Exploration Strategies for Incremental Learning of Object-Based Visual Saliency

Céline Craye, David Filliat, Jean-François Goudou

Exploration Strategies for Incremental Learning of Object-Based Visual Saliency

Céline Craye*†, David Filliat* and Jean-François Goudou†
*ENSTA Paristech - INRIA FLOWERS team,
Unité informatique et Ingénierie des Systèmes
828 boulevard des Maréchaux, 91762 Palaiseau, France
†Thales - SIX - Theresis - Vision & Sensing
1, avenue Augustin Fresnel, 91767 Palaiseau, France
Email: celine.craye@thalesgroup.com, celine.craye@ensta-paristech.fr

***Abstract*—Searching for objects in an indoor environment can be drastically improved if a task-specific visual saliency is available. We describe a method to learn such an object-based visual saliency in an intrinsically motivated way using an environment exploration mechanism. We first define saliency in a geometrical manner and use this definition to discover salient elements given an attentive but costly observation of the environment. These elements are used to train a fast classifier that predicts salient objects given large-scale visual features. In order to get a better and faster learning, we use intrinsic motivation to drive our observation selection, based on uncertainty and novelty detection. Our approach has been tested on RGB-D images, is real-time, and outperforms several state-of-the-art methods in the case of indoor object detection.**

I. Introduction

Visual exploration, object discovery or identification in cluttered environments by mobile robots is still an open problem. Machine learning, and especially deep learning has recently shown impressive results on complex datasets such as IMAGENET [18], but learning is offline, fully supervised, and may not be flexible to new environments or dynamics. Another approach is to learn and discover an environment directly on a robot, in an incremental and autonomous way. Learning is then specialized for a specific environment, but is constantly improved and remains flexible to any change or novelty. In addition, actions of the robot can be guided by intrinsic motivation that will allow better and faster learning, focusing first on simple tasks, and increasing difficulty as learning progresses.

In robotics, visual exploration of the environment is often associated with a visual attention strategy [2], [6]. Thus, the robot's attention is directed towards areas of interest, and irrelevant portions of the visual field are not considered. Visual attention can be driven by purely bottom-up saliency maps [7], [13], [17], [21], or refined by top-down modulation [9], [22]. Bottom-up saliency highlights stimuli that are intrinsically salient in their context, which may sometimes be sufficient for scene exploration [23]. However, top-down modulation, which highlights elements that are relevant for a specific task, is more meaningful for the problem of object detection in indoor environments.

Inspired by human vision, the foveal vision principle is commonly used in robotics for object identification tasks. This principle is naturally consistent with selective visual attention, where target selection is determined in the wide field of view, and focus is analyzed with foveal view. The selection of target is often obtained with a saliency map, but also by novelty detection [12], reinforcement learning techniques [15], or in very simple configuration, competence progress [1].

So far, visual attention in robotics is used as a target selection system for another task. Saliency maps estimation is used as a black box and is not improved during exploration. Yet, learning visual attention directly on the robots would make sense in a developmental perspective. It was found that visual attention significantly varies at different ages [11], but also between cultures. In [3], Boduroglu *et al.* suggest that Eastern and Asian disparities in visual attention allocation may be related to differences in their physical environment. This means that visual attention is learned and modulated by our environment. To our knowledge, no method has ever been presented to learn visual attention incrementally directly on a robotic platform.

We aim to design an algorithm that can learn saliency in a incremental and intrinsically motivated manner, directly within a robot's environment. We restrict the problem to indoor environment and we define saliency as related with objects. For that, we use a mechanism of foveal observation, providing reliable and accurate information in a small portion of the visual field, and a contextual observation providing low resolution information in a wider portion of the field of view. We use foveal vision to reliably identify salient elements, then to learn their visual features on the contextual view. We select the foveal view based on intrinsically motivated criteria to speed up and improve the learning quality. After a learning period, salient elements can be detected directly on the contextual view and used for other tasks.

The article is organized as follows: Section II describes the different components of our method. In section III, we present experimental results and comparison with state-of-the-art. Lastly, we provide concluding remarks and future work directions.

