Affordance Estimation Enhances Artificial Visual Attention: Evidence from a Change Blindness Study

Jan Tünnermann
University of Paderborn, GET Lab
Pohlweg 47–49, 33098 Paderborn, Germany

Wail Mustafa
University of Southern Denmark, The Maersk Mc-Kinney Moller Institute
Niels Bohrs Allé 1, 5230, Odense M, Denmark

Norbert Krüger
University of Southern Denmark, The Maersk Mc-Kinney Moller Institute
Niels Bohrs Allé 1, 5230, Odense M, Denmark

Bärbel Mertsching
University of Paderborn, GET Lab
Pohlweg 47–49, 33098 Paderborn, Germany

Artificial attention models have been proposed to simulate human attentional behavior for the purpose to predict such or to endow technical systems with the ability to filter relevant from irrelevant information in visual scenes. Such models are typically based on the concept of saliency, which reflects the conspicuity of a visual entity regarding features such as color, intensity or orientation. Besides these stimulus-driven processes, a lot of effort has been made to enhance the models with top-down influences, which are known to govern human attentional behavior. Mostly, this aspect is considered in the form of specific targets for visual search or very general characteristics as the gist of a scene. For human attention it has been shown that objects which afford actions—such as graspable items in the action space—attract attention. Here we show that an artificial attention model that estimates such affordances can better predict human performance in a change detection task than a classic bottom-up saliency model. The implications are twofold: (1) The results add further evidence that human attention is highly influenced by affordances, which we can objectively model and compare to an objective control based on visual saliency. (2) The integration of affordance estimation into technical attention systems provides a top-down influence which is not overly specific or general but guides attention to objects which are potential targets of actions and with respect to the physical capabilities of the system; this is an advancement for technical cognitive systems.

Keywords: Visual Attention, Artificial Attention, Affordance, Change Blindness

Introduction

The phenomenon of change blindness (see e.g., Simons & Levin, 1997; Simons & Rensink, 2005), the inability of observers to notice changes in a visual scene when the presentation is interrupted during the change, indicates how sparse and transient visual scenes are represented in human vision. This striking effect is so stable that it also can be observed in real-world interactions and where the change is the (covert) replacement of a person during a conversation (Simons & Levin, 1998). It has been argued that visual attention is required to perceive the change and that in absence of a local signal (as motion or onset) top-down influences such as the scene context determine whether a change is noticed. Rensink et al. (1997) showed this by inducing changes either to objects of high interest or low interest, with the latter leading less often to successful change detection. When natural images are presented upside-down, this effect is reduced because the scene context is disrupted (Shore & Klein, 2000; Kelley et al., 2003). In a more recent study, Sampanes et al. (2008) showed that the gist of scene is stored independently from local features whose alterations remain unnoticed as long as the gist is preserved. Tseng et al. (2010) reported that performing actions to report changes improves the detection frequency, indicating that implicit information from the motor system can be accessed to support this task. Given this evidence for the importance scene context and task-relevance for guiding attention in the change blindness paradigm, it is not surprising that Stirk & Underwood (2007) found that using solely the bottom-up component of a computational attention model Itti & Koch (2001), it is unable to predict change detection for natural scenes.

Tünnermann et al. (in prep.) found an advantage for detecting changes to objects near the observer (possible action targets) over objects of similar perceptual appearance that in the images were out of the action space. Action possibilities that potentially influence perception were coined “affordances” by Gibson (1977). Affordances represent action possibilities that exist in the environment with regard to the action capabilities of the observer. A com-
mon example is a mug with a handle that affords grasping to humans. The important role of affordances in guiding attention has also been investigated (and confirmed) in reaction time tasks (see e.g., Craighero et al., 1999; Bekkering & Neggers, 2002), where facilitatory effects of spatial attention due to action possibilities have been found. These findings are supported by electrophysiological evidence (Handy et al., 2003).

