Integrating Object Affordances with Artificial Visual Attention

Jan Tünnermann, Christian Born & Bärbel Mertsching

GET Lab, University of Paderborn
Pohlweg 47–49, 33098 Paderborn, Germany
{tuennermann, born, mertsching}@get.uni-paderborn.de

Abstract. Affordances, e.g., grasping possibilities, play a role in the guidance of human attention. We report experiments on the integration of affordance estimation with artificial visual attention in a prototypical model. Furthermore, Growing Neural Gases are discussed as a potential framework for future attention models that deeply integrate affordance, saliency and further attentional mechanisms.

Keywords: Attention, Saliency, Affordance

1 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 surroundings. A popular concept is the *affordance of objects*, introduced by J. J. Gibson in 1977 [1]. In this holistic view, objects possess 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, not only to enhance grasping actions, but also to benefit object recognition and semantic scene perception (see e.g., [2–4]). In many cases, objects are better defined by actions the object supports, than by 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 [5]. Therefore, recent research integrates affordance estimation with object recognition [6, 5] and the semantic interpretation of scenes and objects [7, 8].

Artificial visual attention is a concept inspired by cognitive psychology. 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 feature contrasts) [9] or in a top-down manner by incorporating knowledge, task demands [10] or the "gist of the scene" [11].

Findings from psychology suggest that affordances influence human visual attention. This has been shown in reductions of reaction times when affordances

were used to guide attention towards target locations [12] and effects on event related signals in electrophysiological and brain imaging research [13].

This design paper is an update of the report we presented at the First Workshop on Affordances: Affordances in Vision for Cognitive Robotics [14]. The remainder of the paper is organized as follows: Section 2 contains a compressed report of the experiments conducted in [14]. In section 3 we discuss Growing Neural Gases as a framework for artificial attention that we believe has the potential to integrate bottom-up saliency, affordance-based attention and top-down mechanisms in a consistent architecture and improve on several disadvantages of current region-based attention systems. Section 4 concludes the paper.

2 Change Detection Experiments on Saliency and Affordance in Human Attention

In a previous study [15], we employed a "single-shot" change detection task with natural images (see e.g., [16]) to measure the participants' distribution of attention towards salient or affording objects. The phenomenon of change blindness due to short scene interruptions renders the detection of changes in objects difficult. An observer's performance depends on the allocation of attention towards the objects [17]. 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 an affordance-based model than when those selected by the saliency model were changed.

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.

Hence, in this experiment, we test the so called "flicker paradigm" (see e.g., [17]): 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 [15], 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.

2.1 Experiment 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.

Stimuli: The stimulus material reported in [15] 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. [9] had been removed, in the other altered image, 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 2.2) had been removed. Refer to [15], for more details regarding the stimulus generation and the actual pictures.

Design and Procedure: The experiment was conducted on a 12.1" touch-screen laptop[1]. Participants were presented with every original image paired with one of the possible changes. 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 [15], the images cycled back and forth between the original and changed image until the change was reported 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". For every trial, the response time was recorded.

Results and Discussion: Figure 1 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 [15] 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.

In [14] we show response time distributions over the different changes and images. No long response times are found when saliency and affordance of a change were high. Especially for the saliency-based changes, long response times

[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.

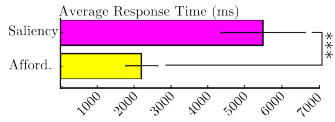

Fig. 1. Average response times for affordance- and saliency-based changes (error bars show the SEM; *** $p < 0.001$).

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.

2.2 Experiment 2

This second experiment is based on predictions from a prototypical 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.

A Combined Model of Saliency and Affordance: The model is outline in figure 2. The left image of a stereo image pair is segmented into homogeneously colored regions ① (see e.g., [18]). These regions can be considered proto-objects at pre-attentional stages.

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.

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 for the feature and saliency computations can be found in [19]).