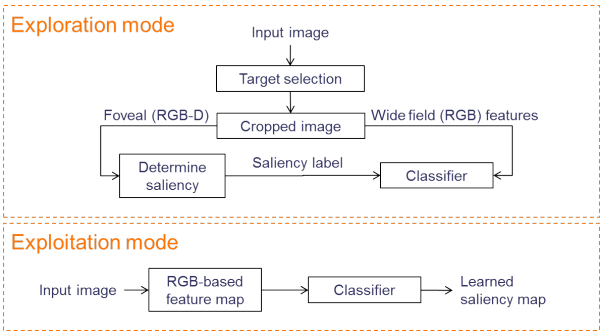

Fig. 1. General architecture of our system

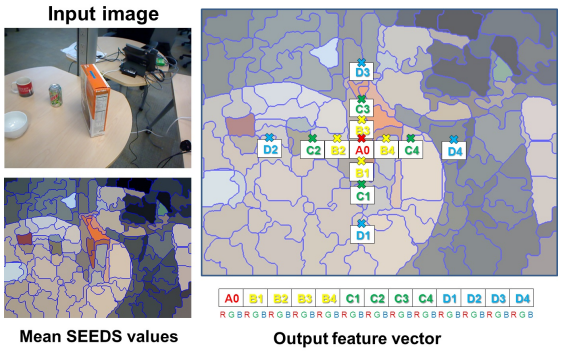

Fig. 2. RGB-based feature extraction

II. Proposed approach

In this section, we describe our incremental learning mechanism of visual saliency. The algorithm is designed for indoor environments with many objects.

A. General architecture

Figure 1 presents the general architecture of our system. In an exploration stage, the system is trained and learns the visual aspect of salient elements within their context. For that, we use the foveal vision to accurately determine whether the observed area (i.e. a small portion of the visual field) is salient or not, and we learn the corresponding visual features in the wide field of view. In a second stage, we exploit the model to generate saliency maps dedicated to the environment that has been previously explored.

In the exploration stage, we first select an area of the input image in which we obtain foveal observation. This selection can be either random or intrinsically motivated. In a biological eye, the foveal observation has a higher resolution than the contextual one. In our method, instead of a higher resolution, the foveal observation is augmented with a depth component. Therefore, we get access to RGB-D data in the fovea, and only RGB in the whole field of view. Using RGB-D data is convenient for evaluation and comparison with state-of-the-art, as a large number of datasets are available. Nevertheless, other types of foveal information could be used, for example a higher-resolution image, stereo-vision or motion analysis.

The foveal RGB-D information is then used to determine whether the selected area is salient or not. For that, we check if the area is consistent with the geometrical criterion, accessible from depth information. If consistent, the area is considered as salient and sent to the classifier with label *salient* (or *not salient* otherwise). This geometrical saliency is computationally expensive and could not run in real time if applied on the whole field of view (See Section II-C).

Lastly, the classifier is trained with RGB-based features that are fast to compute. The label determined by the depth information is associated with the input features, thus enabling saliency learning (More details in Section II-D).

B. RGB-based visual features

We want to use visual features based on the RGB components to reconstruct the saliency map. Those features must be light enough to be computed in real time on the whole field of view, and be representative enough to discriminate between salient and non salient elements. Lastly, the features must be stable enough to avoid saliency misdetection, especially in areas of strong or sharp color variation. For that, we use features inspired by Make3D [19], as these type of features are used to estimate depth from RGB data. Our features are obtained as follows: first, we obtain 150 SEEDS superpixels [20] on the whole image, and we compute the mean RGB value in each of them. Then, for each pixel of the image, we construct the feature vector based on the means superpixel value of the current pixel, and the means superpixel values of the neighbors at three different scales. Figure 2 shows the feature extraction process for one pixel. The three different scales are 25,50 and 100 pixels from the central pixel, thus leading to 39 features. Using the superpixel mean value instead of the pixel value itself makes the feature vector more robust to noise and sharp variations in the image. The RGB feature vector is used both in the learning stage for classification, and in the exploration stage for saliency map reconstruction.

C. Salient elements discovery

In our case, salient elements are defined as objects of limited size (up to 30 centimeters) lying on plane surfaces (typically tables or floor). This definition is consistent with a large nunber of indoor objects in static environments. However, this could be adapted to other object range. More generally, the way to determine saliency can be easily modified without modifying the general architecture of the system.

