Saliency and Affordance in Artificial Visual Attention

Jan Tünnermann & Bärbel Mertsching
GET Lab
University of Paderborn
Pohlweg 47–49, 33098 Paderborn, Germany
{tuennermann, mertsching}@get.uni-paderborn.de

***Abstract*—Affordances, as for example grasping possibilities, are known to play a role in the guidance of human attention but have not been considered in artificial attention systems so far. Extending our earlier work, we investigate the combination of affordance estimation and visual saliency in an artificial visual attention model. Different models based on saliency, affordance estimation, or their combination are suggested and evaluated via their predictions for a change detection task with human observers.**

I. Introduction

With the transition of robots from specialized automata performing predefined tasks to general autonomous agents, the requirements to perceive, reason about, and interact with their environment have drastically increased. A recent development in robotics is to model aspects of environmental psychology, which deal with the interaction between humans and their surrounding. A popular concept from this field, the *affordance of objects*, was introduced by J. J. Gibson in 1977 [7]. In this holistic view, objects posses certain affordances, i.e., objects or their parts can afford certain actions. A common example is a mug, whose handle affords grasping.

This idea has been transferred to technical systems to enhance their performance preparing and executing actions that correspond to such affordances [see e.g., 5, 24, 25]. One result of such studies—as well as of earlier general object recognition studies [e.g., 18, 15]—is that the affordance concept in technical systems not only supports direct grasping actions, it can also benefit object recognition and semantic scene perception. The general idea is that in many cases an object is better defined by the actions which the object supports, than with its visual attributes. Coming back to the example of a mug, even though colors and shapes may differ widely, mugs in general afford grasping (possibly by some kind of handle), containing liquid and drinking [4]. Therefore, recent research integrates affordance estimation with object recognition [e.g., 8, 4] and the semantic interpretation of scenes and objects [e.g., 27, 28].

Visual attention is another concept which is being transferred from cognitive psychology to technical systems. The main idea is to filter relevant from irrelevant information very early in processing, and distribute processing resources accordingly. Attention can be guided bottom-up by saliency—local contrasts with respect to features such as intensity, color, or local orientation [10]—or in a top-down manner by incorporating knowledge, such as task demands [11] or the "gist of the scene" [12].

Findings from psychology suggest that affordances influence visual attention. This has been shown in reductions of reaction times when affordances were used to guide attention towards target locations [16, 6] and effects on event related signals in electrophysiological and brain imaging research [9].

The influence of affordance on visual attention can be regarded as a way to bias the distribution of processing resources towards objects that afford actions. These constitute potential targets of actions or reasoning, even though a specific behavior towards them may not be planned at this stage.

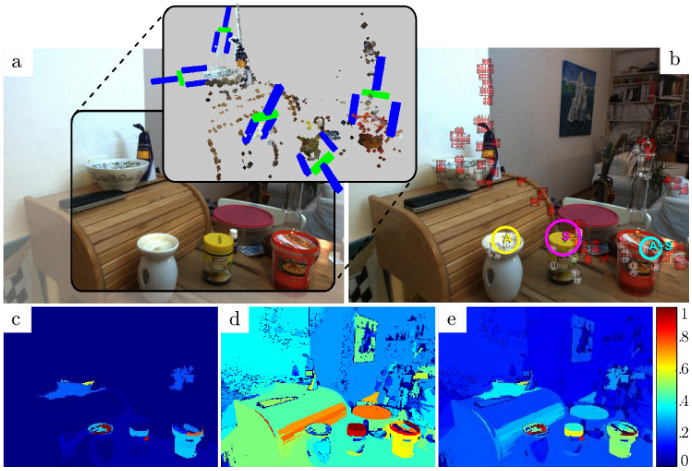

Fig. 1: **a:** A test scene. Inset: exemplary grasps fitted on a sparse 3D scene representation. **b:** Grasp points projected into 2D. White patterns represent grasps towards reachable locations, whereas red patterns indicate locations out of reach. "A" affordance-based, "S" saliency-based, and "A+S" combined selection. **c–e:** Underlying affordance (c), saliency (d), and combined (e) maps.