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 transformed into *3D Texlets* (a3). Small groups of neighboring *Texlets* are created by applying a position-based k-means clustering. Planes are fitted through the *groups* to form *Surflings* (a4), which are further grouped (when close to each other and similarly oriented) to generate *Surfaces* (a5) [20].

Grasping hypotheses in 3D space are generated by fitting a simulated gripper (see figure 3a) to elements of the scene considering the surfaces generated in the process outlined above (a6). Details of this process can be found in [20]. The result, 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.

The combined attention map is then obtained by integrating the individual saliency maps (s4) [19] and the contribution from the affordance stream using the regions from the initial segmentation. While the saliency contribution is already in region-form, the grasping hypotheses (contact points) have to be projected

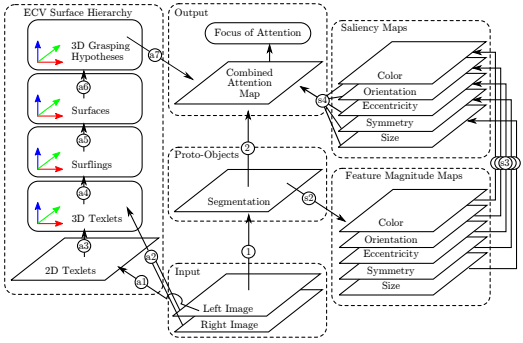

Fig. 2. Structure of the proposed model. The type and flow of data is described in the main text referring to this figure.

into 2D first. All points which fall into a certain region (a7)+(2) are summed and normalized by the region size. 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, with their contribute decreasing with distance from the original location. This can be seen in figure 3b. In this first attempt to combine saliency and affordance in a technical model, we combine both linearly with equal weights. More advanced strategies to combine different feature channels in attention models are discussed in [9].

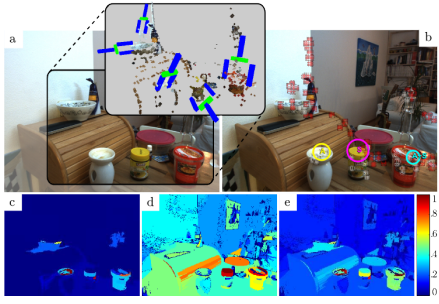

Fig. 3. a: A test scene. Inset: exemplary grasps fitted to a sparse 3D representation. **b:** Grasp points projected into 2D. White patterns represent grasps towards reachable locations, red patterns indicate locations out of reach. "A" affordance, "S" saliency, and "A+S" combined selection. **c–e:**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

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

Stimuli, design and procedure: Stimulus material, experimental design and procedure mostly correspond to the description of the first experiment in section 2.1. The only differences were the use of a new image set (see figure 4a) with an additional third possible change based on the combined prediction. Furthermore, the saliency-based prediction was obtained with the region-based saliency model [19], which constitutes the saliency channel in the combined model, in contrast to the first experiment where the model by Itti and colleagues [9] 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, and the scene was rerecorded, until three distinct predictions were obtained.

Fig. 4. a: Change locations marked in nine exemplary images (all 29 images are shown in [14]). **b:** Average response times for the changes based on affordance, saliency (region-based), and combined predictions (error bars show the SEM).

Results and Discussion: Figure 4b 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. Inspection of the distribution of the individual changes' average response times (refer to [14]), hints that saliency-based changes may benefit from increasing affordance, while the same seems not to be the case for affordance-based changes.

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.

Hence, whether and how saliency and affordance enhance each other remains unclear. In the prototypical model, affordance and saliency have been processed based on different representations (ECV vs. region-based), normalized in different ways, and integrated eventually. Due to this process it becomes difficult to assess the relative contributions of both channels and relate them to response times (we attempted this to some degree in [14]). Furthermore, for future technical applications, calculating and maintaining to separate representations is rather unpractical. Therefore, in the remainder of the paper Growing Neural Gases as a potential structure for a fully integrated representation and attention model is discussed.