First of all, we must select the portion of the visual field where salient element geometrical checking will be done. We call this area the foveal view, and we restrict the use of depth information to this area. The choice of the foveal target is described in section II-E.

After selecting our foveal target, we check the depth of the central pixel and we adjust the size of the foveal view so that a 30 centimeters object can be entirely seen. We then select the SEEDS superpixels computed in Section II-B falling in this area. For each superpixel, we create a point cloud based on the depth map, and we use RANSAC [8] algorithm to find the major plane equation of the superpixel. We then find

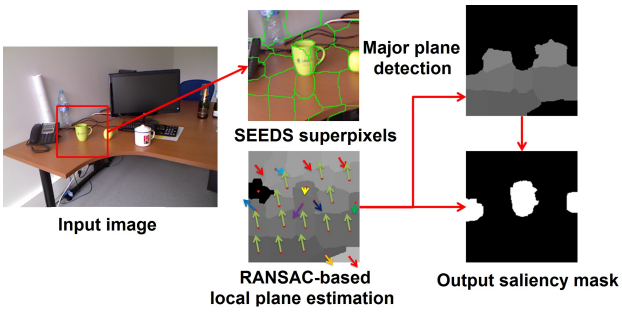

Fig. 3. Main steps of salient element discovery.

the global major plane of the foveal area based on the most frequent local (superpixel-based) plane. Lastly, we check if this plane has objects lying on it by finding sets of pixels whose distance to the major plane is less than 30 cm and that are not occluded by the plane. With this definition, walls are sometimes detected as salient if they are close enough to the table. Therefore, we add the constraint that plane elements that are in contact with the border of the foveal view are not salient. Figure 3 summarizes the main steps of the algorithm.

D. Saliency learning

Salient element discovery is made possible by checking the geometrical consistency in the fovea, which is a slow and expensive process (see section III-C) . We now want to use the information extracted from this step, and reconstruct saliency from a much lighter processing. For that, we use the 39 visual features described in II-B and a classifier that will incrementally learn the saliency provided by each foveal observation.

Each new foveal observation produces a segmentation of salient and non salient pixels within the observed area. Given this segmentation, we produce a set of samples that will be used in our learning process. First, we cut the observed area into small squares of twenty pixels (we found this value to be a good trade-off between speed and accuracy), and we determine a *salient* or *not salient* label for each of them, depending whether or not the square has a majority of *salient* pixels. Then, we determine for each square the feature vector of the central pixel. For each square, we then produce a features-label vector that we send to the classifier for learning. At each new observation, a new set of samples is added to the training dataset and the classifier is retrained.

Our classifier is a random forest [4] of 20 trees with a maximum depth of 10 per tree. This choice was motivated by the good performance of these classifiers, their fast training computation, their good generalization capacity and their ability to efficiently handle unbalanced data (In practice, we discover much more *not salient* elements).

In the exploitation stage, the saliency map is constructed as follows: we compute the RGB features on the whole input image, then we cut the image in squares of 20 pixels. We compute the center pixel feature vector, and we send it to the classifier. For each square, we obtain a score from the classifier, estimating the saliency of the square. We associate the classifier score to each square of the input image.

E. Intrinsically motivated target selection

Target selection of the foveal view has a critical impact in the learning quality and efficiency. Selecting a foveal view driven by intrinsic motivation should enable a more efficient learning than pure random selection, by focusing for example on elements that are still unknown or harder to learn. We propose two simple criteria to select the foveal area, thus enabling a better learning of the saliency.

A first strategy is to choose the foveal target so as to reduce uncertainty. For each new frame and before selecting the foveal view, we construct a saliency map based on the classifier state. Each pixel of the estimated saliency map (sal) reprensents the probability to be *salient* given corresponding features f:

$$sal(x,y) = Pr(salient|f(x,y)) = 1 - Pr(\overline{salient}|f(x,y)) \quad (1)$$

We select the target in the region where uncertainty is maximum, or in other words, where probability to be *salient* is neither high nor low, the most uncertain probability being 0.5. To determine in practice the most uncertain area, we randomly select 20 potential targets in the whole visual field. For each target, we determine the associated foveal area, and we calculate in this area the average uncertainty value, provided by equation 2:

$$Unc(x,y) = \sum_{i,j\in[-S..S]} -|sal(x+j,y+i) - 0.5| \quad (2)$$

where S is half the size of the fovea. The saliency discovery process is then applied in the area with highest Unc value among the 20 candidates.