Recently, we investigated whether an artificial visual attention model can better predict human performance in an attention guided task when the prediction is based on affordance estimation instead of saliency estimation [22]. We found that this indeed is the case. Utilizing a change detection task, we compared a popular saliency model [10] with an attention model based on grasping hypotheses (estimated from a sparse scene representation [13], see figure 1a). The phenomenon of change blindness due to short scene interruptions renders the

detection of changes in objects difficult, but the performance is increased when attention is allocated to the changing part of the scene [14]. For the evaluation of psychologically inspired computer vision systems, the change blindness paradigm has the great advantage that images with natural scenes can be used, whereas many other psychophysical tasks require the use of highly artificial synthetic stimuli. We found that human observers performed better in reporting the changes that were made to objects selected by the affordance-based model than when those selected by the saliency model were changed.

In the present study, we extend our previous work [22] to investigate a model based on combined affordance and saliency, as opposed to the models that used either individual component. If this enhances the prediction of change detection, it may better reflect the allocation of attention. This may be the case, for example, when there are several objects of high affordance and their individual saliencies tip the scale for selection priority.

The rest of this paper is structured as follows: In section II, we report a new change detection experiment that uses the stimulus material already used in [22] but a different task. This is to test this task and get a preliminary insight how saliency and affordance are related in this stimulus material. Then, in section III, we describe a new model of combined saliency and affordance which is used to create new test images to compare this combined model (in section III-B) with separate saliency and affordance models in the change blindness task. The implications are discussed in section IV.

II. Saliency and Affordance in Change Detection

A. Experiment I - Motivation

In our previous study [22], we employed a "single-shot" change detection task with natural images [see e.g., 20] to measure the participants' distribution of attention towards salient or affording objects. The single-shot paradigm shows a single change from the original to the altered image which are shown only briefly (usually between 100 and 500 ms) and a blank screen is shown between the two images. The presentation usually lasts for less than a second and participants respond afterwards, when the image is already gone. Thus, there is only a single binary hit-or-miss measurement per change. Furthermore, the same images cannot be repeated and therefore the amount of trials is limited to the number of available images. Their creation is quite an effort, due to editing in the changes (object removals in our case). Because of the limited number of trials and the binomially distributed response, a large number of subjects is required (40 - 80) to obtain reliable results. In our previous study there were two possible changes for every image—saliency or affordance—and thus two subjects were required to obtain an affordance and saliency measurement for every image. In the experiment we report in section III-B, three changes were possible. This further increases the number of subjects required.

Hence, in this experiment, we test the so called "flicker paradigm" [see e.g., 14]: the presentation is similar as described above, but it is repeated until the change is reported. Therefore, a more informative measure, namely the time it takes the subject to detect the change, can be obtained. This not only reduces the number of participants required, but may also allow to relate the degree of affordance and saliency to the response time. Therefore, the objective of this first experiment is, using the stimulus material from [22], to investigate whether the effect that affordances are more important than saliency in change detection can be replicated using the flicker paradigm. Furthermore, a first insight in the influence of the saliency and affordance values on the response time is provided.

1) Participants: Twelve volunteers (average age of 26.82, $SD = 3.6$) participated in this experiment. All had normal or corrected-to-normal vision and not seen the images before.

2) Stimuli: The stimulus material reported in [22] was used. This consisted of 28 natural scenes, mostly pictures of office environments that contained a number of objects in the reachable action space and some in background areas which would not be reachable by the observer of the scene. For every image, two changed versions were created by locally altering the image (locally blending in an identical image in which the object had been removed at that location). In one altered image of the same scene, an object selected by the saliency model by Itti et al. [10] had been removed, in the other one object selected by an affordance-based prediction (density of grasping possibilities per image area; this corresponds to the affordance stream described in see section III-A) had been removed. Refer to [22], for more details regarding the stimulus generation and the actual pictures.