Here we describe a computational attention model that includes an affordance measure as feedback from an higher level scene representation and show its ability to predict change detection in the change blindness paradigm. To our knowledge, this is the first artificial attention system that successfully predicts change detection in natural images. This does not only add to the evidence that action possibilities play an important role in attention-driven tasks, such as change detection, it is also a significant advancement for technical attention models that are often limited to bottom-up saliency processing or require specific target descriptions (templates) as a top-down influence.

### A Model of Affordance and Attention

To create an attention model that considers the aspects of affordance, we integrate a model of mid- and high-level visual representations (ECV; early cognitive vision; Pugeault et al., 2010) with a visual attention model. ECV is used to calculate potential grasping actions for objects in the scene and feed this information into the attention system. Because attention modulates visual processing at very early stages, the influence of grasping hypotheses estimated with ECV should be considered a feedback loop. The following briefly outlines the concepts and points to references for technical details and more extensive discussion of the different approaches used in this work. Then we describe the integration of artificial attention and ECV which we used in the experiments reported in this paper.

**Region-based artificial visual attention**

As a framework for the proposed attention system, we employ a region-based approach (Aziz & Mertsching, 2008; Tünnermann & Mertsching, 2013). Region-based artificial visual attention is a concept that is grounded in the field of computer vision for mobile robots, where attention must be integrated in a flexible framework and interact with other components of the agent. It does not necessarily simulate human attention and it does no mimic biological mechanism on a low level (in contrast to e.g. Itti & Koch, 2001). Instead, it applies abstract concepts of visual attention on the level of regions, coherent groups of pixels, which can be used for efficient conspicuity calculations and passed on to post-attentional processes.

The process is outlined in figure 1. After a pixel-based input image is segmented into coherent regions, which are stored in a region list (figure 1a, b, and c), different feature magnitudes are computed for each region. These are average color, orientation, symmetry, eccentricity and size. We refer to Aziz & Mertsching (2008) for detailed descriptions of how these features are computed. Once the feature magnitudes are available, saliency can be computed for each region and for each feature dimension. A top-down influence with regard to a specific template (itself a region) can enhance the saliency of every region that has similar attributes (Aziz & Mertsching, 2008); see Tünnermann et al. (2013) for a description of how more complex templates can be used). Bottom-up saliency is computed as a competition among the regions, where every region collects votes from its neighbors depending on its difference regarding a feature dimension (Aziz & Mertsching, 2008). In these processes, feature saliency maps (top-down and bottom-up, see figure 1d and 1e) are generated for each of the five feature dimensions, which are then fused into a single overall saliency map (see figure 1f). In this map fusion, weights can be considered that represent the relative importance of the feature dimensions and the bottom-up and top-down streams. The focus of attention is obtained by selecting the region with the maximum saliency in the overall saliency map. For selecting subsequent foci, inhibition of return can be applied (Aziz & Mertsching, 2007).

The region-based framework for artificial attention allows to easily integrate further influences (e.g., motion saliency, see Tünnermann & Mertsching, 2012) by assigning further attributes to the regions. We make use of this property in the present paper, by assigning an affordance value to each region. The calculation of the affordances requires a (sparse) 3D scene representation at higher levels, whose computation is described in the following section.

**Grasp Affordance Generation Based on an Early Cognitive Vision System**

The visual representation used for the computation of grasps affordances is based on the Early Cognitive Vision (ECV) system Pugeault et al. (2010). The ECV system produces a hierarchy of visual entities in both 2D and 3D spaces with a multi-modal description for each entity similar to the hierarchy computed by the human visual system (for a detailed discussion we refer to Krüger et al., 2013). This description contains geometric attributes, appearance attributes and uncertainty estimates (for details see Pugeault et al., 2010).

The ECV hierarchy works in two domains; the edge domain and surface domain where entities lie in different levels, from low-level features (such as line segments and texlets) to high-level ones (such as contour and surflings) (Kootstra et al., 2012). In this paper, the surface domain hierarchy only has been used. With respect to sensors, ECV supports both stereo vision as well as Kinect cameras (Olesen et al., 2012). For this study, we make use of the surface domain’s high-level entities extracted from a stereo camera. Figure 2a shows the surface hierarchy applied in this work.