A second strategy is to direct the foveal view preferably to novel stimuli. We use an improved version of the proximity measure suggested by Breiman *et al.* [4], based on random forests. The proximity of samples s and p is defined as the number of common terminal nodes in all trees, weighted by the number of elements in the common terminal nodes:

$$prox(s,p) = \sum_{i=1}^{nTree} \frac{\delta(l_s^i, l_p^i)}{card(l_p^i)} \quad (3)$$

where $nTree$ is the number of trees used in the random forest, δ is the Dirac function, l_s^i is the terminal node index of sample s for tree i, and $card(x)$ is the number of elements in terminal node x. Novelty of a sample is caracterized by the proximity with the nearest neighbor in the training set. If the proximity is low, then the sample is far from the training set and the novelty is high. Novelty is therefore obtained by equation 4:

$$Nov(s) = 1 - \frac{1}{N(s)} \max_{p\in TrSet} prox(s,p) \quad (4)$$

where $TrSet$ is the training set and $\frac{1}{N(s)} = \sum_{i=1}^{nTree} card(l_s^i)$ is a normalization factor. We proceed to target selection based on novelty exactly the same way as uncertainty, except that we replace the term in the sum of equation 2 by $Nov(s(x,y))$.

To avoid blocking situation, we adopt an epsilon-greedy strategy by selecting the target randomly 30% of the time.

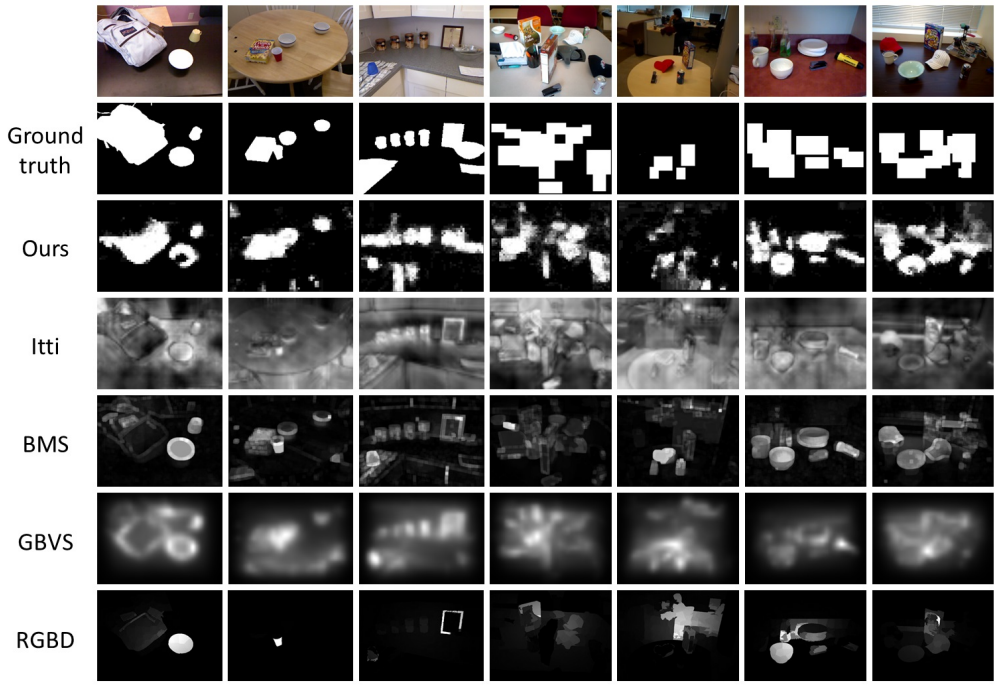

Fig. 4. Sample saliency maps for the five evaluated methods.