3) Design and Procedure: The experiment was conducted on a 12.1" touchscreen laptop[1]. The participants were instructed to normally sit in front of the laptop and to adjust the distance and display angle for optimal viewing and touching.

Participants were presented with every original image paired with one of the possible changes. Thus, a single participant saw 14 original–affordance and 14 original–saliency pairs in random order. The number for which each changed image appeared with the affordance or saliency change was balanced over all subjects.

In contrast to the task used in [22], where participants saw the change only once (from original to the changed version), here the images cycled back and forth between the original and the changed image until the subject discovered the change and responded by touching the screen at the location of the change. If no response was made within one minute, the current trial was aborted and the next trial started. The timing of the image sequence was: "1000 ms initial blank"–"300 ms original image"–"300 ms blank"–"100 ms changed image"–"300 ms blank". The sequence after the initial blank was repeated until the participant responded. The changed image was shown for a shorter time as we are interested in the distribution of attention in the original image, reducing the risk that highly salient (or affording) objects that become visible

[1]Note that the change blindness effect is very robust and does not require highly accurate timing that can only be established with CRT monitors or specialized equipment, which is the case for many other psychophysical paradigms.

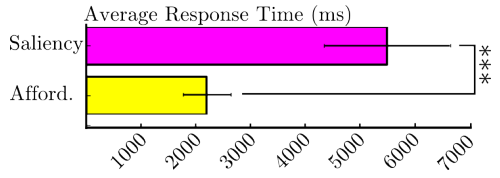

Fig. 2: Average response times for affordance- and saliency-based changes (error bars show the SEM; *** $p < 0.001$).

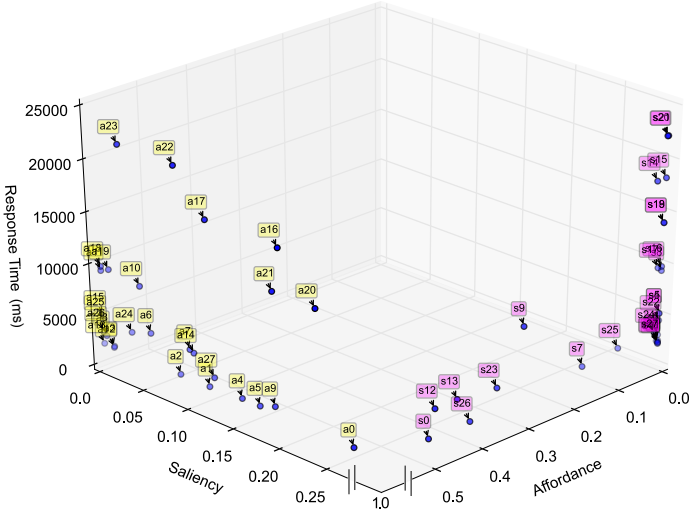

Fig. 3: Average response times (in milliseconds) of each change plotted into the affordance–saliency space with the saliency and affordance values at the changed locations in the respective maps. Points with pink labels ("sX") are saliency-based changes, whereas yellow labels ("aX") mark affordance-based changes in images X. Because the most salient and most affording objects had been selected for a change, they are set to maximum (1.0) in their respective dimension. The value for the other dimension was obtained by averaging the respective map under the removed object's bounding rectangle.

in the changed image substantially influence the response. For every trial, the response time was recorded.

4) Results and Discussion: Figure 2 shows the average response time to saliency- and affordance-based changes. Affordance based changes are reported significantly faster, $t(11) = -5.03, p < 0.001$, confirming our earlier results [22] from the single shot hit-or-miss task. This provides further evidence for the importance of affordances in the deployment of attention. The one minute time limit was reached only four times in the 336 changes presented over all subjects.

