Learning and Recognition of Objects Inspired by Early Cognition

Maja Rudinac, Gert Kootstra, Danica Kragic and Pieter P. Jonker

Abstract— **In this paper, we present a unifying approach for learning and recognition of objects in unstructured environments through exploration. Taking inspiration from how young infants learn objects, we establish four principles for object learning. First, early object detection is based on an attention mechanism detecting salient parts in the scene. Second, motion of the object allows more accurate object localization. Next, acquiring multiple observations of the object through manipulation allows a more robust representation of the object. And last, object recognition benefits from a multi-modal representation. Using these principles, we developed a unifying method including visual attention, smooth pursuit of the object, and a multi-view and multi-modal object representation. Our results indicate the effectiveness of this approach and the improvement of the system when multiple observations are acquired from active object manipulation.**

I. INTRODUCTION

Bringing artificial systems to real-world environments poses many different problems that must be solved. One of the challenges is to recognize objects despite the uncontrolled nature of the real world. Variations in object appearance due to viewpoint or environmental conditions need to be overcome by the system. In this paper, we approach this challenge by taking inspiration from object learning in infants. As defined in the cognitive theory of Piaget [18], infants learn representations of objects by actively exploring them. Doing so, allows to observe the objects from different viewpoints, and thus exploring the possible variations in appearance. In early stages in child development, the infant's visual attention is directed primarily to salient parts of the environment [22]. The child will first be able to learn representations of the objects that are actively shown by the caregivers [9]. In later stages, the infant will learn to manipulate and explore the objects independently [20].

In this paper, we aim to mimic the early stage of object learning on an artificial cognitive system, using a caregiver to demonstrate objects by manipulating them. We believe it is important to start at an early stage, in order to develop and test important concepts in object learning. Future versions of our system will develop in line with child development, as advocated in [28].

Figure 1 shows our cognitive model of object learning and recognition. The model is based on Baddeley's model of working memory [1] and Knudsen's model of attention [10]. We narrow both models down to the parts that deal with visual information. The central executive is responsible for the control of cognitive processes and, in our approach, has the coordinating role in visual learning and recognition, involving the long-term memory, which stores the object representations. Additionally, it is involved in the control of attention. The visual memory (termed visio-spatial sketchpad in [1]) holds the visual information of the attended regions, such as color, texture, and shape information. Our model furthermore includes an attention mechanism as an interface between the working memory and the outside world, as proposed in Knudsen's model of attention [10]. Attention is focused on relevant parts of the visual field based on different types of visual information, such as color, intensity, and motion. The focus of attention can be changed to different parts of the visual field through saccadic eye movements or in order to track an object through smooth pursuit.

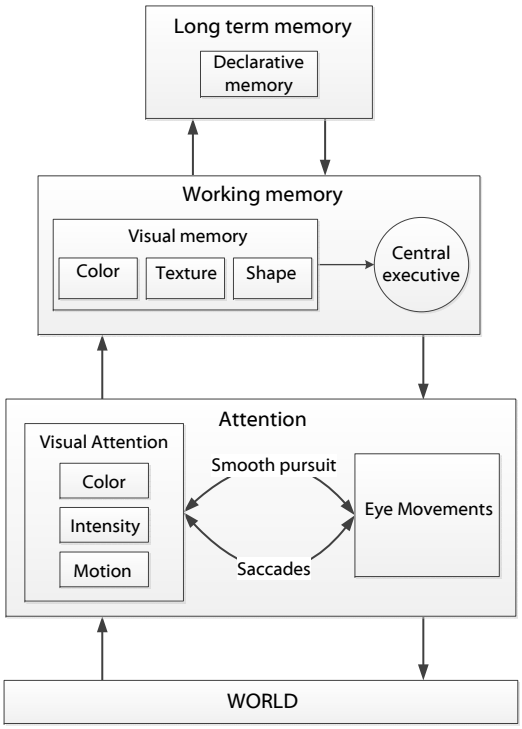

Fig. 1: Cognitive model for object learning and recognition based on Baddeley's model of working memory [1] and Knudsen's model of attention [10].

The first problem that arises in learning, is how to localize and segment unknown objects from the background. For the detection of unknown objects in a scene, no top-down knowledge can be used. Object-detection methods based on 3D point clouds calculated from stereo-image pairs [2] provide good results in the case of the textured objects. However, they fail in the case of uniform colored objects which are widely present in the environment. As a solution to this challenging problem, we therefore consider bottom-up

visual-attention methods.The saliency method presented in [8] has, for instance, been used in [21] to guide the attention of a robot. An attention method based on local symmetry in the image has been proposed in [11] to fixate on objects in the scene. Finally, following method [24] provides fast segmentation of objects based on their saliency. Since this method assumes no prior information about the scene and only requires input from a single camera, we will further exploit it in the initial step of our method.