3 Growing Neural Gases: An Architecture for Combining Saliency, Top-Down Attention and Affordances?

In the long term, a fully integrated architecture for artificial attention is desirable. In such a framework, feedback loops, which are known to be highly im-

portant in biological vision, can be established: results from higher levels of the architecture, such as affordances or top-down information can be used to influence the generation or propagation of the scene representation from lower levels. Furthermore, the integration of different dimensions, such as various saliency dimensions, specific and gist-based top-down influences and object affordances benefits from a common consistent representation which can preserve the relative strength of each dimension's contribution (a prerequisite for more advanced combination strategies as suggested in [9]).

We discuss Growing Neural Gases (GNG; e.g. [21]) as pre-attentional structures for a fully integrated approach. GNGs have not been applied to artificial attention before but exhibit promising features. On the one hand, they can be seen as related to the already mentioned region-based approaches [19, 22, 23] as pre-attentional structures are employed. On the other hand, they implement a basic perceptual learning in the sense that the current representation is updated with new data instead of recalculating it entirely as in region-based artificial attention. We briefly describe the main concepts of GNGs and discuss its potential application to saliency, top-down and affordances calculations.

GNGs constitute an unsupervised learning technique. Nodes (neurons) may be connected by edges and possess attributes representing properties of the state space (e.g., x- and y-positions). Examples are presented to the algorithm and it determines the closest node according to a distance measure on the respective properties. For this node and, with a reduced strength its topological neighbors, the properties are updated. Edges carry an age value, which is increased in every update. A new edge is inserted between the closest node and the runner-up. If such an edge already exists, its age is reset. Edges which are too old are deleted. Furthermore, there is a domain dependent error value for each node, which must have the characteristic that it is reduced when a new node is inserted in proximity. At fixed intervals, the node with the highest error value is determined, as well as its neighbor with the highest error. A new node is inserted in between and connected with these two nodes, replacing their original connection. The error value is redistributed between all three nodes reducing the probability that the next insertion is performed nearby, guiding the growth of the network. An additional utility term quantifies a node's usefulness and may result in the deletion of the node to avoid infinite growth of the network.

The algorithm is initialized with two connected nodes with random properties. The dynamically adapting set of nodes with changing neighborhoods can form multiple independent graphs.

Pre-Attentional Structures and Saliency based on GNGs: GNGs can potentially be adapted to generate pre-attentional structures. Pixels of the image are chosen uniformly at random and used as examples to train a GNG as described above. Distances to the pixels can be calculated by using, e.g., a weighted euclidean distance. Figure 5a depicts the result for a simple synthetic test image. For the resulting graphs, feature magnitudes and saliency can be calculated as described for region-based saliency in [19].

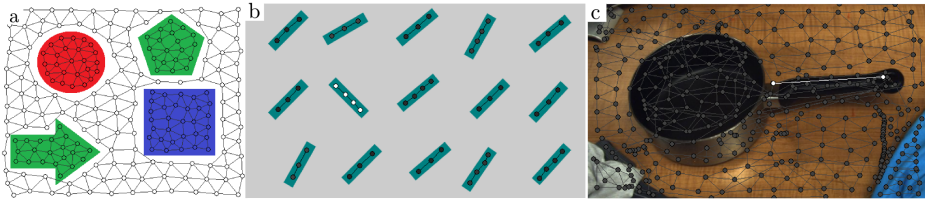

Fig. 5. a: Exemplary results using GNG-based pre-attentional structures. Node colors reflect the represented object. **b:** The result of a saliency computation. Gray levels of the nodes represent the orientation saliency of a graph (background graph removed in calculation). **c:** GNGs may include useful candidates (two-node networks; white) for estimating grasp affordances. Networks with more than two nodes are colored in gray.