Surface-based grasps are constructed around individual surfaces in 3D space, see figure 2b. This method creates simple actions aiming at grasping a surface as a whole. We perform these towards the positions of the surflings belonging to a specific surface. The grasps are generated with respect to the main directions of the surface derived by means of PCA. For details about this and a second surface-based method, we refer to Kootstra et al. (2012).
Figure 1: a Exemplary pixel image. b The image is segmented by region-growing; in this example, region 2 is currently growing. Adjacent unsegmented pixels are selected as candidates C and their color is compared to region’s seed (white 2) and the current region border (black 2s). If the candidates do not pass thresholds regarding these color differences they are added to the region (labeled with the region number) and the process continues. When no new pixel can be added to the region, the next unsegmented pixel becomes the seed for the next region; this is repeated until the whole image is segmented into coherent regions. c These regions are represented as a region list and their features are calculated (here illustrated as the average color, which is one feature that is stored for the region). d For each feature dimension (here illustrated for color) top-down saliency can be calculated by assigning activation to the region depending on the similarity with templates (t); the strength of the arrows represents the similarity of the template and the region. e Bottom-up saliency is calculated by collecting votes for each region (here region 2) from neighboring regions that are weighted by the difference (arrow strength) regarding the feature. f The results of these processes (from all feature dimensions) are combined in the overall saliency map.

Integrating ECV and region-based Attention

As first step in combining the grasp affordances described in the previous section with the region-based attention framework, the hypothesized grasps in 3D space are filtered to exclude grasps that cannot be performed (see figure 3a). Here we use a simple distance criterion and exclude every hypothetical grasp that was further than 0.7 m away from the observer. A more accurate performability analysis may be conducted by including concrete mechanical models of the grasping system and apply inverse kinematics and consider the complete scene geometry, as it might block certain trajectories. In the context of early affordance influences on attention, we regard the simple distance-based criterion as sufficient. Every hypothetical grasp that has not been removed by the filtering is the projected into the 2D image space of the saliency map (figure 3b). The position used for back-projection is the center of the surface that is approached by the grasp. Alternatives, as e.g. the contact points of the simulated gripper, do often fail to hit the corresponding region, as they are located close to the region border and sometimes integrate with the background. The number of grasping hypotheses is normalized by the area covered by a region. This value constitutes the affordance estimate that is ascribed to a region; the higher the value, the more grasps are possible towards the object represented by the region. Note that in the idealized depictions in figure 3, each region corresponds to one object, which must not necessarily be the case for real images, so it is possible that parts of objects have different affordance values.

For the experiments we report in this paper, the focus of attention is obtained by only considering the affordance values, as shown in figure 3. Because of the universal nature of the region lists, top-down and bottom-up saliency calculations can be performed as described in section Region-based artificial visual attention. Such contributions can be combined as indicated by the faint portions of figure 3 and fused with the affordance results.

Figure 2: a ECV surface hierarchy b Encompassing (EGA\textsubscript{e}) and pinch (EGA\textsubscript{p}) elementary grasps. c Exemplary grasps computed on image data used in the psychophysical experiment.
Figure 3: a Grasping hypotheses that have been computed in 3D space (see figure 2) are filtered with regard to the distance. If they are too far (here drawn in gray) they are excluded from further processing. 
b The affordance value of each region is calculated as the density of grasping hypotheses, estimated by the number of hypothetical grasps projected into 2D per region area. The position that is projected into 2D is based on the 3D surface position that was approached by a grasp.

Experiment 1

In this experiment, the change blindness paradigm was used to measure whether observers deploy more attention towards objects which are highly afford ing according to the proposed model or whether more attention is directed towards highly salient objects according to the classic bottom-up saliency model by Itti & Koch (2001).

Participants. In this experiment, 40 subjects with an average age of 23.23 (SD = 3.40) participated. All had normal or corrected-to-normal vision and had not seen the images before. Because the session required less than five minutes, the subjects participated without any reward, however, some did the experiment in addition to longer unrelated experiments for which they received course credit or money (6 € per hour).