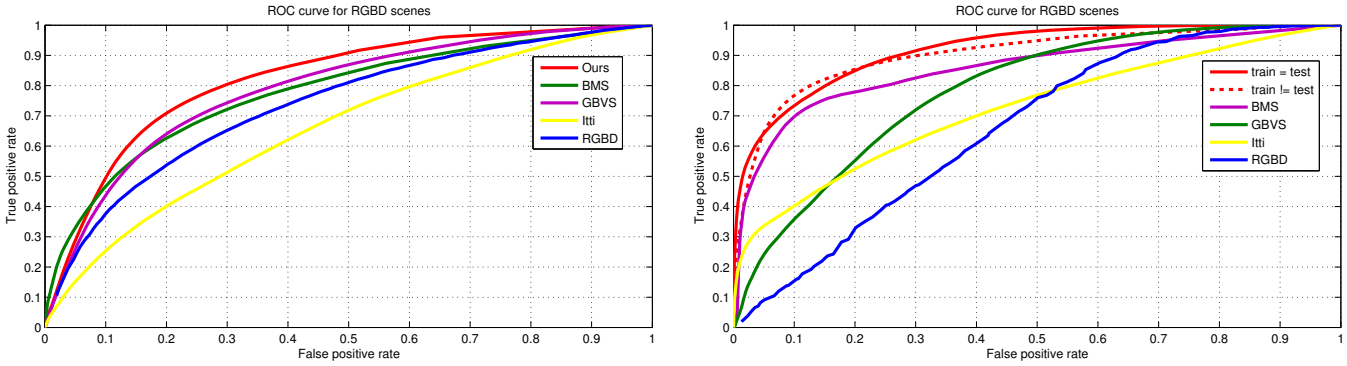

Fig. 5. ROC curves for five different approaches. (a) Training a dedicated classifier for each image of the dataset. (b) On *table-small-2* sequence, comparison between"train==test" and "train!=test" modes.

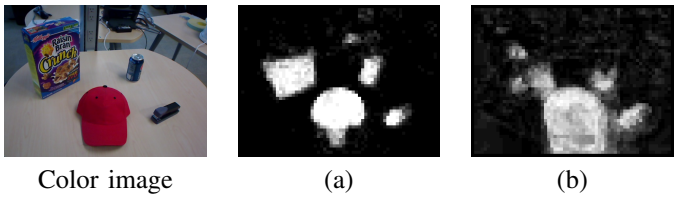

Fig. 6. Saliency maps in (a) Train = test and (b) Train $\neq$ test modes. Although (b) has more noise, learning saliency based on various points of view can recover salient areas that are not geometrically consistent in the current image.

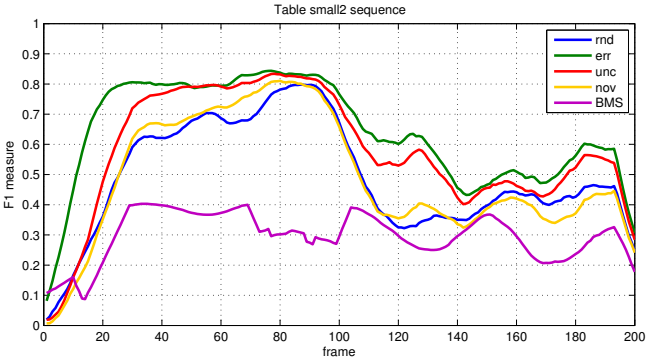

Fig. 7. F1 measure for several target selection criteria.

III. Experimental results

A. Comparison to state-of-the-art

A few bottom-up saliency databases are available with RGB-D images [17], however, many images in these datasets are not consistent with our definition of saliency (outdoor scenes, large objects, no flat surfaces). Other RGB-D datasets focus on evaluating object pose estimation and/or recognition [14]. Objects on those datasets are consistent with our definition, but the labeling is not always well-suited for saliency evaluation (use of bounding boxes, geometrically salient objects considered as distractors). We therefore evaluate our method on *RGB-D scenes* dataset [14], but to obtain the ground truth saliency maps, we add to the provided ground truth masks the objects considered as distractors in the dataset but that are salient given our geometrical definition.

We compare our results with four approaches. BMS [21] and GBVS [10] are among most accurate RGB saliency methods according to the MIT *saliency benchmark* [5]. Peng *et al.* [17] is a state-of-the-art method for RGB-D saliency, and Itti&Koch [13] is used as a baseline.