Once objects are located, they need to be explored and manipulated so that the system can learn them properly. Active exploration in robotics is a hot topic and several systems have been recently proposed [12], [15], [5]. For example, in the work of [17] manipulation of object has been used in a bottom up attention system as a top-down knowledge to control visual search of that object. Furthermore, by changing its viewpoint, the robot can actively test the robustness of its object representations. This has also been used, for instance, to select stable interest points for robust object representations [12]. Object exploration can also be used to build a more complete object model by integrating different viewpoints, e.g., in [19] and in [6] In our research we combine similar approaches to obtain a robust system for active learning.

Finally, objects need to be robustly described and later recognized in constantly changing illumination settings and cluttered environments. The best results in challenging settings were obtained using local features [13] and their extensions to color [27]. Nevertheless, using keypoints to describe novel objects will only work in the case of textured objects, and a combination with other methods is required for recognition of uniformly colored ones. Therefore, we propose to utilize the method of [23], which combines both approaches and automatically calculates dominant features of the object.

The main contribution of this paper is a novel unifying system for learning of object representations when no prior knowledge is available. All system knowledge is bootstrapped online by object manipulation. Mimicking cognitive development in infants, the system first localizes unknown objects in the environment as salient regions. For this, an adapted version of our previous saliency method is used, which assumes no initial information on the scene and provides fast scene segmentation. Once localized, the object, manipulated by the caregiver, is tracked, and segmented to obtain different observations of the object. Based on these observations, a multi-view multi-modal representation of the object is build, which is used for learning and recognition. The performance of the system is extensively tested, and the benefits of object manipulation to obtain a sequence of observations is shown.

II. System layout

We assume a setup with a single camera where a human caregiver presents objects to the system by actively manipulating them. Initially, the system has no knowledge of the objects, but over time, the system will learn from the demonstrations to recognize the objects. Our system depends on four modules for both learning and recognition of objects in unstructured environments: visual attention, smooth pursuit, object description, and novelty detection. The visual-attention module makes a first estimate of the location and the initial segmentation of the objects in the scene. This module detects visual saliency without any prior knowledge of the objects or the scene. The resulting object segments initialize the smooth-pursuit module, which tracks the object that is actively shown to the system by the human caregiver. Throughout the manipulation, the segmentation of the tracked object is incrementally improved. The segmentation is continuously used by the object-description module for an incremental and multi-modal visual description of the object. The feature vector consists of color, texture, and shape information. Based on all stored feature vectors, the dominant features are emphasized to improve performance. Finally, based on all observations of the object during manipulation, the novelty detector classifies the object under inspection as either novel or known, based on which the sequence of observations are either used to learn the new object, or to recognize the known object.

The active manipulation of the objects has several advantages. Firstly, the motion can be used to track the object and to improve object segmentation. Secondly, through the manipulation of the object, the system acquires novel viewpoints, which enriches the object representation. Finally, multiple viewpoints of the same object, allow the system to better estimate the dominant features and the intra- and inter-object variability.

The four modules are described in the next subsections, followed by a description of the incremental learning method and the active recognition method.

A. Visual Attention

The visual-attention module finds the salient parts of the image in order to detect and segment objects without any prior knowledge on the objects or their backgrounds. We extend our saliency method for fast object segmentation, proposed in [24], to benefit from multiple observations during the manipulation of the object.

At its basis, the method calculates the spectral residuals in three different color channels, red-green, blue-yellow, and illumination as proposed in [7]. These residuals are calculated by taking the Fourier transform of an image, and taking the difference between the magnitude of the frequency spectrum and a low-passed version of this magnitude. A saliency map is obtain by transforming the residuals back to the spatial domain and summing the three color channels. The resulting saliency map defines for every pixel in the image how much it stands out from the background.

Next, salient regions are found by applying the MSER blob detector [14] to the saliency map. For a given image frame at time t, this results in a set of salient regions, R_t. We observe the salient regions over a number of s consecutive frames, resulting in a combined set of n regions, $R = \bigcup_{j=1}^{s} R_j = \{r_1, \ldots, r_n\}$ where r_i is the center of i-th salient region.

Fig. 2: Results of the visual attention module. The yellow contours indicate the salient regions and the black boxes the regions of interest.

Figure 2 shows several contours of salient regions in yellow. Combining the salient regions over several frames improves the robustness and quality of object detection. However, a too large number of frames increases the chance of obtaining overextended segments. In our experiments, we obtained good results with $s = 5$.

