Let it Learn
A Curious Vision System for Autonomous Object Learning

Pramod Chandrashekhariah*, Gabriele Spina* and Jochen Triesch
Frankfurt Institute for Advanced Studies,
Johann Wolfgang Goethe University, Frankfurt am Main, Germany

Keywords: Active Vision, Unsupervised Learning, Autonomous Vision System, Vision for Robotics, Humanoid Robot, iCub, Object Recognition, Visual Attention, Stereo Vision, Intrinsic Motivation.

Abstract: We present a "curious" active vision system for a humanoid robot that autonomously explores its environment and learns object representations without any human assistance. Similar to an infant, who is intrinsically motivated to seek out new information, our system is endowed with an attention and learning mechanism designed to search for new information that has not been learned yet. Our method can deal with dynamic changes of object appearance which are incorporated into the object models. Our experiments demonstrate improved learning speed and accuracy through curiosity-driven learning.

1 INTRODUCTION

One of the hallmarks of biological organisms is their ability to learn about their environment in a completely autonomous fashion. Future generations of robots assisting humans in their homes should similarly be able to autonomously acquire models of their working environment and any objects in it. While computer vision has made much progress in developing object recognition systems that can deal with many object classes, these systems need to be trained with supervised learning techniques, where a large number of hand-labeled training examples is required. Only recently, researchers have started addressing how a robot can learn to recognize objects in a largely autonomous fashion, *e.g.*, (Kim et al., 2006), how learning can be made fully online (Wersing et al., 2007; Figueira et al., 2009) and how the need for a human teacher can be minimized (Gatsoulis et al., 2011). To this end, current attention systems of robots (Begum and Karray, 2011) have to be extended such that they support an efficient autonomous learning process.

The central inspiration of our approach is the concept of intrinsic motivation (Baranes and Oudeyer, 2009; Schmidhuber, 2010; Baldassarre, 2011). Children learn and build internal representations of the world without much external assistance. Instead, they are intrinsically motivated to explore and play and thereby acquire knowledge and competence. In short, they are curious. It has been proposed that infants' interest in a stimulus may be related to their current learning progress, *i.e.*, the improvement of an internal model of the stimulus (Wang et al., 2011). We adopt the same idea to build a "curious" vision system whose attention is drawn towards those locations and objects in the scene that provide the highest potential for learning. Specifically, our system pays attention to salient image regions likely to contain objects, it continues looking at objects and updating their models as long as it can learn something new about them, it avoids looking at objects whose models are already accurate, and it avoids searching for objects in locations that have been visited recently. We show that our system learns more efficiently than alternative versions whose attention is not coupled to their learning progress.

2 OBJECT LEARNING

Our system is implemented on the iCub robot head (Metta et al., 2008), Fig. 1. Its basic mode of operation is as follows. An attention mechanism generates eye movements to different locations. Any object present at the current location is segmented and tracked while learning proceeds. If the object is unfamiliar then a new object model is created. If the

Chandrashekhariah P., Spina G. and Triesch J..
Let it Learn - A Curious Vision System for Autonomous Object Learning.
DOI: 10.5220/0004294101690176
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 169-176
ISBN: 978-989-8565-48-8

Figure 1: The iCub robot (a) and its learning environment (b).

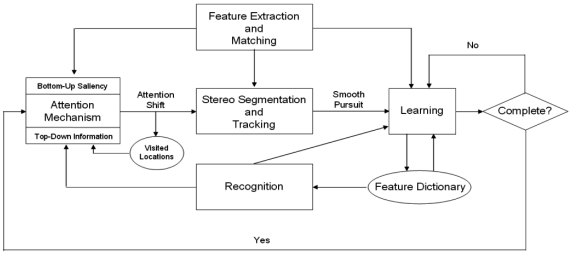

Figure 2: System architecture.

object is already familiar, then its model is updated if necessary. Learning proceeds for as long as the model can be improved. Then a new focus of attention is selected. Figure 2 shows the system architecture, which is explained in detail in the following sections.