In a first test, we use 100 selected images from *RGB-D scenes*. We train a dedicated classifier for each image of the dataset so as to validate the accuracy of our saliency definition and the consistency of our architecture. For each image, we perform 500 random observations to train our classifier. Figure 4 shows a few results. Our approach not only selects consistent salient regions, but also provides a much more accurate shape and size of the salient objects than other techniques.

In a second test, we evaluate the generalization capability of our method when inputs are taken from a new point of view or with slightly different elements of the same environment. As *RGB-D scenes* is composed with video sequences, we select one of them (*table-small-2* in our case). Instead of training a dedicated classifier for each image in the sequence, we train a classifier on a portion of the sequence (60%), and we evaluate the performance on the rest of the sequence (40%).

Figure 5 represents the ROC curves of the five aforementioned approaches, for both tests. In all cases, our results outperforms state-of-the-art results, which makes sense as our saliency definition is optimized for this type of input content versus pure bottom-up saliency. In the second test, we compare the case where a different classifier is trained for each image (curve "train==test"), and the case where a global classifier is trained on a subset of the sequence and tested on another one (curve "train!=test"). Surprisingly, the "train!=test" configuration performs slightly better for a certain range of threshold. We intuitively explain this result based on Figure 6. On those images, the cap is partially missing in the "train==test" configuration, probably because they were not geometrically consistent with our definition of saliency under this point of view. However, those missing parts are found in the "train!=test" mode, as they are inferred from various points of view, and therefore more likely to have been detected at least a few times. On the other hand, the saliency map is more noisy, thus leading to a higher false positive rate.

B. Intrinsically motivated exploration

To validate the use of intrinsic motivation, driven either by uncertainty or novelty for target selection, we use again the *table-small-2* sequence. We evaluate four different criteria. Novelty and uncertainty are the one described in Section II-E, random is a pure random target selection and is a lower bound for evaluation. We also use a fourth criterion to get a higher bound. The target is selected based on the ground truth and aims to minimize prediction error. For that, we compare at each new frame our saliency prediction with the ground truth, and, using the same procedure as in Section II-E, we select the target where the prediction error is the worst.

To obtain an average value of the performance of each criterion, we run the saliency learning algorithm on the whole sequence four times. We perform a single foveal observation per frame, and the classifier is updated at each new frame, based on the foveal observation. We evaluate for each frame the average F1 measure[1] of the saliency estimation given the ground truth. Figure 7 shows the frame by frame evaluation for each criterion. We also display the F1 measure of BMS on this whole sequence for comparison.

Except at the very beginning of the sequence, all our methods outperforms BMS. As expected, random selection (called *rnd* in the figure) has the worst performance, and prediction error minimization (called *err*) is the best. The criteria based on novelty (*nov*) and uncertainty (*unc*) are in between, with a much better performance for the uncertainty criterion. The significant drop in the performance around frame 100 is due to a drastic change of point of view and strong illumination variations, making previous learned samples not enough to accurately predict new saliency. Nevertheless, the performance is lowered down for all the methods and the rank remains the same, still higher than BMS.

C. Real-time implementation

Our algorithm was implemented to capture a Kinect RGB-D stream, perform a frame by frame foveal observation, and train a classifier online in a separate thread as new observation are obtained. Our implementation was tested on Ubuntu 10.12 with an Intel Core i3-3240, CPU at 3.4GHz quadcore processor. For each step of the algorithm, the following processing times were obtained:

- Foveal observation and salient element discovery: 50ms to 500ms depending on the size of the observed area.
- Random forest training (separate thread) : 10s for 100 000 samples (100 to 1000 per frame)
- SEEDS and feature maps computation : 90ms
- Classifier-based saliency estimation and saliency map construction : 30ms

The total saliency map computation time (90ms+30ms) is fast enough for processing images at 8Hz and could be further

[1]The F1 score is the harmonic mean of precision and recall, defined as $F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. The F1 measure is a more meaningful score than accuracy, as the percentage of salient pixels in the sequence is quite low and the two classes are therefore unbalanced.