Since multiple salient regions can correspond to one object, the regions are clustered using adapted Parzen-window density estimation [26] followed by mean-shift clustering. The density estimation is made by fitting a Gaussian kernel to each of the centers of the salient regions, r_i. For each point in the image x, the probability density function, $p(x)$, is defined by:

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp^{-\frac{(x-r_i)^2}{2\sigma^2}} \quad (1)$$

where σ represent the width of the Gaussian kernel optimized by maximizing the likelihood [26] and n is the number of contour centers. Subsequently, outlier points that have low probability values and belong to isolated clusters are removed when

$$\log(p(x)) < \frac{1}{n} \sum_{i=1}^{n} \log\left(p(r_i)\right) - 3 \underset{i=1}{\overset{n}{\text{var}}} \left(\log(p(r_i))\right) \quad (2)$$

where var_i gives the variance over i. Finally, the salient regions are clustered by segmenting $p(x)$ using mean-shift segmentation[4]. As final result, we find the regions of interest around each object in the scene, as illustrated in Fig. 2. The regions of interest serve as the initial segmentations for learning and tracking of the objects during manipulation.

B. Smooth pursuit

The visual-attention module provides an initial segmentation of the objects in the scene, which is used by the smooth-pursuit module to track the object during manipulation. We combine two supplementary methods to robustly obtain good segmentations of the object. The first method follows the moving object using a model-based tracker [16]. The tracker builds a color histogram of the object based on the region of interest provided by the attention module. Using the mean-shift framework, the object is tracked over successive frames. [16].

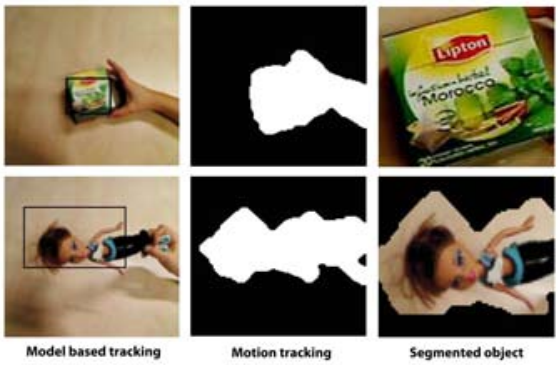

Fig. 3: Results of the smooth-pursuit module. The results of model-based tracking and motion tracking are combined to segment the object.

The second method finds the object by detecting motion in the image using motion-history images (MHI) [3]. The MHI combines the motion changes over a sequence of images. For a single frame, the regions of motion are found by calculating the frame difference with the previous frame. By thresholding and subsequently applying morphological dilation and erosion, a motion silhouette is obtained. Especially for low-textured object or when objects are moved slowly, this single-frame silhouette will not correspond well to the object. However, by continuously adding the silhouettes to the MHI, this problem is solved. Using a decay over time, the MHI defines the temporal history of motion at each point in the image. The MHI is then thresholded and the largest connected component is returned.

The bounding box returned by the model-based tracker and the silhouette returned by the motion tracker are combined to get the segmented object. The two methods complement each other. On the one hand, due to fast motions or illumination changes, the tracker can lose the object. This can be compensated for by the motion detector. On the other hand, using only motion segmentation would result in including the manipulator (in our case the human hand) in the object segment. By combining it with the model-based tracker, initialized by the attention module, only the object is segmented.

Figure 3 gives some examples of the model-based tracking, the motion tracking, and the combined object segmentation, illustrating the benefit of the combination of both methods. Figure 4 shows some examples of different viewpoints of objects while being tracked during manipulation.

C. Object description

Once a viewpoint of the object is segmented, it is described using visual features and temporarily stored in the visual memory. Since the objects appearances can vary significantly both in color, texture and shape, we propose to extract all corresponding feature vectors and to automatically calculate dominant features of the object. For this we utilize our previous research on fast and robust feature descriptors [23]. There we used a color histogram including hue, saturation and value

as a color descriptor, a Gray Level Co-occurrence Matrix (GLCM) as a texture descriptor, and an edge histogram for shape information, all combined into a single $D = 256$ dimensional vector $\mathbf{f}$. This feature vector thus describes both textured and untextured objects.

Some features might be more descriptive than others. We therefore perform a normalization step, in order to emphasize the dominant features. We have shown that this has advantages over methods that treat every element in the feature vector as equally important in [23], where we focused on object recognition in challenging situations, such as occlusions and variable illumination conditions. In this paper, we adapt this method to an online situation, where multiple viewpoints of the objects are acquired.