Furthermore, based on visual inspection, the distribution of responses times per image in the affordance–saliency space (see figure 3), suggest that no long response times are found when a change has also high activity in the respective other dimension. Especially for the saliency-based changes, long response times occur where the affordance is close to zero. Therefore, an attention model that combines saliency and affordance could show a better performance than models based on each individual component. Such a combined model is described in the following section and tested experimentally in section III-B.

III. Combining Saliency and Affordance

A. A Combined Model of Saliency and Affordance

The proposed combined model of saliency and affordance is presented in figure 4. The processing begins with the acquisition of a stereo image pair. The saliency maps and the eventually combined attention map are based on the left image. Therefore, the left image is segmented into homogeneously colored regions ① (see [2] or [3] for segmentation methods for region-based attention on which this model is based). These regions can be considered proto-objects, which at pre-attentional stages are used for the feature and saliency computations and later integrate the different saliency dimensions and the affordance estimation.

In the saliency stream, the regions are used to generate feature magnitude maps for *color*, *orientation*, *eccentricity*, *symmetry* and *size* (s2). The feature *color* is obtained as the average color of all pixels of a region. *Orientation*, *eccentricity* and *symmetry* are calculated based on 2D central moments of the spatial distribution of a region's pixels. *Size* is the number of pixels in a region. For details on the calculation of feature magnitudes, please refer to [1]. As a next step in this stream, saliency maps are calculated for each feature dimension individually (s3). This is done by applying a voting style procedure, where each region collects votes from its neighbors, regarding the dissimilarities in every feature dimension. Details on how the difference between two regions in a specific feature dimension is measured can be found in [1].

In the affordance stream, the left image is used to generate *2D Texlets* which are small local texture patches (a1). Using stereo disparities (a2), the *2D Texlets* are transferred into 3D space to form *3D Texlets* (a3). There, position-based k-means clustering is used to form small groups of neighboring *Texlets*. Planes are fitted through these *groups* to form so called *Surflings* (a4), which are then further grouped (when close to each other and similarly oriented) to generate *Surfaces* (a5).

Grasping hypotheses in 3D space are generated by fitting a simulated gripper (see figure 1b) to elements of the scene considering the surfaces generated in the process outlined above (a6). For details on the generation of this feature hierarchy and the grasping hypotheses, see [13]. The result of this process, which we make use of in this study, is the estimated contact points of the gripper on the surfaces. Note that in the present study the simulated gripper performs simple two-fingered grasping and does not include a bio-mechanic model of grasping. The latter can be used to filter out grasping hypotheses which may not be performable by the observer. The only criterion which is used to filter grasping hypotheses with regard to performability is a distance limit of 70 cm, which roughly corresponds to human reaching distance.

The combined attention map is then obtained by integrating the individual saliency maps (s4) (according to the combination strategy described in [1], which assigns higher weights to channels with sharp saliency peaks) and combining this contribution from the saliency stream with the contribution from the affordance stream using the regions which have been obtained during the initial segmentation. While the saliency contribution is already in region-form, the grasping hypotheses (represented by the contact points) are projected from 3D into 2D. All points which fall into a certain region (a7)+(2) are summed and

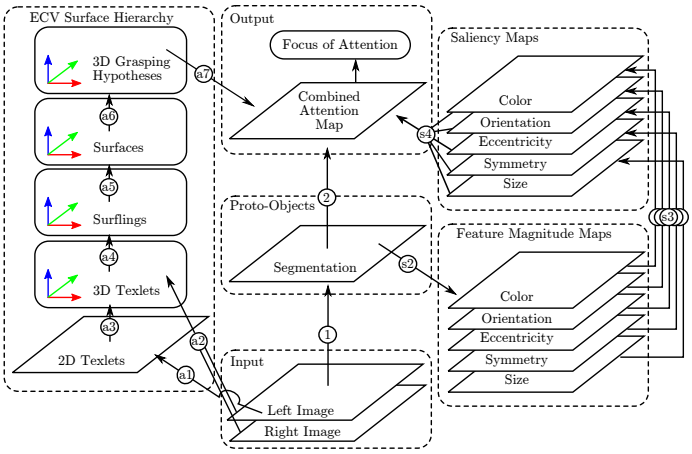

Fig. 4: Structure of the proposed model. The type and flow of data is described in the main text referring to this figure.