The feature *orientation* is used as an example here. As in the aforementioned paper, 2D central moments are calculated for each node $n(x, y)$ in graph G_i:

$$m^i_{1,1} = \sum (x - \overline{x})(y - \overline{y}),\ m^i_{2,0} = \sum (x - \overline{x})^2 \text{ and } m^i_{0,2} = \sum (y - \overline{y})^2\ \forall n(x, y) \in G_i \tag{1}$$

where $(\overline{x}, \overline{y})$ denotes the center of G_i. The orientation ϕ^i is then computed as

$$\phi^i = \frac{1}{2} tan^{-1} \left(\frac{2m^i_{1,1}}{m^i_{2,0} - m^i_{0,2}} \right), \tag{2}$$

resulting in an orientation value ϕ^i between 0° and 180°. The orientation saliency s_{ϕ^i} of every graph G_i is then obtained as

$$s_{\phi^i} = \sum_{G_i, i \neq j} \frac{|\phi^i - \phi^j|}{90°}. \tag{3}$$

The result of this process is an orientation saliency map (saliency value associated with every graph) and shown for an orientation pop-out stimulus in figure 5b.

Applying Top-Down Information in GNGs: As argued above, mechanisms from region-based attention can be transferred to GNG structures. Therefore, templates could be used as described in [22, 24] for region-based attention.

Furthermore, as the transfer from pixels to the substantially smaller number of neurons provides a simplified problem space, rough heuristics, such as for determining the gist of a scene, may also benefit from a GNG-based representation.

Integrating Object Affordances in GNGs: Pre-attentional structures obtained from GNGs may prove sufficiently stable to apply appearance-based affordance estimation in local and global contexts as proposed by [25]. Furthermore, GNGs obtained with the described procedure may provide easily identifiable candidates for graspable elements. Figure 5c shows GNGs obtained from a picture of a cooking pot extracted from Song et al. [25]'s figure 4b (the pink dot was

removed). Highlighting only two-node networks, in agreement with the ground-truth for such handles (see Song et al. [25]'s figure 2a), successfully detects the pot's handle. Such rough heuristics could be directly useful for generating local graspability estimates which can then be fused with global estimates [25], or provide candidates for more expensive follow-up processing.

4 Conclusion

The results of our first experiment further supports the idea that object affordances are important for the spatial deployment of visual attention [23]. In the second experiment we did not find additional enhancements by combining saliency and affordance. This is in line with another change blindness study [26], 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. The second experiment, however, did also fail to replicate the advantage of the affordance-based predictions over the saliency-based predictions. This may arise from limitations of the prototypical model (see section 2.2). Alternative possibilities are discussed in [14].

Hence, an important next step in this line of research is a deeper integration of affordance in attention systems. The fact that affordance-based advantages are present in 2D images presented to humans, which depict foregrounds and background (experiment 1 and experiments reported in [15]), proves that binocular cues are not necessary for the effect in biological vision. Thus, a 2D dimensional retinotopical structure would provide a useful domain for such a fully integrated approach. We discussed Growing Neural Gases as a framework for this in section 3. These may allow to integrate appearance based affordance-estimation as suggested by Song et al. [25] with bottom-up and top-down attention in future work to allow more sensitive experiments and practical evaluation in a robot.

References

1. Gibson, J.J.: The Theory of Affordances. In Shaw, R., Bransford, J., eds.: Perceiving, Acting, and Knowing. (1977) 67–82
2. Stark, L., Bowyer, K.: Generic Recognition Through Qualitative Reasoning About 3-D Shape and Object Function. In: IEEE Conference on Computer Vision and Pattern Recognition. (1991) 251–256
3. Detry, R., Kraft, D., Kroemer, O., Bodenhagen, L., Peters, J., Krüger, N., Piater, J.: Learning Grasp Affordance Densities. Paladyn **2**(1) (2011) 1–17
4. Varadarajan, K.M., Vincze, M.: Affordance Based Part Recognition for Grasping and Manipulation. In: ICRA Workshop on Autonomous Grasping. (2011)
5. Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B.: Using Object Affordances to Improve Object Recognition. IEEE Transactions on Autonomous Mental Development **3**(3) (2011) 207–215
6. Gijsberts, A., Tommasi, T., Metta, G., Caputo, B.: Object Recognition Using Visuo-Affordance Maps. IEEE/RSJ International Conference on Intelligent Robots and Systems (2010) 1572–1578