Stimuli. The stimulus material consisted of 28 photographs of indoor scenes, which are depicted in figure 4. The majority were office scenes and all setups contain objects in the near action space and the background areas, where no direct action can be executed towards the objects. The images were recorded as stereo-pairs with a bumblebee stereo camera and processed with the rivaling attention models. For our affordance-based model, we processed the left and right images as described in section Integrating ECV and region-based Attention for every scene, with the results that one region of the initial segmentation had the highest activation. The corresponding object was removed from the rectified left image to produce the change-blindness pair “original to affordance-based prediction” (OA in the following). The right images of the stereo-pairs were only used in the process but were never shown to participants. The predictions from the model by Itti & Koch (2001) were obtained by running its publicly available implementation1 on the rectified left image. The object which was hit by the model’s focus (location of highest saliency peak) was then removed for the post-change images to create an “original to bottom-up saliency prediction” pair (OS). Figure 5 shows an exemplary scene and the saliency- and affordance-based changes. Whenever both models made the same prediction the images were discarded and the scene rearranged and rerecorded, however, this happened only three times. On another three occasions scenes had to be rearranged because the models suggested objects that could not be removed 2.

Figure 5: a One image with objects marked that were removed ( for OA and for OS). The left images of b to c relate to Itti & Koch’s saliency model, the right images to the proposed affordance-based approach. b Corresponding saliency maps. c Changed image with the object removed. d Absolute intensity difference between original and changed images normalized over both images.

1http://ilab.usc.edu/toolkit/downloads-virtualbox.shtml
2The affordance model suggested edges of shelves twice, which could not be removed without also removing all their contents or leaving the contained objects suspiciously floating; the saliency model also suggested such element once by pointing to freespace on a table between objects.
the objects to leave no suspicious looking locations. Influences on
distance scene parts (global illumination changes or, distant reflec-
tions) were not removed, to retain an unambiguous change at the
predicted location. The resulting inconsistencies are not expected
to be distracting or suspicious during the short exposure durations
in the experiment. An example can be seen in the right-row image of
figure 5c, where the water boiler still shows the reflection of the
removed cup, but no local conspicuity is created nor is the gist of
the scene altered.

Figure 5a shows a scene with the objects marked that were re-
moved with regard to the model predictions, whose saliency maps
are shown in figure 5b. In figure 5c the corresponding images with
the predicted objects removed are shown and in figure 5d the mag-
nitude of the absolute intensity difference between original and
changed image is shown for each version. The affordance-based
change resulted in a relatively low difference, which is not sur-
prising because the high intensity of the object does not neces-
sarily contribute to its conspicuity in the affordance-based predic-
tion as it does for the bottom-up saliency prediction. Figure 6 em-
phasizes this, showing that accumulating the intensity differences
over all images leads to substantial lower values at locations of the
affordance-based predictions (figure 6b) compared to the bottom-
up saliency prediction (figure 6a). The figure also shows that the
changes are distributed equally over the scenes. The affordance-
based changes appear a little biased towards the center, which is
likely to depend on the distance from camera to object which is re-
stricting the estimated graspability (objects in the corners are often
too far away). Furthermore, a vertical strip on the left in figure 6b
shows no changes. This is because the left image of the stereo pair
was used and in the far left are no stereo-correspondences (and thus
no object representations and no estimated grasping possibilities).

Though fainter in intensity, the affordance-based changes are often
a bit larger in spatial extent (compare ellipses in figure 4). How-
ever, because both models processed exactly the same input im-
ages, and thus had the same features as intensity, size and location
available, this should not be considered a confound but rather re-
flects inherent properties of the different processing methods.

**Design.** In a within-subjects design, image pairs were shown with
the change either being the object removed that was suggested by
the affordance model (OA) or the object suggested by Itti & Koch’s
bottom-up saliency model (OS). Whether a specific image was
shown to a participant with an OA or an OS was varied between
subjects, but there were always 14 (the half of all trials) of each
condition. Over all 40 participants, each scene was shown 20 times with the OÅ and 20 times with the OS change.