All viewpoints of all objects are merged in one feature matrix, $\mathbf{F}$, where $\mathbf{F}(i,j)$ is the value in row i and column j, which gives the j-th feature of feature vector $\mathbf{f}_i$. As the different features are obtained in a different way, the columns in the feature matrix have significantly different values. We therefore first normalize $\mathbf{F}$ by dividing all values by the maximum value in their respective column:

$$\bar{\mathbf{F}}(i,j) = \mathbf{F}(i,j)/\mathbf{m}(j) \tag{3}$$

$$\mathbf{m}(j) = \max_{i=1}^{N} \mathbf{F}(i,j) \tag{4}$$

where $\mathbf{m}$ is the vector with the maximum values per feature, N is the number of observed feature vectors in the matrix, and $\max_{i=1}^{N}$ gives the maximum over all rows. The second step in the normalization procedure emphasizes dominant features, inspired by text-retrieval approaches [25]. The dominance of each feature is captured in the weight vector $\mathbf{w}$, which is calculated using the variance over all observations:

$$\mathbf{w}(j) = \frac{1}{m_j} \log_2 \left(\frac{1}{m_j} \operatorname*{std}_{i=1}^{N} \bar{\mathbf{F}}(i,j) + 2 \right) \tag{5}$$

$$m_j = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{F}}(i,j) \tag{6}$$

where $\mathrm{std}_{i=1}^{N}$ calculates the standard deviation over all rows. The feature dominance is then used to reweigh $\mathbf{F}$, so that more dominant features get emphasized:

$$\hat{\mathbf{F}}(i,j) = \bar{\mathbf{F}}(i,j) \cdot \mathbf{w}(j) \tag{7}$$

The dominance weighting makes the object descriptors more robust to changes in viewpoint and illumination conditions, allowing more robust object recognition. Over time, when more object are learned by the system, a better selection of the more dominant features can be made, since there is more data available for the statistical analysis of the feature matrix. Similar process also appears in human working memory, since one important characteristic of learning is pruning only the most relevant information that will be stored in the long term memory [1].

D. Novelty detection

For a given observation of the object, the system needs to be able to decide whether it is a novel object or not. This is done by a learned novelty classifier. First, the feature vector extracted from the new observation, $\mathbf{g}$ is normalized based on the stored normalization vectors $\mathbf{w}$ and $\mathbf{m}$:

$$\hat{\mathbf{g}}(j) = \mathbf{g}(j) \cdot \frac{\mathbf{w}(j)}{\mathbf{m}(j)} \tag{8}$$

The normalized feature vector $\hat{\mathbf{g}}$ is then matched to the stored dominance-weighted feature matrix, $\hat{\mathbf{F}}$, and the distance to each of the feature vectors in the matrix is determined using the L1 distance:

$$\mathbf{d}(i) = \sum_{j=1}^{D} \|\hat{\mathbf{g}}(j) - \hat{\mathbf{F}}(i,j)\| \tag{9}$$

Using the distance vector $\mathbf{d}$, the M best matches are found and used to calculate an average distance value, v:

$$v = \frac{1}{M} \sum_{i=1}^{M} \mathbf{d}'(i) \tag{10}$$

where $\mathbf{d}'$ is the distance vector sorted in ascending order. Based on the distance value v, the newly observed object is classified as novel when:

$$v > \mu_{\mathrm{n}} + \sigma_{\mathrm{n}} \tag{11}$$

To obtain the values for μ_{n} and σ_{n}, the novelty classifier is continuously learned and updated based on the intra-object feature distances. These distances are obtained at the end of each object-manipulation sequence, by acquiring one more observation and calculating the average L1 distance between that last observation and the other observations in the sequence, similar to Eq. 10. For a given object k, this results in the intra-object feature distance, v_k. Using the values for all objects, $\{v_1, \ldots, v_K\}$, the novelty classifier fits a Normal distribution to obtain the mean μ_{n} and standard deviation σ_{n}.

Over time, when more objects are learned by the system, the classification performance will improve, since a more accurate estimation of the classification boundary can be made. If the object under observation is classified as novel, the observation sequence is used to learn the object. If the object is classified as known, the sequence is used to recognize the object. Both processes are described in the following two subsections.

E. Incremental learning of objects

Initially, the system has no knowledge of the objects or the environment. This knowledge is formed incrementally throughout the manipulation of the objects. Using the attention module (Sect. II-A), the object(s) are localized. Next, when the object is moved from the initial position, the smooth-pursuit module (Sect. II-B) tracks the object, and a total of M different observations of the object are segmented from the background. For every segment, the combined descriptor (Sect. II-C) is extracted online and stored in working memory. Next, the novelty detector (Sect. II-D) classifies the object as novel or known.