We describe objects as spatial arrangements of local image features, an approach that is robust to occlusions, local deformations, variation in illumination conditions, and background clutter, *e.g.*, (Agarwal and Roth, 2002). To this end, image features are extracted at interest points detected with the Harris corner detector (Harris and Stephens, 1988). We use Gabor wavelet features, which have the shape of plane waves restricted by a Gaussian envelope function. At each interest point we extract a 40-dimensional feature vector, which we refer to as a Gabor-jet, resulting from filtering the image with Gabor wavelets of 5 scales and 8 orientations, *e.g.*, (Wiskott et al., 1997). The choice of the features is motivated by the fact that they have a similar shapes as the receptive fields of simple cells found in the primary visual cortex of mammals (Jones and Palmer, 1987).

2.1 Stereo Segmentation and Tracking of the Object

To segment a potential object at the center of gaze from the background, we make use of stereo information. We find correspondences between interest points detected in the left and right image by exhaustively comparing Gabor-jets extracted at the interest points from left and right image, see Fig. 3a,b. Each interest point in the left image is associated with the best matching interest point in the right image if the similarity S between the two jets (we use the normalized inner product) is above a preset threshold (0.95 in our current implementation). We then cluster the matched interest points from the left image (that is used for learning) into different groups according to their image location and disparity (Fig. 3c). We use a greedy clustering scheme that starts with a single interest point and adds new ones if their x-position, y-position, and disparity are all within 5 pixels of any existing cluster member. Figure 3d shows how the object at the center of gaze is properly segmented from other objects which are at a similar depth but different spatial location or at a close-by spatial location but different depth.

After segmentation the cameras are moved to bring the object to the center of view and keep it there — in case the object is moving — by a tracking scheme. To this end, the mean location of foreground features is calculated, then this location is tracked with both eyes using a model-free tracking scheme called Democratic Integration (DI) (Triesch and Malsburg, 2001). DI is a multi-cue tracking system that provides a fast and robust way of tracking unknown objects in a changing environment. Once the object is at the center of gaze, model learning starts.

2.2 Learning Object Models

Once an object has been segmented and fixated, its novelty or familiarity is determined by the recognition system described in section 2.4. If the object is already familiar, the recognition module provides the unique identity of the object, *i.e.*, an object index that was assigned when the object was first encountered. Otherwise a new object index is assigned.

Object learning involves the generation of a model that has a set of associations between the Gabor wavelet features and the object index (Murphy-Chutorian and Triesch, 2005). An association is made between a feature and an object index if they occur together during learning and it is labeled with the distance vector between the location of the feature and the center of the object, *i.e.*, the point on the object on which gaze is centered.

2.3 Feature Dictionary

Object learning is carried out in an on-line fashion. There are no separate training and testing/recognition phases. As the system starts learning, the models for all the objects are learnt incrementally using a shared feature dictionary accumulating information