Fig. 8. Evolution of saliency as new observations are obtained. Number of observation is displayed under the saliency map.

optimized. Moreover, the foveal observation is sometimes time-consuming, but it only occurs in the exploration stage. Lastly, training the random forest in a separate thread allows acquisition of new examples without slowing down the execution time of saliency computation.

Figure 8 shows the state of saliency learning after some iterations. A video presenting the algorithm as well as robustness to motion and novelty is available online [2].

IV. Conclusion and future work

In this paper, we have presented an approach to learn visual saliency in an incremental and intrinsically motivated way. Using the mechanism of foveal view and visual attention, we acquire accurate saliency estimation in a small portion of the visual field of view, and incrementally learn saliency estimation in the large field of view based on simple and computationally inexpensive visual features.

This method shows good performance in the case of detecting small objects lying on flat surfaces, as a clear geometrical definition can describe these situations. In this case, our results outperform state-of-the-art. The use of simple intrinsic motivation criteria such as novelty or uncertainty can drive the selection of new targets and allow a better and faster learning. Lastly, our method has been implemented to run in real-time and is therefore well-suited for robotics applications.

In a future work, we will investigate further the intrinsic motivation strategy for target selection. In particular, we would like to investigate the use of competence progress. Suggested by Kaplan and Oudeyer [16], this type of intrinsic motivation focuses on progress rather than novelty or uncertainty. We believe that this type of criteria would improve the learning while avoiding unlearnable situations. Moreover, we would like to apply this framework for other types of foveal data, for example, an image with a better resolution and small field of view. Lastly, we plan to use this saliency learning on a mobile robot and exploit this saliency for object discovery or recognition.

References

[1] Adrien Baranès and P-Y Oudeyer. R-iac: Robust intrinsically motivated exploration and active learning. *Autonomous Mental Development, IEEE Transactions on*, 1(3):155–169, 2009.

[2] Mårten Björkman and Danica Kragic. Active 3d scene segmentation and detection of unknown objects. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3114–3120. IEEE, 2010.

[3] Aysecan Boduroglu, Priti Shah, and Richard E Nisbett. Cultural differences in allocation of attention in visual information processing. *Journal of Cross-Cultural Psychology*, 40(3):349–360, 2009.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Zoya Bylinskii, Tilke Judd, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. http://saliency.mit.edu/.

[6] José M Cañas, Marta Martínez de la Casa, and Teodoro González. An overt visual attention mechanism based on saliency dynamics. *International Journal of Intelligent Computing in Medical Sciences & Image Processing*, 2(2):93–100, 2008.

[7] Erkut Erdem and Aykut Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, 13(4):11, 2013.

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[9] Fred H Hamker. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Computer Vision and Image Understanding*, 100(1):64–106, 2005.

[10] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[11] MR Harter and L Anllo-Vento. Visual-spatial attention: preparation and selection in children and adults. *Electroencephalography and clinical neurophysiology. Supplement*, 42:183–194, 1990.

[12] Xiao Huang and John Weng. Novelty and reinforcement learning in the value system of developmental robots. 2002.

[13] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[14] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824. IEEE, 2011.

[15] Silviu Minut and Sridhar Mahadevan. A reinforcement learning model of selective visual attention. In *Proceedings of the fifth international conference on Autonomous agents*, pages 457–464. ACM, 2001.

[16] P-Y Oudeyer, Frédéric Kaplan, and Verena Vanessa Hafner. Intrinsic motivation systems for autonomous mental development. *Evolutionary Computation, IEEE Transactions on*, 11(2):265–286, 2007.

[17] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *Computer Vision–ECCV 2014*, pages 92–109. Springer, 2014.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.

[19] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2005.

[20] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. Seeds: Superpixels extracted via energy-driven sampling. In *Computer Vision–ECCV 2012*, pages 13–26. Springer, 2012.

[21] Jianming Zhang and Stan Sclaroff. Saliency detection: a boolean map approach. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 153–160. IEEE, 2013.

[22] Qi Zhao and Christof Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, 11(3):9, 2011.

[23] Jun-Yan Zhu, Jiajun Wu, Yichen Wei, Eric Chang, and Zhuowen Tu. Unsupervised object class discovery via saliency-guided multiple class learning. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3218–3225. IEEE, 2012.

[2] http://perso.ensta-paristech.fr/∼ craye/