When the object is classified as novel, the observations of the object are added to the long-term memory. To do so,

Fig. 4: Examples of learned viewpoints

Fig. 5: Subset of tested objects

the M new feature vectors are added as new rows to the original feature matrix, $\mathbf{F}$. Including the new observations, the new dominant features are calculated and the matrix is normalized as given by Eq. (3) to Eq. (7). The intra-object feature distance for the novel object is furthermore calculated and used to update the novelty classifier, as explained in Sect. II-D.

F. Active recognition of objects

When the novelty detector classifies the object under inspection as known, the system will attempt to recognize the object using the sequence of M observations obtained during the manipulation of the object.

We use a voting scheme for recognition. For each observation i, the dominance-weighted feature vector, $\hat{\mathbf{g}}_i$, is obtained as in Eq. (8). The vector is matched with the feature matrix $\hat{\mathbf{F}}$ stored in the long-term memory, and the distance vector $\mathbf{d}$, giving the distance measures of $\hat{\mathbf{g}}_i$ to each of the stored feature vectors in the matrix, is obtained as in Eq. (9). Next, the M best matching feature vectors in the matrix are found and each of these vectors cast a vote for the object they are associated with. This process is repeated for each of the observations in the sequence, resulting in $M \times M$ votes for objects. The object with the most votes is returned as the recognized object.

III. Experimental Setup

In this section, we explain the experimental setup we adopted to test the performance of our system, and describe the dataset. Further, we show how we applied the SIFT descriptor [13] in order to compare the performance with the proposed multi-modal description module.

A. Learning and recognition approaches

To show the benefits of object manipulation, we test the proposed incremental learning (IL) and active recognition (AR), and compare it to passive learning (PL) and active learning (AL). The difference is that in incremental learning and active recognition, $M = 10$ observations of the object are acquired resulting from object manipulation. In passive learning and passive recognition, the objects are described from only a single observation $M = 1$. In the case of passive learning, due to the insufficient amount of data, we cannot calculate the intra-object feature distance. We therefore replaced Eq. (11) by $v > 10 \cdot \bar{v}$, where $\bar{v}$ is the inter-object feature distance, that is, the average distance between the feature vectors of all objects in the long-term memory. The multiplier was empirically established to give optimal novelty-classification results.

B. Dataset

In testing of our system, we used used 40 objects in total, 20 learned objects and 20 unknown. Examples of the objects from the database are shown in Fig. 5 and include both textured and uniformly colored objects. The objects are first learned online in settings with a single object and daylight conditions. For each experiment, we further test the recognition performance of each approach on 100 different scenes with 1-4 objects present in them. As will be shown in the result section, to investigate the robustness of the system we performed tests under significant illumination changes as well as in the cluttered scenes.

C. Applying the SIFT descriptor

To show that our description module can be implemented using other visual descriptors as well, we test it with the Scale-Invariant Feature Detector (SIFT) [13]. SIFT keypoints are detected in the object segment and described using the SIFT descriptor. For each observation of the object, this results in a set of descriptors, which are stored together as a *set descriptor* of the object. When matching a current observation with the stored descriptors in long-term memory, the similarity is measured by the maximum number of matching keypoints with any of the stored set descriptors. Keypoint matching is performed as described in [13]. Object matching and novel detection is done similar to what is described in Sect. II-D, with the difference that v in Eq. (10) is replaced by, u, the average of the M best keypoint-matching similarity measures. Furthermore, since the direction is reversed – higher values now indicate a better match – the novelty threshold (Eq. (11)) is changed to $u < \frac{1}{4} \cdot \bar{u}$, where $\bar{u}$ is the inter-object similarity measure, and the multiplier is empirically determined to get optimal novelty detection.

We compared the performance of our description module with SIFT for all tested learning and recognition approaches. Experimental results are provided in the next section.