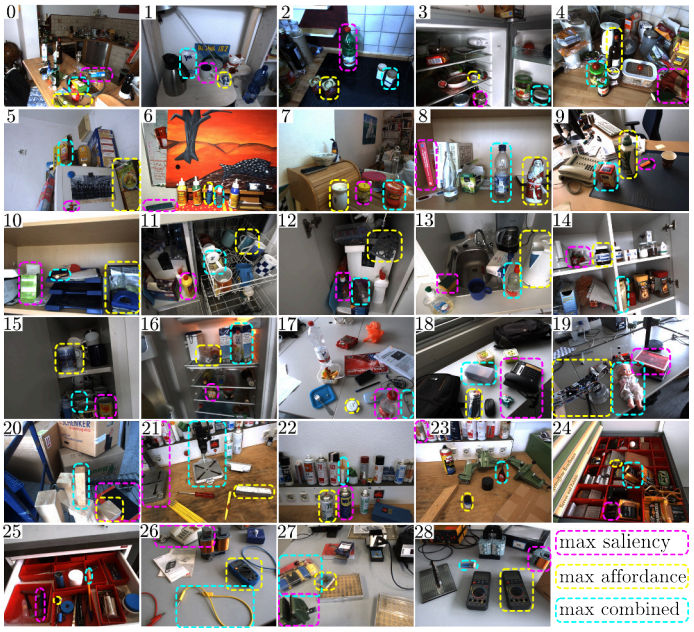

Fig. 5: Changed locations marked in the original images (best viewed magnified in the digital version).

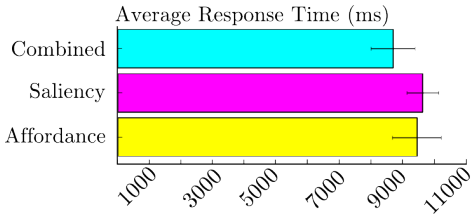

Fig. 6: Average response times for the changes based on affordance, saliency (region-based), and combined predictions (error bars show the SEM).

normalized by the region size to yield the affordance value of a region. Because the contact points can often be found on the edges of objects (in their 2D projection), instead of considering a single point in this process, each back-projected point is expanded to 5×5 points in a square region surrounding the initial location, who contribute decreasingly weighted with regard to their distance from the original location. This can be seen in figure 1b.

When the estimated affordance is combined with the saliency, this can be done in different ways. In principle, feature combination strategies as described in [10] can be applied. For instance, the map providing a small number with strong peaks could be preferred. Nevertheless, given that there is not yet any knowledge on how affordance and saliency are supposed to combine, we perform a naive combination. This is done by normalizing the values into the same dynamic range $(0\ldots1)$ and then calculating a weighted average of affordance and saliency for each region.

For the experiment reported in section III-B, equal weights are used to obtain the combined prediction. Weighting saliency one and affordance zero yields the pure saliency-based prediction while the reverse results in the purely affordance-based prediction.

B. Experiment II - Evaluation

This second experiment is based on predictions from the newly introduced model that combines affordance and saliency estimation. It is intended to investigate whether predictions based on combined saliency and affordance better reflect human attention than the individual components.

1) Participants: Thirty volunteers (average age of 28.46, $SD = 5.86$) participated.