**Procedure.** The participants were instructed to stand at comfortable reaching distance in front of a 22" touchscreen monitor. They were asked keep their arms close to the body between trials (as we observed in previous experiments that some participants tend to extend their hand close towards the screen before the trials start, occluding parts of the images). Subjects then read an on-screen instruction that explained the task and they performed three example trials. The example trials contained outdoor scenes unlike any of the stimulus material of the main experiment.

During the experiment 28 trials were shown, half of them containing OÅ changes and the other half OS changes. They were shown randomly intermixed, so different participants had different sequences with regard to the image order and the change condition.

The start of each trial was initiated by the previous touch response, followed by a 1000 ms interval. Then the original image was shown for 400 ms followed by a 100 ms blank and then the changed image was shown for 200 ms (see figure 7). After this sequence, and empty screen was shown to record the response, for which the participant touched the region of the screen at the location where the change was noticed or at a guessed location in case the change had gone unseen. Because the monitor had a 19 to 10 aspect ratio and the images a 4 to 3 ratio, they did not fill the whole screen. So during the empty presentation for the response, a rectangular border of where the image had been was shown as a frame of reference. Other than that, the background was kept at a bright gray before the trials, during blanks and for the response. A hit was regarded when the touch response was located within a 400 × 400 area centered at the center of the removed object’s bounding box.

This supports our hypothesis that affordance, as measured by our model, plays a more important role for the deployment of visual attention than pure bottom-up saliency, as measured with the model by Itti & Koch (2001). Indeed it could be strong evidence, however, it can be argued that it is only an assumption that the success of the proposed model is due to its ability to capture object affordances. The objects predicted by both models are relatively salient, as compared for example to empty space or homogeneously textured background. Therefore, it is possible that the proposed model is simply a better saliency predictor than the model by Itti & Koch (2001). To test whether this is the case or if the model’s success is genuinely based on the inclusion of object affordances is tested in a second experiment.

**Experiment 2**

If the advantage of the proposed model originates from its ability to capture affordances, the prediction performance of the affordance model should be reduced by a manipulation of the stimulus material that disturbs the perception of affordances. We hypothesize that in contrast the performance of the saliency model should remain constant, given that the manipulation does not alter the local contrast. Such manipulation is the rotation of the images by 180°. The local contrast remain unchanged, but the perception of affordances is disturbed because the positions of objects within the image and their relationships are affected. For example, for the scene shown in figure 9, the geometry of the rotated scene suggests that the area which is the working space in the original image is now a wall, and vice versa. The apparent orientation of the object predicted by the affordance model no longer offers the possibility of a direct power grip in the upside-down version of the image. On the contrary, local contrasts of color, intensity, or ori-

---

3 All t-tests and ANOVAs reported in this paper assume an alpha level of 0.05 and were performed on the arcsine-transformed relative frequencies. Whenever differences are not significant, we additionally report the mean and standard deviation of the differences and 95% confidence intervals around the mean difference.
orientation remain unchanged (with regard to the transformed image coordinates). Hence, when observers perform the same task as in the previous experiment, the changes made due to the affordance-based prediction should be detected less frequently as on the correctly oriented images whereas the success of the saliency-based predictions should not change. The same logic has been applied by Tünnermann et al. (in prep.) to distinguish between detection success contributions due to saliency and due to object location in the action space.

**Participants.** In the second experiment, 40 participants with an average age of 26.13 (SD = 6.17) participated. All had normal or corrected-to-normal vision and had not seen the images (or their upright original versions) before. As in Experiment 1, the subjects participated without reward.

**Stimuli, Design and Procedure.** The experimental design, stimulus material and the presentation procedure was kept exactly as in Experiment 1, with the only difference being that all images (including the changed ones) were rotated by 180°, that is, they were presented upside-down.

