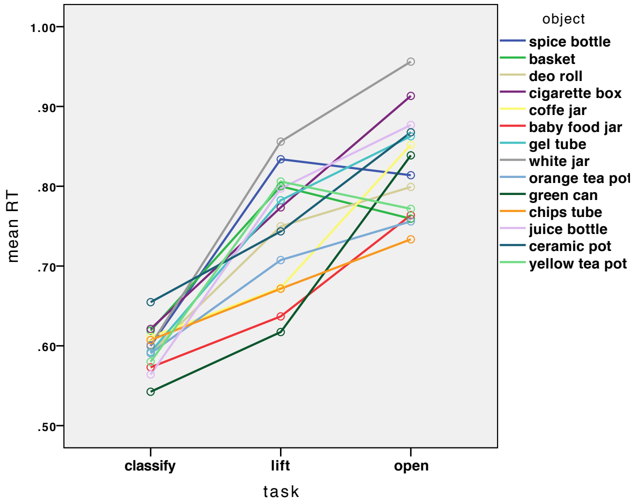

Figure 7: Mean reaction times of each object across task.

General Discussion

The presented experiment was aimed at assessing different eye movement strategies employed in identifying an object in contrast to tasks in which actual interactions had to be performed on the object. The distinct tasks as well as the object-specific affordance points were expected to strongly influence the distribution of eye fixations on each object. Indeed, we found significant differences in the scanpath behavior in the 3 conditions, suggesting for each one the construction of a specific attentional landscape around the informative/affording points.

In the classification task, the mean position of the first three fixations was mostly in the direction of the COG of the object. When grasping an object to lift it, fixations concentrated on a position to the left of and slightly higher than the COG. On the one hand, 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 in (Desanghere & Marotta, 2011). On the other hand, for up-right objects a preference to fixate more on the side of the thumb could be observed, while horizontal objects were on average fixated closer to the rest of the fingers. In the case of the two objects with a handle, there was a smaller peak in the center of the object (suggesting a first brief fixation there) and a higher mode on the handle, where later, longer fixations concentrated. In both cases it is possible that due to the objects' reduced size, subjects first considered lifting them with a power grasp and then went for the handle. In the case of opening, the fixation distribution presented a clear peak well localized on the opening region, which required the most processing for the planning of the finer motor operation (usually performed with a precision grip). 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. Beyond that, however, our data indicate that tracking eye movements 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 to timely infer the intention of impaired patients and possibly inform assistive interfaces to control prosthetic devices without the need of cumbersome training. The reliability with which the fixation mode consistently fell within the specific ROI supports considerations.

It seems plausible that the general flow of processing is first concerned with locating the object of interest (first fixation close to the COG). Next, it moves towards the most informative points – either for decision making in the case of the classification task, or for the purpose of executing anticipatory behavior control (Hoffmann, 2003; Butz, Sigaud, Pezzulo, & Baldassarre, 2007) towards interaction-relevant points (for lifting/opening) with proper behavioral interaction routines. In the former case, just 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 motions with properly aligned hand shapes (Jeannerod, Arbib, Rizzolatti, & Sakata, 1995; Cisek, 2007; Herbort & Butz, 2011). The consequentially more elaborate motion planning is also confirmed by significantly longer reaction times when an active motor task, different for every object, has to be planned anew.

In conclusion, as for more complex behavior, 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 intentions.

References

Baldauf, D., & Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision Research*, *50*(11), 999–1013.

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

Brouwer, A.-M., Franz, V. H., & Gegenfurtner, K. R. (2009). Differences in fixations between grasping and viewing objects. *Journal of Vision*, *9*(1).

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*(1485), 1585-1599.

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

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*(2).

Foulsham, T., & Underwood, G. (2009). Does conspicuity enhance distraction? saliency and eye landing position when searching for objects. *Quarterly journal of experimental psychology*, *62*(6), 1088–1098.

Geusebroek, J.-M., Burghouts, G. J., & Smeulders, A. W. M. (2005). The amsterdam library of object images. *Int. J. Comput. Vision*, *61*(1), 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*(13), 1567–1587.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20–25.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *J Vis*, *3*(1), 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.* Elsevier.

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

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* (p. 44-65). Springer-Verlag.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in neurosciences*, *18*(7), 314–320.

Johansson, R. S., Westling, G., Backstrom, A., & Flanagan, J. R. (2001). Eye-Hand coordination in object manipulation. *J. Neurosci.*, *21*(17), 6917–6932.

Kwok, R., & Braddick, O. (2003). When does the titchener circles illusion exert an effect on grasping? two- and three-dimensional targets. *Neuropsychologia*, *41*(8), 932-40.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*(11), 1311–1328.

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

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

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.

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