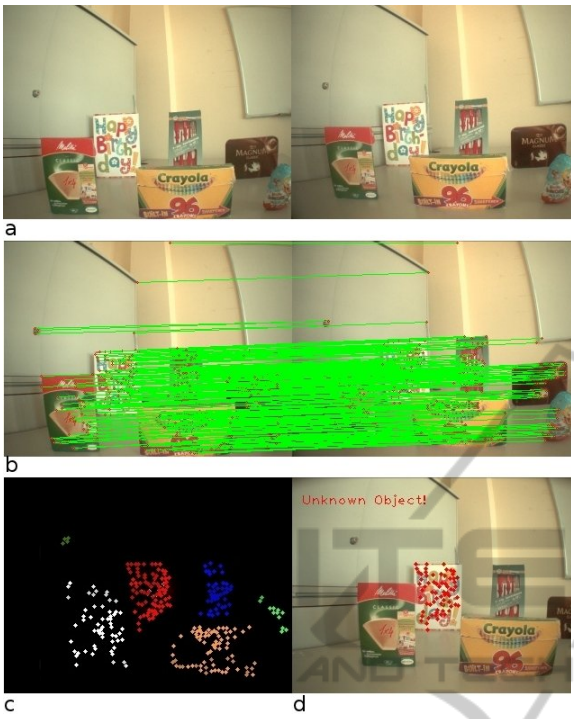

Figure 3: Several objects are placed in front of the robot that analyzes the scene using its cameras (a). Harris corner points are detected and matched across left and right image using Gabor-jets (b). A low resolution saliency map is used to select the most salient interest points in the scene. Interest points on left image are clustered based on their location and stereo disparity (c). Spurious clusters with less than 3 features are removed. Attention shifts to the most salient object that is segmented out from the scene (d).

about objects and the associated feature vectors. We use a single-pass clustering scheme that updates the feature dictionary for every input feature vector. Let $\mathscr{C}$ be the set of clusters and n be the number of clusters in the feature dictionary. Once the system starts learning it adds features from the objects in the scene. Each input feature vector $\mathscr{J}$ has an associated object index k and the distance vector (x,y) to the object center measured in pixels. In the beginning, when the dictionary is empty, a cluster is created and it will be represented by the input vector. Subsequently, when the number of clusters grows, the algorithm decides to either assign a feature to an existing cluster (without altering its representation) if the similarity value S is higher than a threshold θ (equal to 0.95)(see $\diamond$ in Fig. 4) or make it a new cluster otherwise ($\star$ in Fig. 4). During each update, object index and distance vector are associated to the same cluster. When a feature matches an existing cluster, a possible duplicate association of this cluster to the current object is avoided. If the object index is the same and if the feature locations are within a euclidean distance of 5.0

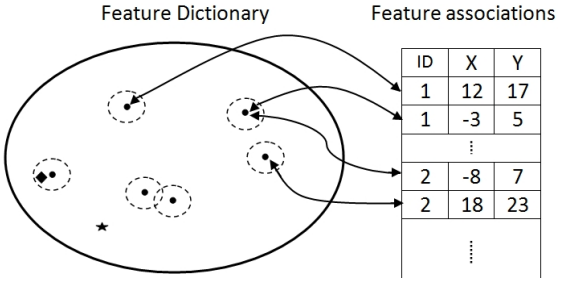

Figure 4: $\bullet$: Cluster centers (dotted lines indicate the boundaries). $\star$: Input for which new cluster is created. $\diamond$: Input for which no new cluster is created.

pixels the association is neglected. The algorithm can be summarized as follows:

Algorithm 1: Online learning of feature dictionary.

Initialize $n \leftarrow 0$, $\theta \leftarrow 0.95$.
loop
  Provide new feature vector $\mathscr{J}$ and distance vector (x,y).
  Obtain object index k from recognition (new or existing)
  Calculate $i_{win} = \arg\max_i S(\mathscr{J}, \mathscr{C}_i)$
  if $S(\mathscr{J}, \mathscr{C}_{i_{win}}) < \theta$ **then**
    $n \leftarrow n+1$, $\mathscr{C}_n \leftarrow \mathscr{J}$
    Store association of $\mathscr{C}_n$ with object k at (x,y)
  else
    if $\mathscr{C}_{i_{win}}$ not associated with object k at (x,y) **then**
      Store association of $\mathscr{C}_{i_{win}}$ with object k at (x,y)
    end if
  end if
end loop

2.4 Recognition

In our work recognition is an integral part of the learning process. When the robot looks at an object the features on the segmented portion are sent to the recognition module and compared with the features in the dictionary. We use a generalized Hough transform (Ballard, 1987) with a two dimensional parameter space for recognition. Each feature votes in the space of all object identities and possible centroid locations based on their consistencies with the learned feature associations. Features with a similarity value higher than 0.95 will cast one vote each for the object identities that they match in the feature dictionary. Votes having information about object's identity as well as object's location are then aggregated in discretized bins in Hough space. We use bins of size 5×5 pixels in our work. If the number of votes in a bin favor-

ing a particular object index is greater than a predefined threshold (10 in this implementation) we declare the object as being present at the corresponding location. However, if there are different bins voting for the same object at different locations in the scene due to possible false feature matching, the location with the maximum number of votes is marked as the expected location. In the end, the recognition module returns a set of locations corresponding to those objects in the model whose voting support was sufficient.