2) Stimuli, design and procedure: Stimulus material, experimental design and procedure mostly correspond to the description of the first experiment in section II-A. The only differences were the use of a new image set (see figure 5) with an additional third possible change based on the combined prediction with equally weighted affordance and saliency contributions. Furthermore, the saliency-based prediction was obtained with the region-based saliency model [1], which constitutes the saliency channel in the combined model, in contrast to the first experiment where the model by Itti and colleagues [10] was used.

Due to the fact that three predictions (affordance, saliency, combined) are required for each image, and the images focus mainly on the action space where saliency and affordance are both expected to be relatively high, sometimes the same object was selected by two or all three predictions. In such a case, the scene was slightly rearranged by unsystematically shifting objects or the camera system, and the scene was rerecorded, until an image with three distinct predictions was obtained.

3) Results and Discussion: Figure 6 shows the average response times to changes based on the (region-based) saliency, affordance, and combined predictions. According to an one-way repeated measures ANOVA, no effect of prediction type was found, $F(2, 29) = 0.63$, $p = 0.54$. This is in contrast to the result of our first experiment, where the responses to affordance-based changes were significantly faster. Furthermore, the saliency conditions (from experiment 1 and experiment 2), as well as the affordance conditions (from each experiment), differ significantly, $t(40) = 5.82$, $p < 0.001$ (affordance), $t(40) = 4.3$, $p < 0.001$ (saliency) according to Holm-Bonferroni corrected two-sided t-tests.

The long response times in the second experiment indicate that the task was more difficult than in the first experiment. Moreover, the scenes were arranged to contain a large number of affording objects in the action space and thus also the saliency-based selections were mainly such foreground objects, whereas the stimulus material used in the first experiment contained saliency-based changes which were frequently in the background. In addition to the different scene arrangement, the region-based saliency model may less often select locations in the background where the contrasts are low, because such parts are often merged into larger background regions with low saliency (see figure 1d).

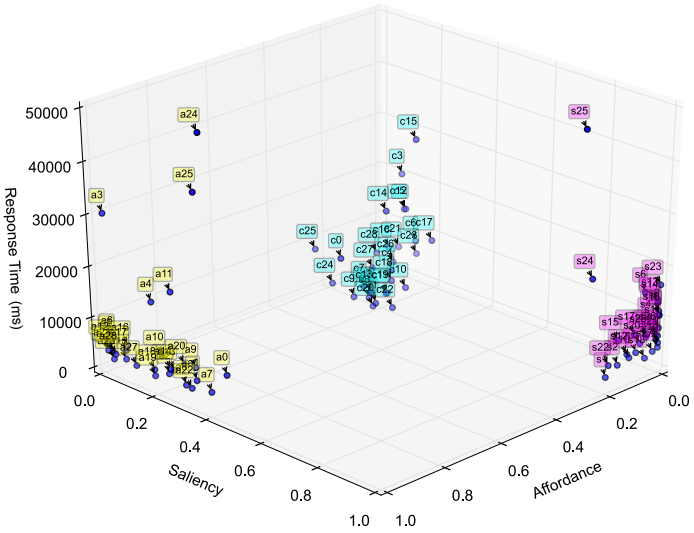

Fig. 7: Distribution of average response times (in milliseconds) per image and change, visualized as in figure 3. The numbers correspond to figure 5

.

Inspecting the distribution the individual changes' average response times in the affordance–saliency space (figure 7), it can be seen that while the saliency-based changes benefit from increasing affordance (response times decrease along the affordance axis), the same is not the case for the affordance-based changes which do not show a similar pattern. The changes based on the combined prediction cluster at lower saliency and affordance values (which were obtained by averaging the respective maps under the bounding rectangle of the removed combined object). Those, which are particularly low, resulted in longer response times. Notably, the one minute limit was reached twelve times for affordance- and nine times for saliency-based changes, and only once in the combined condition.