7. Yao, B., Ma, J., Fei-Fei, L.: Discovering Object Functionality. In: IEEE International Conference on Computer Vision. (2013) 2512–2519
8. Zhao, Y., Zhu, S.C.: Scene Parsing by Integrating Function, Geometry and Appearance Models. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3119–3126
9. Itti, L., Koch, C.: Feature Combination Strategies for Saliency-Based Visual Attention Systems. Journal of Electronic Imaging **10**(1) (2001) 161–169
10. Navalpakkam, V., Itti, L.: A Goal Oriented Attention Guidance Model. In: Biologically Motivated Computer Vision. (2002) 453–461
11. Oliva, A., Torralba, A.: Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research **155** (2006) 23–36
12. Garrido-Vásquez, P., Schubö, A.: Modulation of Visual Attention by Object Affordance. Frontiers in Psychology **5** (2014) 59
13. Handy, T.C., Grafton, S.T., Shroff, N.M., Ketay, S., Gazzaniga, M.S.: Graspable Objects Grab Attention When the Potential for Action is Recognized. Nature Neuroscience **1** (2003) 1–7
14. Tünnermann, J., Mertsching, B.: Saliency and Affordance in Artificial Visual Attention. In: RSS 2014, First Workshop on Affordances: Affordances in Vision for Cognitive Robotics. (2014)
15. Tünnermann, J., Krüger, N., Mertsching, B., Mustafa, W.: Affordance Estimation Enhances Artificial Visual Attention: Evidence from a Change Blindness Study. (in review)
16. Tseng, P., Tünnermann, J., Roker-Knight, N., Winter, D., Scharlau, I., Bridgeman, B.: Enhancing Implicit Change Detection Through Action. Perception **39**(10) (2010) 1311–1321
17. Rensink, R.A., O'Regan, J.K., Clark, J.J.: To See or Not to See: The Need For Attention to Perceive Changes in Scenes. Psychological Science **8**(5) (1997) 368–373
18. Backer, M., Tünnermann, J., Mertsching, B.: Parallel k-Means Image Segmentation Using Sort, Scan and Connected Components on a GPU. In Keller, R., Kramer, D., Weiss, J.P., eds.: Facing the Multicore-Challenge III. (2013) 108–120
19. Aziz, M.Z., Mertsching, B.: Fast and Robust Generation of Feature Maps for Region-Based Visual Attention. **17**(5) (2008) 633–644
20. Popović, M., Kootstra, G., Jørgensen, J.A., Kragic, D., Krüger, N.: Grasping Unknown Objects Using an Early Cognitive Vision System for General Scene Understanding. IEEE/RSJ International Conference on Intelligent Robots and Systems (2011) 987–994
21. Fritzke, B.: A Self-Organizing Network That can Follow Non-Stationary Distributions. In: Artificial Neural Networks – ICANN'97. (1997) 613–618
22. Aziz, M.Z., Mertsching, B.: Visual Search in Static and Dynamic Scenes Using Fine-Grain Top-Down Visual Attention. In: Computer Vision Systems. (2008) 3–12
23. Tünnermann, J., Mertsching, B.: Region-Based Artificial Visual Attention in Space and Time. Cognitive Computation **6**(1) (2014) 125–143
24. Tünnermann, J., Born, C., Mertsching, B.: Top-Down Visual Attention with Complex Templates. In: International Conference on Computer Vision Theory and Applications. (2013) 370 – 377
25. Song, H.O., Fritz, M., Gu, C., Darrell, T.: Visual Grasp Affordances from Appearance-Based Cues. In: IEEE ICCV Workshops. (2011) 998–1005
26. Stirk, J.A., Underwood, G.: Low-Level Visual Saliency Does not Predict Change Detection in Natural Scenes. Journal of Vision **7** (2007) 1–10