3 ATTENTION MECHANISM

Our attention mechanism controls what the robot will look at, for how long it will keep looking at it, and where it should avoid looking. We embody curiosity in the attention mechanism by introducing the following ways of guiding attention to where learning progress is likely.

3.1 Bottom-up Saliency at Interest Points

We have adapted a bottom-up saliency model developed by Itti et al. (Itti and Koch, 2001). In this model the conspicuity of each image location in terms of its color, intensity, orientation, motion, etc. is encoded in a so-called saliency map. We make use of stereo information to select the most salient point in the scene. Images from both eyes are processed to obtain left and right saliency maps. Since objects are represented as features extracted at interest points, our attention mechanism only considers points in the saliency map that are associated with a pair of interest points matched between left and right image (all other points are neglected). In this way we restrict attention to locations of potential objects that the system could learn about. The saliency values for the matched interest points are computed using a 2-dimensional gaussian centered on them, with σ = 1.5 and a cutoff value of 0.05. This has the effect of bringing out clusters of high salience more than just isolated pixels of high salience.

When there are no other variations in the visual characteristics of the scene it is very likely that the attention mechanism continues to select the same location as the most salient point. To avoid this we temporarily inhibit the saliency map around the current winner location by subtracting a Gaussian kernel at the current winner location. This allows the system to shift attention to the next most salient location. To avoid constant switching between the two most salient locations, we also use a top-down inhibition of already learned objects below.

3.2 Attention based on Learning Progress

It has been argued that infants' interest in a stimulus is related to their learning progress, *i.e.*, the improvement of an internal model of the stimulus (Wang et al., 2011). We mimic this idea in the following way. When the robot looks at an object, it detects whether the object is familiar or not. If the object is new it creates a new object model making new associations in the shared feature dictionary. If the object is known, the model is updated by acquiring new features from the object. The attention remains focused on the object until the learning progress becomes too small. As a side effect, the robot continues learning about an object when a human interferes by rotating or moving it, exposing different views with unknown features.

3.3 Top-down Rejection of Familiar Objects

The third mechanism to focus attention on locations where learning progress is likely makes use of the system's increasing ability to recognize familiar objects. A purely saliency-based attention mechanism may select the same object again and again during exploration, even if the scope for further learning progress has become very small. Therefore, once there are no more new features found on certain objects, our system inhibits their locations in the saliency map whereever they are recognized (Fig. 5a). To this end, the models of these objects are used to detect them in every frame using the recognition module. The interest points on the saliency map that are in the vicinity of the object detections are removed from being considered for the winner location.

3.4 Top-down Rejection of Recently Visited Locations

We have incorporated an inhibition-of-return mechanism that prevents the robot from looking back to locations that it has recently visited. To this end, the absolute 3D coordinates of the visited locations are saved in the memory and they are mapped onto the pixel coordinates on images from the cameras in their current positions to know the locations for inhibition. In our experiments, a list of the 5 most recently visited locations is maintained and close-by interest points are inhibited for the next gaze shift (Fig. 5b).

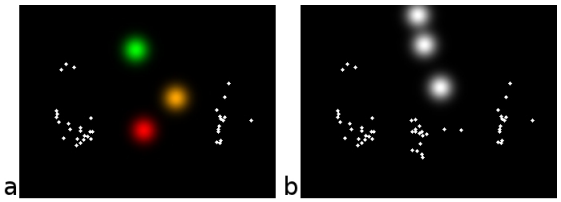

Figure 5: (a) Top-down rejection of familiar objects: When objects become familiar to the robot they will be inhibited for further selection by removing the corresponding interest points. Color blobs indicate recognized objects whose interest points have been removed. (b) Top-down rejection of visited locations: The robot inhibits recently visited locations (white blobs).