IV. Discussion and Conclusion

Taken together, the results of the first and second experiment further support the idea that object affordances are important for the spatial deployment of visual attention. In the second experiment we did not find additional enhancements by combining saliency and affordance. This is in line with another change blindness study [19], where saliency did not further enhance the detection of changes in objects which are shown in unusual contexts. Early attention appears to be strongly influenced by the environment represented in the scene. Future models that combine saliency and affordance could therefore employ strategies strongly biased towards affordances when these are available in the scene and fall back to saliency only when no other information is available.

The second experiment did not replicate the advantage of the affordance-based predictions over the saliency-based predictions we found in the first experiment using the stimulus material from [22]. This can originate from the use of a different saliency model, from the use of a different set of stimuli, or a combination of both. The region-based saliency model is not expected to yield predictions so much different in general (see [21] for comparisons with the model from [10]). However, as explained in section III-B3, the arrangements in the test scenes differ from those used in [22] in important aspects: The scenes used in [21] contained salient—but ungraspable—objects in the background, which were frequently selected by the saliency model. The scenes used in experiment 2 focus on objects in the action space, as the goal was to provide a variety of action space objects, which can exhibit different saliencies. Several scenes, especially scenes 24 and 25 (see figure 5), contain a large amount of competing targets resulting in extreme response times for saliency as well as affordance changes (see figure 7). Hence, this stimulus material may not be sensitive enough to reveal differences which may be small for these rather similar action space objects.

An important next step in this line of research is a deeper integration of affordance in attention systems. In the present study, affordance and saliency were estimated based on separate low-level scene representations (a 3D feature hierarchy and simple image regions). In future work, attention could be integrated and guide the creation of a 3D feature hierarchy. Alternatively, affordances could be estimated locally in 2D, as suggested in [17], for integration with classic attention models that are based on 2D retinotopical maps.

Our approach focuses on estimating the affordance of scene elements with respect to actions the system is able to perform towards these elements. This affordance-based channel can be integrated with existing region-based saliency channels (as shown in the present study) and top-down template mechanisms [e.g., 23]. Following a different approach, work by Varadarajan & Vincze [26] includes further aspects of environmental psychology, such as the affordance relations between the represented objects, in an attention model. A form of semantic saliency is directly derived from such affordances. The procedure presented in [26] obtains saliency from affordance aberrations, a measure of how unusual an object appears in its local semantic context. In addition to successful application in artificial vision, such models well predict the findings of [19] and others, which indicate that semantic scene context outranks pure stimulus driven saliency in early human vision.

With regard to our specific approach, the implementation of this attention system based on grasp affordances in a robot with an articulated arm and gripper will allow a practical evaluation of the proposed concepts.

References

[1] M. Z. Aziz and B. Mertsching. Fast and Robust Generation of Feature Maps for Region-Based Visual Attention. 17(5):633–644, 2008.

[2] M. Z. Aziz, M. S. Shafik, B. Mertsching, and A. Munir. Color Segmentation for Visual Attention of Mobile Robots. In *IEEE Symposium on Emerging Technologies*, pages 115–120, 2005.

[3] M. Backer, J. Tünnermann, and B. Mertsching. Parallel k-Means Image Segmentation Using Sort, Scan and Connected Components on a GPU. In R Keller, D Kramer, and J-P Weiss, editors, *Facing the Multicore-Challenge III*, pages 108–120. 2013.

[4] C. Castellini, T. Tommasi, N. Noceti, F. Odone, and B. Caputo. Using Object Affordances to Improve Object Recognition. *IEEE Transactions on Autonomous Mental Development*, 3(3):207–215, 2011.

[5] R. Detry, D. Kraft, O. Kroemer, L. Bodenhagen, J. Peters, N. Krüger, and J. Piater. Learning Grasp Affordance Densities. *Paladyn*, 2(1):1–17, 2011.

[6] P. Garrido-Vásquez and A. Schubö. Modulation of Visual Attention by Object Affordance. *Frontiers in Psychology*, 5:59, 2014.