TABLE I: The results of object recognition in conditions of uniform illumination

Combined descriptor	Precision(%)	Recall(%)	F measure(%)
IL-AR	98.00	100.00	98.99
IL-PR	100.00	100.00	100.00
PL-AR	85.71	97.67	91.33
PL-PR	68.75	94.29	79.50
SIFT	Precision(%)	Recall(%)	F measure(%)
IL-AR	81.25	97.50	88.64
IL-PR	68.33	93.33	78.90
PL-AR	64.86.	64.86	64.86
PL-PR	68.57	61.54	64.87

TABLE II: The results of classification of unknown objects

Combined descriptor	Precision(%)
IL-AR	92.00
IL-PR	66.00
PL-AR	46.00
PL-PR	44.00
SIFT	Precision(%)
IL-AR	82.00
IL-PR	78.00
PL-AR	58.00
PL-PR	40.00

TABLE III: Results of the object recognition at significant illumination variations

Combined descriptor	Precision(%)	Recall(%)	F measure(%)
IL-AR	90.00	100.00	94.73
IL-PR	80.00	97.50	87.64
PL-AR	66.00	97.50	78.57
PL-PR	56.00	100.00	71.57
SIFT	Precision(%)	Recall(%)	F measure(%)
IL-AR	68.00	100.00	80.95
IL-PR	53.33	82.76	64.86
PL-AR	59.46.	62.86	61.11
Pl-PR	48.94	88.46	63.01

IV. Experimental Results

In this section we provide detailed description of various experiments and give discussion on obtained results.

A. System performance in uniform illumination conditions

As a first experiment, we test the system's performance of recognizing single objects learned online. The testing is performed immediately after the objects were learned, so the illumination conditions remain approximately the same. The performance is given in Table I for both our combined descriptor and SIFT. One can conclude that for both descriptors, incremental learning significantly improves the system precision; around 15% compared with the passive learning approach. We can also see that in the case of single objects described with our combined descriptor both active and passive recognition give a very high precision. This proves that the calculated dominant features are very discriminative between different appearances of the object and confirms results from our previous analysis [23]. However, active recognition significantly improves the performance in the case of using the SIFT descriptor, since the exploration increases the number of matched keypoints. If we compare the performance of the SIFT descriptor and our combined descriptor, the combined descriptor shows much higher precision and recall rates. One of the reasons for this is that SIFT cannot detect weakly textured objects.

B. System performance in classification of novel objects

As a second experiment, we measure the system's ability to correctly classify unknown objects, in the same settings as for the first experiment. The results are given in Table II. The learned threshold for incremental learning gives best result in this case: 92%, followed by the actively learned threshold for SIFT keypoints: 82%. This test depicts that an adaptive threshold learned from observations greatly improves performance over fixed thresholds. It also shows that it is very difficult to conclude just from a single viewpoint that an object is novel; so active recognition is beneficial in such a case.

C. System performance in the variable illumination conditions

In a third experiment, we tested the system's ability to work in challenging illumination conditions, which is often the case in real-world setups. We learned the objects in sunny natural light and tested it later in 4 different illumination conditions (cloudy natural light, combination of natural and halogen light, only halogen light, and sunny natural light). The results are displayed in Table III for both the combined descriptor and for SIFT. The incremental learning and active recognition approach give good results for the combined descriptor, with a precision of 90%. This can be explained by the fact that the dominant features are robust to illumination changes and that by using the active method, enough viewpoints of the object are acquired. However, passive learning and passive recognition obtain much lower values. The precision of SIFT significantly drops due to the illumination changes while the recall rates for both descriptors are not so much affected by the light conditions.

D. System performance in the presence of clutter

In a fourth experiment we test the system performance in the presence of clutter. Scenes contain multiple known and unknown objects (distractors), where the number of objects ranges from 2 to 4. The performance results are shown in Table IV. The best results are obtained for the incremental learning - active recognition approach for a combined descriptor, followed by the same method for a SIFT descriptor. The incremental learning brings an improvement of 20% compared to passive learning. From Table IV, one can also see that our combined descriptor outperforms the SIFT descriptor. However, higher results are noticed for precision than for recall. This is due to a wrong initial object localization caused by a high clutter in the scene in which case we obtained false negatives for all tested methods. In total around 12% of all mistakes in the last test are due to

TABLE IV: Object recognition results in cluttered scenes with both known and unknown objects

Combined descriptor	Precision(%)	Recall(%)	F measure(%)
IL-AR	97.44	80.28	88.03
Il-PR	80.58	66.40	72.81
PL-AR	78.72	59.20	67.58
PL-PR	73.40	57.50	64.49
SIFT	Precision(%)	Recall(%)	F measure(%)
IL-AR	81.36	78.05	79.67
IL-PR	68.32	61.06	64.49
PL-AR	60.19	59.62	59.90
PL-PR	59.41	57.69	58.48

a false segmentation. The main reason for this lies in the fact that if there is an uniformly colored object placed next to several textured objects, it will be averaged out by the spectral residual operation. The object localization method is very precise in the case of 1 or 2 objects in the scene (almost 100%), while the precision rates drop to 85% for 3 objects and 80% for 4 objects in the scene.