In order to ease exploration of regions beyond the current field of view, we have also added a mechanism to occasionally turn the head in a new direction. To this end, the space of possible head directions is parcellated into 4 quadrants. Whenever the robot has visited ten locations in one quadrant it shifts to the opposite quadrant.

4 EXPERIMENTS AND RESULTS

The system described above incorporates several mechanisms to make it intrinsically motivated to seek out new information or, simply put, to make it curious. To evaluate the benefits of this curiosity, we test the performance of the system by incorporating one or more of the attention mechanisms in a staged manner. We will label the full system including all mechanisms as the IM (intrinsic motivation) system.

4.1 Experimental Setup

The model is implemented on an iCub robot head (Metta et al., 2008) (Fig. 1a). It has two pan-tilt-vergence eyes mounted in the head supported by a yaw-pitch-twist neck. It has 6 degrees of freedom (3 for the neck and 3 for the cameras). Images are acquired from the iCub cameras at 27 fps with resolution of 320×240 pixels. Experiments are performed placing iCub in a cluttered environment with various objects in the scene that are placed at different depths with partial occlusions. The background comprises walls, doors and book shelves. Figure 6 shows the objects, which have different sizes and shapes.

4.2 Evaluation Method

To evaluate the system, we let the robot autonomously explore its environment for 5 minutes and then test its performance using previously recorded and manually segmented ground truth images. During ground truthing we manually control the robot to look at each object present in the scene. The robot will extract features on the objects, that are manually segmented, until it does not find any new feature. This period was observed to be less than 10 frames on an average for static objects, but more for rotating/moving objects (see below). Once all the features are collected on all the objects, they are tested with the model generated by the system at the end of the learning process. To evaluate the performance of the system we consider the following parameters: Number of objects learnt, number of visits on an object (to test the exploration efficiency), accuracy of the object models (in terms of repeated object identities, missed/wrong detections, recognition rate), and time taken for learning the objects. Since the object identities depend on the order in which objects are learnt, we programmed the systems to store representative images of the object together with the self-assigned object ID. These images are displayed while testing and allow a visual verification of the correctness of the recognition.

Figure 6: Objects used in the experiments. Black frames indicate the objects used in the dynamic object scenario.

4.3 Two Experimental Scenarios

In the following we describe two testing scenarios using static and dynamically changing scenes.

In the first scenario, objects are static and iCub has to actively explore the scene and learn about the objects. We set a time span of 5 minutes during which iCub learns as many objects as possible. We place 12 objects in the scene allowing partial occlusions. Object locations are varied from one experiment to another.

In the second scenario we tested the ability of the system to update the model of an object with new features (Fig. 10). We used only 3 objects that are rotated by a human to dynamically change the objects' ap-

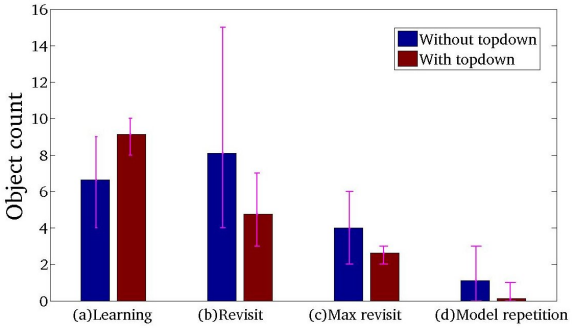

Figure 7: Comparison of system performance with and without top-down information in the static object scenario. (a) Total number of objects learnt. (b) Total number of revisits of objects. (c) Maximum number of object revisits. (d) Number of objects whose models were duplicated

pearance while iCub learns about them. The learned object models are evaluated with separate test images showing the objects in four different poses.

4.4 Results

In this section we illustrate the performance of our system in a staged manner. We have employed bottom-up saliency in all the experimental scenarios. We will demonstrate a further improvement in attention and learning mechanism by using top-down information and learning progress parameters on top of this.