[7] J. J. Gibson. *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. 1977.

[8] A. Gijsberts, T. Tommasi, G. Metta, and B. Caputo. Object Recognition Using Visuo-Affordance Maps. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1572–1578, 2010.

[9] T. C. Handy, S. T. Grafton, N. M. Shroff, S. Ketay, and M. S. Gazzaniga. Graspable Objects Grab Attention When the Potential for Action is Recognized. *Nature Neuroscience*, 1:1–7, 2003.

[10] L. Itti and C. Koch. Feature Combination Strategies for Saliency-Based Visual Attention Systems. *Journal of Electronic Imaging*, 10(1):161–169, 2001.

[11] V. Navalpakkam and L. Itti. A Goal Oriented Attention Guidance Model. In *Biologically Motivated Computer Vision*, pages 453–461, 2002.

[12] A. Oliva and A. Torralba. Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Progress in Brain Research*, 155:23–36, 2006.

[13] M. Popović, G. Kootstra, J. A. Jørgensen, D. Kragic, and N. Krüger. Grasping Unknown Objects Using an Early Cognitive Vision System for General Scene Understanding. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011.

[14] R. A. Rensink, J. K. O'Regan, and J. J. Clark. To See or Not to See: The Need For Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5):368–373, 1997.

[15] E. Rivlin, S.J. Dickinson, and A. Rosenfeld. Recognition by Functional Parts [Function-Based Object Recognition]. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 267–274, 1994.

[16] K. L. Roberts and G. W. Humphreys. Action-Related Objects Influence the Distribution of Visuospatial Attention. *Quarterly Journal of Experimental Psychology*, 64 (4):669–88, 2011.

[17] H. O. Song, M. Fritz, C. Gu, and T. Darrell. Visual Grasp Affordances from Appearance-Based Cues. In *IEEE ICCV Workshops*, pages 998–1005, 2011.

[18] L. Stark and K. Bowyer. Generic Recognition Through Qualitative Reasoning About 3-D Shape and Object Function. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–256, 1991.

[19] J. A. Stirk and G. Underwood. Low-Level Visual Saliency Does not Predict Change Detection in Natural Scenes. *Journal of Vision*, 7:1–10, 2007.

[20] P. Tseng, J. Tünnermann, N. Roker-Knight, D. Winter, I. Scharlau, and B. Bridgeman. Enhancing Implicit Change Detection Through Action. *Perception*, 39(10): 1311–1321, 2010.

[21] J. Tünnermann and B. Mertsching. Region-Based Artificial Visual Attention in Space and Time. *Cognitive Computation*, 6(1):125–143, 2014.

[22] J. Tünnermann, N. Krüger, B. Mertsching, and W. Mustafa. Affordance Estimation Enhances Artificial Visual Attention: Evidence from a Change Blindness Study. (in review).

[23] J. Tünnermann, C. Born, and B. Mertsching. Top-Down Visual Attention with Complex Templates. In *International Conference on Computer Vision Theory and Applications*, pages 370 – 377, 2013.

[24] K. M. Varadarajan and M. Vincze. Affordance Based Part Recognition for Grasping and Manipulation. In *ICRA Workshop on Autonomous Grasping*, 2011.

[25] K. M. Varadarajan and M. Vincze. Object Part Segmentation and Classification in Range Images for Grasping. In *15th International Conference on Advanced Robotics*, pages 21–27, 2011.

[26] K. M. Varadarajan and M. Vincze. Semantic Saliency Using k-TR Theory of Visual Perception. In *21st International Conference on Pattern Recognition*, pages 3676–3679, 2012.

[27] B. Yao, J. Ma, and L. Fei-Fei. Discovering Object Functionality. In *IEEE International Conference on Computer Vision*, pages 2512–2519, 2013.

[28] Y. Zhao and S.-C. Zhu. Scene Parsing by Integrating Function, Geometry and Appearance Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3119–3126, 2013.