**Results and Discussion.**

The difference between the prediction based on the affordance model and the saliency model is reduced in Experiment 2, however, the difference is still significant (see figure 10), t(39) = 5.33, p < 0.001 in a paired two-tailed t-test. Comparing the results from the first and second experiment, the hit frequency of for the affordance-based prediction is significantly lower, while the saliency-based prediction is not significantly different. This shows in a mixed-design ANOVA with the data from Experiment 1 and Experiment 2 and repeated measures regarding which model made the prediction.

![Image](image_url)

**Figure 9:** Rotated image as used in Experiment 2. The figure shows the original image (pre-change) with the potentially changing objects marked for illustration. The affordance-based prediction ( ) and the saliency-based prediction ( ) are based on the model output regarding the unrotated images. In other words, exactly the same objects as in Experiment 1 were changed.

![Image](image_url)

**Figure 10:** Average relative change detection frequencies for the bottom-up saliency condition and the affordance-based condition (** p < 0.001). Error bars show the standard error of the mean.

There is main effect for model F(1, 78) = 126.61, p < 0.001, as well as for image orientation F(1, 78) = 16.66, p < 0.001, and a significant interaction, F(1, 78) = 7.61, p < 0.01. As a post test, Welch’s two-sample t-test confirms that the affordance-based prediction is significantly worse when the images are shown upside down, t(74.68) = 4.38, p < 0.001. The difference of the saliency-based predictions (the pink and purple bars in figure 11) is not significant t(78) = 1.4, p = 0.18. The mean difference is is only 0.04 (SD = 0.03, 95% CI [−0.02, 0.1]).

![Image](image_url)

**Figure 11:** Average relative change detection frequencies for the bottom-up saliency condition and the affordance-based condition (** p < 0.001). Error bars show the standard error of the mean.

This second experiment provides support for the claim that the proposed model performs better because it is based on a measure of affordance. That a significant difference in favor for the affordance model remains when the images are turned upside down (see figure 10), could be due to different reasons: the concept of affordance is...
based on several attributes, such as graspability and location within the action space. By turning images upside down, not all of them are disturbed. For example, the tendency of affordance-based predictions being closer to the image center (due to the fact that the corners often depict scene elements farther away from the observer and therefore not in grasping reach) persists even when images are turned upside down. Similarly, local properties such as visibility (also saliency) and whether the image portion is cluttered with obstructing objects are also related with affordances and are preserved when images are turned upside down. Finally, it must be said that it is still possibility that affordance-based model predicts object of higher saliency than the saliency model by Itti & Koch (2001). Experiment 2 however shows that this cannot be the only reason and that a substantial part of the advantage is due the relation of objects and scene geometry, which is part of the affordance concept.

### General Discussion

We proposed a technical attention system that includes a measure of affordance obtained from an artificial graspability estimation. It is included in a framework of region-based artificial attention and can therefore be combined with traditional region-based bottom-up and top-down mechanisms. In this paper, we used the pure affordance component of the proposed model to compare its predictions in change blindness experiments to predictions of the bottom-up saliency model by Itti & Koch (2001). The results of Experiment 1 show a significant advantage in change detection prediction for proposed affordance-based model. Experiment 2 supports the interpretation that it is really the affordances and not an improved saliency detection leading to this result. Thus it can be accepted that the artificial affordance estimation enhances the proposed model.

However, another potential objection must be discussed: Is it really the deployment of attention, which is in favor for the objects selected by the affordance model? In the forced response paradigm participants must select a location on the screen, even when they did not notice the change (and thus have not attended the location). In this situation they may tend to select certain objects (instead of empty spaces) they consider conspicuous. In this decision they may be biased to select objects in the action space over objects at other locations in the scene and therefore the affordance model’s advantage could be based on its ability to better predict this decision and not on better predicting the deployment of attention. Regarding this concern, we analyzed the hit frequencies for the object that did not change. For the images for which a participant saw the version with the affordance-based change, we counted the hit frequency for the locations of the saliency-based predictions (where no change appeared for the participant in these trials). Inversely, hit frequencies for the affordance-predicted locations were obtained when the bottom-up saliency change was shown. In the following, these positions are referred to as unchanged candidates. An example for an unchanged candidate is the blue hole puncher in figure 7.