We will first illustrate the effect of top-down information on the system's performance in the static object scenario. Figure 7 compares the system's performance with and without top-down information. We report average values over 10 experiments carried out with different objects, locations, and lighting conditions. Error bars represent maximum and minimum values. Figure 7a shows the number of objects learnt by the system in 5 minutes that were validated by ground truth. Figure 7b shows the number of revisits of objects during exploration. In the absence of top-down information the system visits some objects repeatedly although little new information is available there. Similarly, Fig. 7c shows the maximum number of revisits across all objects. Figure 7d shows the number of objects whose models were incorrectly duplicated, *i.e.*, the system did not recognize the object when visiting it at a later time and created a second object model for the same object. Figure 8 shows the comparison in terms of time taken by the system to learn the first n objects. Across all measures, the system using top-down information is superior to the one without. One can expect a higher performance on a robot that has higher visual range and resolution covering more objects in the scene.

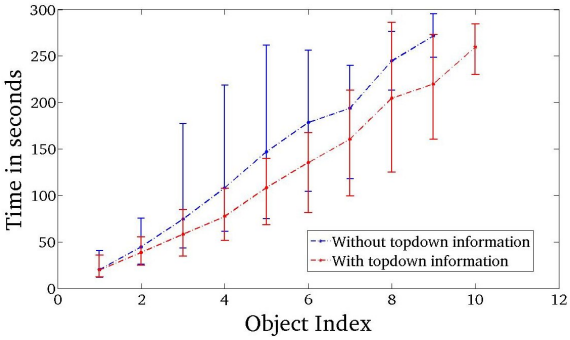

Figure 8: Comparison of the system with and without top-down attention in terms of the time taken by the system to learn the first n object models.

Our system looks at an object for as long as it finds something new to learn about. To evaluate the benefit of this feature we compare the full system (IM) to a version that only looks at an object for a fixed duration (equal to 3 seconds which was observed to be sufficient for learning an arbitrary object) before shifting gaze (No IM). The advantage of the full IM system is illustrated in the rotating object scenario. For this experiment we used the three objects marked by black rectangles in Fig. 6. The objects are rotated by a human operator as the robot learns about them (see Fig. 10). It is observed that the full IM system avoids duplicate representations for the same object. Figure 9 shows feature to object associations after learning. The features corresponding to an object model are collected and their distance vectors are marked from the center of the object. Figure 9a shows that for the IM case the features are densely populated covering most of the parts of the object. As our object models are pose invariant what is depicted in

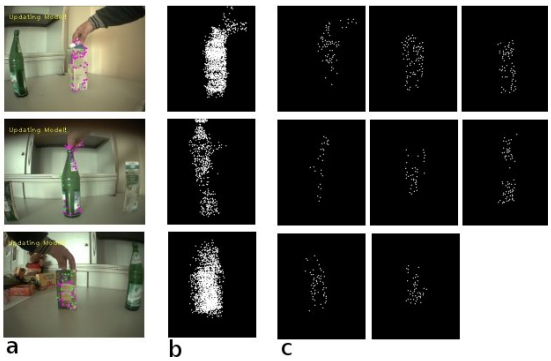

Figure 9: Features belonging to the model for the learnt object are marked at locations given by distance vectors from the object center that were saved in the feature dictionary. (a) Objects during their learning progress. (b) For IM: Features cover the object densely and the object model is not duplicated. (c) For No IM: Features are sparse and there are duplications of object representations in the feature dictionary.

Figure 10: The IM system updates the model of the object finding new features on it while it is rotated by a human operator. Red dots represent features found on the object when the model is first created, purple dots represent new features found on the object during the model update, green dots represent shared features that have previously been associated with this object but found at a different location on the object.

Table 1: Object representation in the feature dictionary.

	Milk packet		Water bottle		Tea box	
	No IM	**IM**	**No IM**	**IM**	**No IM**	**IM**
Model 1	63	1511	27	535	55	1601
Model 2	82	–	35	–	44	–
Model 3	69	–	65	–	–	–

Table 2: Recognition Accuracy (Rotating Objects).