E. Performance dependence on the number of learned observations and dominance weighting

As a fifth experiment, we measure the system dependence on the number of learned object observations. We test recognition of single objects learned online in similar illumination conditions as in the first experiment. In total, we tested 20 objects, for which the number of learned viewpoints per object is varied from 1 to 20. For each case, we calculate the precision and recall of the system. To show the influence of the dominant features, we have compared the performance of the combined descriptor with and without the weighting step (7). From the performance graphics depicted in Fig. 6 and Fig. 7, we can observe that the performance of the system increases with the number of observations. Using our dominance weighting, results in significant better performance than when the raw feature vectors are used. For the dominance-weighted descriptor, maximum performance is already achieved at 10 observations. The raw descriptor also benefits from multiple observations, but more then 20 observations are necessary to reach maximum performance. These results motivate our setting of $M = 10$.

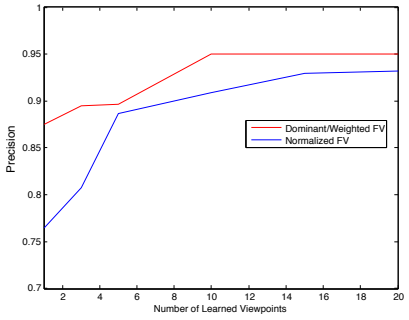

Fig. 6: System precision as a function of the number of observations used in learning.

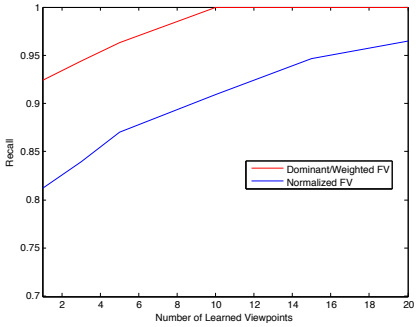

Fig. 7: System recall as a function of the number of observations used in learning.

TABLE V: Multi modal analysis of different components of the combined descriptor

Descriptor	Precision(%)	Recall(%)	F measure(%)
Color	94.44	89.47	91.89
Texture	41.67	76.92	54.05
Edge	57.14	1	72.73
Combined descriptor	95.00	1	97.44

F. Multi-modal descriptor analysis

In the final experiment, we test the contributions of different modalities of the combined descriptor – color, texture, and shape – to the overall performance. We use the same settings as in the first experiment, and test the recognition performance in the case of uniform illumination conditions and $M = 10$ learned observations per object. The performance is tested for each of the different feature modalities and compare to the performance using the combined descriptor. For each separate feature modality the dominance-weighted feature vector is obtained as in Eq. (5) to enhance the dominant features. It can be observed from the results in Table V that the largest contribution and the highest precision is from the color descriptor, followed by the edge descriptor. However, by combining all the components and calculating the dominant features, the performance of the descriptor is significantly improved, and all the good characteristics of separate components are inherited.

V. DISCUSSION

In this paper we presented an unifying approach for incremental learning and recognition of objects in an unstructured environment based on a model of early cognition. Inspired by how young infants learn objects, we focus on object learning in an artificial cognitive system, using a caregiver to demonstrate objects by manipulating them.

Based on a cognitive model for object learning and recognition, our system consists of several modules. At first, the visual-attention module is deployed to detect salient regions in the scene and localize unknown objects. The smooth-pursuit module is then utilized to track the object and obtain multiple observations. For the object description, a multi-modal descriptor is used which combines color, texture, and shape information and improves robustness through the emphasize on dominant features. Finally, using the sequence of

observations, the novelty detector classifies the object under inspection as novel or know, in which case the object is either learned and added to the long-term memory, or is recognized through matching with stored object representations.

The proposed approach was extensively tested and compared to a passive learning and passive recognition approach under challenging illumination settings and in cluttered scenes. The main conclusion from these experiments is that our system obtains good recognition performance in the different experimental conditions. Comparing the active methods to passive methods, we can conclude that object manipulation during learning and recognition greatly boosts recognition performance and the classification of novel objects.

The main benefit of object exploration is the fact that through the manipulation, a variety of different appearances of the object is obtained, which boosts the object description, and allows to reliable gather statistics to find the dominant features and to learn the novelty classifier. The motion furthermore improves the detection and segmentation of the object.

When comparing to a description of the objects using SIFT [13], we observe an improvement in performance using our multi-modal and dominance-weighted descriptor. However, the system also benefits from object exploration when using the SIFT descriptor. Our system shows a large improvement when the features in descriptor are weighted according to dominance. Investigations of the contribution of the different feature modalities in the descriptor reveal an improvement of the combined descriptor over single feature modalities.