Indeed the unchanged candidates are selected more often ($M = 0.15$, $SD = 0.07$, in Experiment 1 and $M = 0.14$, $SD = 0.07$, in Experiment 2) than by chance (chance level = 0.09) reflecting the participants guessing strategies that avoid unlikely places (e.g., in the corners) which are included in the calculation of the chance level. These above-chance detections are significant in both experiments, $t(39) = 5.17$, $p < 0.001$ and $t(39) = 4.13$, $p < 0.001$, respectively. However, note that as shown in figure 4, the locations of the affordance- and saliency-based predictions sometimes overlap. Consequently, for this analysis a hit for the changed object is sometimes also a hit for the object that was predicted by the other model but which did not change. More importantly, there is no significant difference between hit rates for unchanged candidates from the affordance-based predictions and the saliency-based predictions which would be a hint for a decision bias contribution. The differences of the means of the affordance- and saliency-based unchanged candidates is only 0.03 ($SD = 0.02$, 95% CI $[-0.02, 0.07]$) in Experiment 1 and 0 ($SD = 0.02$, 95% CI $[-0.05, 0.05]$) in Experiment 2. Thus it can be concluded that the advantage of the affordance-based model is genuinely due to its ability to better predict the deployment of visual attention and not based on a decision bias contribution.

These results are in line with Stirk & Underwood (2007) who showed that a context factor, scene consistency, plays a role in change detection while low-level saliency does not. Similar context effects as discovered by Rensink et al. (1997) have been shown to be disturbed when the images were presented upside down (Kelley et al., 2003). The same is true for the enhanced change detection for objects depicted in the action space (Tünnermann et al., in prep.). In this study we were able to conjointly represent such factors by the concept of affordance and provide technical measure for it. Thus, by contrast to Kelley et al.’s work and (Tünnermann et al., in prep.), the present study did not require to previously rate the level interest or to manually arrange and decide fore- and background objects which are changed. With the objective measure we replicated the disturbance of the context effect by scene inversion.

The technical graspability estimation was done in a very approximative manner using a simple simulated gripper with no restrictions on its kinematics. Refinement of this simulations allows to further enhance the affordance estimation with regard to system of interest. Bringing it closer to human physiology may lead to better prediction of human attention in action contexts. Simulating the action possibilities and restrictions of a specific technical system (e.g. some robot with a certain grasping device) can improve artificial attention systems for such machines.

In terms of efficiency, however, the current implementation has a conceptual drawback: We make use of the ECV system as described by Pugeault et al. (2010) and feed back the grasping possibilities into the region-based attention framework by Aziz & Mertsching (2008). Thus, the representation used to obtain the grasping possibilities is generated (based on ECV) in a completely attention-less manner. A goal for the future is an early integration of both systems, in a way such that attention also guides the creation of the scene representation. Then high-level processes, such
as the graspability estimation, feed back to the attention system. The dynamics of such a system have to be studied.

To summarize, we presented a technical attention model that includes an affordance measure based on graspability estimation. The model is able to better predict the deployment of visual attention for indoor scenes with action possibilities than bottom-up saliency does. This confirms the importance of scene context for attention-guided tasks such as the in the change blindness paradigm. Furthermore, it provides an interesting concept to account for the physical abilities of technical systems for artificial attention systems. Up to now, artificial attention systems were limited to bottom-up saliency and top-down influences in the form of perceptual templates of varying complexity (e.g., “attend color”, “attend red things” or “attend stop signs”). A system which considers affordances in the control of attention, allows to deploy attention to potential action targets and does not require to include any prior knowledge about the appearance of potential targets.

Acknowledgements

This work was supported by the EU Cognitive Systems project Xperience (FP7-ICT-270273).

References


Rensink, R. a., O'Regan, J. K., & Clark, J. J. (1997, September). To see or not to see: the need for attention to perceive changes in scenes. Psychological Science, 8(5), 368–373.


Tünnermann, J., Hilkemeier, F., & Scharlau, I. (in prep.). Change detection is enhanced for objects in action space.