	Milk packet		Water bottle		Tea box	
Pose	**No IM**	**IM**	**No IM**	**IM**	**No IM**	**IM**
Pose 1	57.14%	100%	18.86%	52.57%	–	100%
Pose 2	–	96.10%	32.14%	57.14%	27.58%	100%
Pose 3	20.58%	100%	–	–	–	100%
Pose 4	30.88%	67.64%	–	–	–	100%

the picture is the aggregation of feature vectors from all poses that are captured in the model. Figure 9b shows that for the other case there are duplicate models for the same object in the feature dictionary as the system in this case fails to realize that an object seen sometime later exhibiting different pose is the same object hence learning a new object model with new identity. The features are also not dense enough to identify the objects with high reliability. This is evident from Table 1 that lists the number of associated features in the feature dictionary for every object and the corresponding models. As shown in Table 2, the full IM system also has superior recognition accuracy. Recognition accuracy is defined as the percentage of features of the object model matched with ground truth. Four different poses of every object are shown to the system to see how well it can recognize. We observe that the recognition accuracy is substantially higher for the IM case.

5 CONCLUSIONS

We have presented a "curious" robot vision system that autonomously learns about objects in its environment without human intervention. Our experiments comparing this curious system to several alternatives demonstrate the higher learning speed and accuracy achieved by focusing attention on locations where the learning progress is expected to be high. Our system integrates a sizeable number of visual competences including attention, stereoscopic vision, segmentation, tracking, model learning, and recognition. While each component leaves room for further improvement, the overall system represents a useful step towards building autonomous robots that cumulatively learn better models of their environment driven by nothing but their own curiosity.

ACKNOWLEDGEMENTS

This work was supported by the BMBF Project "Bernstein Fokus: Neurotechnologie Frankfurt, FKZ 01GQ0840" and by the "IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots" project, FP7-ICT-IP-231722.

We thank Richard Veale, Indiana University for providing the code on saliency.

REFERENCES

Agarwal, S. and Roth, D. (2002). Learning a sparse representation for object detection. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 113–130, London, UK, UK. Springer-Verlag.

Baldassarre, G. (2011). What are intrinsic motivations? a biological perspective. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1 –8.

Ballard, D. H. (1987). Readings in computer vision: issues, problems, principles, and paradigms. chapter Generalizing the hough transform to detect arbitrary shapes, pages 714–725. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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

Begum, M. and Karray, F. (2011). Visual attention for robotic cognition: A survey. *Autonomous Mental Development, IEEE Transactions on*, 3(1):92 –105.

Figueira, D., Lopes, M., Ventura, R., and Ruesch, J. (2009). From pixels to objects: Enabling a spatial model for humanoid social robots. In *Robotics and Automation, 2009. ICRA 2009. IEEE International Conference on*, pages 3049 –3054.

Gatsoulis, Y., Burbridge, C., and Mcginnity, T. (2011). Online unsupervised cumulative learning for life-long robot operation. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, pages 2486 –2490.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.

Jones, J. and Palmer, L. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, 58(6):1233 –58.

Kim, H., Murphy-Chutorian, E., and Triesch, J. (2006). Semi-autonomous learning of objects. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on*, page 145.

Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, PerMIS '08, pages 50–56, New York, NY, USA. ACM.

Murphy-Chutorian, E. and Triesch, J. (2005). Shared features for scalable appearance-based object recognition. In *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 16 –21.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *Autonomous Mental Development, IEEE Transactions on*, 2(3):230 –247.

Triesch, J. and Malsburg, C. V. D. (2001). Democratic integration: Self-organized integration of adaptive cues.

Wang, Q., Chandrashekhariah, P., and Spina, G. (2011). Familiarity-to-novelty shift driven by learning: A conceptual and computational model. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1 –6.

Wersing, H., Kirstein, S., Gtting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J., Ritter, H., and Krner, E. (2007). Online learning of objects in a biologically motivated visual architecture.

Wiskott, L., Fellous, J.-M., Kuiger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):775 –779.