Although the exploration in this paper was carried out by a caregiver, the proposed approach is directly applicable when the robot independently starts to explore objects. Inspired by the next step in the development of infants, we are currently extending our system to take the step from a supervised setting to fully autonomous object exploration and learning.

VI. ACKNOWLEDGMENTS

This research was sponsored by the Dutch government through the Point One project PNE09003 (Bobbie). This work was also supported by the EU through the project eSMCs, IST-FP7-IP-270212 and the Swedish Foundation of Strategic Research.

REFERENCES

[1] A. Baddeley and S. D. Sala, "Working memory and executive control," *Philosophical Transactions: Biological Sciences, Royal Society London*, vol. 351, no. 1346, pp. 1397–1403, 1996.

[2] M. Björkman and D. Kragic, "Active 3d scene segmentation and detection of unknown objects," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[3] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[4] D. Comaniciu, P. Meer, and S. Member, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[5] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philosophical transactions of the royal society: Mathematical, Physical and Engineering sciences*, vol. 361, p. 2003.

[6] C. Goerick, H. Wersing, I. Mikhailova, and M. Dunn, "Peripersonal space and object recognition for humanoids," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, pp. 387 –392, 2005.

[7] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR07)*, pp. 1–8, 2007.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] R. Kestenbaum, N. Termine, and E. S. Spelke, "Perception of objects and object boundaries by three-month-old infants," *British Journal of Developmental Psychology*, vol. 5, pp. 367–383, 1987.

[10] E. I. Knudsen, "Fundamental components of attention," *Annual Review of Neuroscience*, vol. 30, pp. 57–78, 2007.

[11] G. Kootstra, N. Bergström, and D. Kragic, "Using symmetry to select fixation points for segmentation," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2010.

[12] G. Kootstra, J. Ypma, and B. de Boer, "Active exploration and keypoint clustering for object recognition," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Pasadena, CA: IEEE, pp. 1005–1010, 2008.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust side-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[15] D. Meger, M. Muja, S. Helmer, A. Gupta, C. Gamroth, T. Hoffman, M. Baumann, T. Southey, P. a. Fazli, W. Wohlkinger, P. Viswanathan, J. J. Little, D. G. Lowe, and J. Orwell, "Curious george: An integrated visual search platform," in *Proceedings of the 2010 Canadian Conference on Computer and Robot Vision*, ser. CRV '10. Washington, DC, USA: IEEE Computer Society, pp. 107–114, 2010.

[16] J. Ning, L. Zhang, D. Zhang, and C. Wu, "Robust object tracking using joint color-texture histogram," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 7, pp. 1245–1263, 2009.

[17] F. Orabona, G. Metta, and G. Sandini, *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, L. Paletta and E. Rome, Eds. Berlin, Heidelberg: Springer-Verlag, 2008.

[18] J. Piaget, *The Grasp of Consciousness: Action and Concept in the Young Child*. Cambridge, Mass.: Harvard University Press., 1976.

[19] N. Pugeault and N. Krüger, "Temporal accumulation of oriented visual features," *Journal of Visual Communication and Image Representation*, vol. 22, pp. 153–163, 2011.

[20] D. H. Rakison and G. Lupyan, "Developing object concepts in infancy: An associative learning perspective," *Monographs of the Society for Research in Child Development*, vol. 73, no. 1, pp. vii, 1–110, 2008.

[21] B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic, "An active vision system for detecting, fixating and manipulating objects in the real world," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 133–154, 2010.

[22] J. Richards, "The development of visual attention and the brain," in *The cognitive neuroscience of development*, M. D. H. . M. Johnson, Ed. East Sussex, UK: Psychology Press, 2003.

[23] M. Rudinac and P. P. Jonker, "A fast and robust descriptor for multiple-view object recognition," in *Proceedings of the International Conference on Control Automation Robotics & Vision (ICARCV)*, pp. 2166–2171, 2010.

[24] M. Rudinac and P. P. Jonker, "Saliency detection and object localization in indoor environments." in *Proceedings of the International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 404–407, 2010.

[25] Y. Rui, T. S. Huang, and S. Mehrotra, "Content-based image retrieval with relevance feedback in mars," in *IEEE International Conference on Image Processing*, p. 815818, 1997.

[26] D. Tax, "One-class classification," phd, Delft University of Technology, Delft, June 2001.

[27] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[28] D. Vernon, C. von Hofsten, and L. Fadiga, *A Roadmap for Cognitive Development in Humanoid Robots*, 1st ed., ser. Cognitive Systems Monographs. Springer-Verlag Berlin Heidelberg, vol. 11, 2011.