International Journal of Neural Systems, 2007. in press.

Online Learning of Objects in a Biologically Motivated Visual Architecture

Heiko Wersing[1], Stephan Kirstein[1], Michael Götting[2], Holger Brandl[2], Mark Dunn[1], Inna Mikhailova[1], Christian Goerick[1], Jochen Steil[2], Helge Ritter[2], Edgar Körner[1]

[1] Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach/Main, Germany
[2] University of Bielefeld, Faculty of Technology PO Box 100131, D-33501 Bielefeld, Germany

July 3, 2007

Abstract

We present a biologically motivated architecture for object recognition that is capable of online learning of several objects based on interaction with a human teacher. The system combines biological principles such as appearance-based representation in topographical feature detection hierarchies and context-driven transfer between different levels of object memory. Training can be performed in an unconstrained environment by presenting objects in front of a stereo camera system and labeling them by speech input. The learning is fully online and thus avoids an artificial separation of the interaction into training and test phases. We demonstrate the performance on a challenging ensemble of 50 objects.

1 Introduction

The human visual system shows an outstanding capacity for learning and robust recognition of numerous objects, at a level far superior to all currently existing technical recognition approaches. A particular feature of human object perception is the capability of quickly analyzing and remembering completely unknown new objects. We refer to this ability in this contribution as *online learning*, which is of high relevance for cognitive robotics and computer vision. A typical application domain is to increase the knowledge of an assistive robot in a changing and unpredictable environment [1, 2]. The capability of learning online constitutes a fundamental difference to offline learning, since it enables

an interactive process between teacher and learner. The immediate feedback about the current learning state can induce an instantaneous and active learning process that reduces the amount of necessary training data and allows an iterative error correction based on user feedback.

In order to achieve online learning of many complex-shaped objects, we present a system combining a flexible neural object recognition architecture with a biologically motivated attention system for gaze control, and a speech understanding and synthesis system for intuitive interaction. The target is to obtain a flexible object representation system that is capable of high-performance appearance-based object recognition of complex objects together with a particularly rapid online learning scheme that can be carried out by cooperative training with a human teacher. A high level of interactivity is achieved by avoiding an artificial separation into training and testing phase, which is still the state-of-the-art for most current trainable object recognition architectures. We do this by using an incremental learning approach that consists of a two-stage memory architecture comprising a context-dependent sensory memory and a persistent object memory that can also be trained online.

Previous approaches to fast interactive object learning often had to resort to simple histogram-based object representations [3], or strong assumptions on the environment for figure-background separation [4]. We relax many of these constraints and do not impose any preconditions on the environment, except that objects are presented to the system by showing them by hand. To allow online learning in this difficult scenario, we use a dynamic segmentation approach that performs a fast figure-ground separation based on an initial stereo-based coarse object hypothesis. The object recognition architecture is motivated from the ventral pathway of the human visual cortex and can be applied to arbitrary complex-shaped objects. Fast online learning can be achieved with this architecture, because object-specific learning occurs only on the highest levels of the hierarchical feature detection stages. The lower stages of the model correspond to earlier and intermediate feature detection stages in the visual cortex and are trained by sparse coding learning rules [5]. This results in a particularly robust appearance-based representation of objects using a consistent library of typical local shape elements. As was shown recently by Serre et al. [7] for a related, but more biologically detailed model, such visual representation architectures achieve a highly competitive recognition and detection performance on current computer vision benchmarks for offline learning.

In the following we review related work in Section 2 and give an overview over our system in Section 3. In Section 4 we describe the components of the visual memory in more detail and show results on the performance and learning behavior in Section 5. We give a discussion in Section 6 and conclude with Section 7.

2 Related Work

Although offline training of model-free object recognition architectures has become an established technique in pattern recognition and applications, only few work has been done until now on online learning for complex-shaped objects. The main problems are poor generalization due to the inherent high dimensionality of visual stimuli, and the difficulty to achieve incremental online learning with standard classifier architectures like multi-layer perceptrons or support vector machines.

To make online learning feasible, the complexity of the sensorial input has been reduced to simple blob-like stimuli [8], for which only positions are tracked. Based on the positions, interactive and online learning of behavior patterns can be performed. A slightly more complex representation was used by Garcia et al. [9], who have applied the coupling of an attention system using features like color, motion, and disparity with a fast learning of visual structure for simple colored geometrical shapes like balls, pyramids, and cubes.

Histogram-based methods are another common approach to tackle the problem of high dimensionality of visual object representations. Steels & Kaplan [3] have studied the dynamics of learning shared object concepts based on color histograms in an interaction scenario with a dog robot. Another model of word acquisition that is based on multidimensional receptive field histograms for shape representation and color histograms was proposed by Roy & Pentland [10]. The learning proceeds online by using a short-term memory for identifying reoccurring pairs of acoustic and visual sensory data, that are then passed to a long-term representation of extracted audiovisual objects.

Arsenio [11] has investigated a developmental learning approach for humanoid robots based on an interactive object segmentation model that can use both external movements of objects by a human and internally generated movements of objects by a robot manipulator. Using a combination of tracking and segmentation algorithms the system is capable of online learning of a few objects by storing them in a geometric hashing representation.

Bekel et al. [12] proposed an approach to supervised online learning for object recognition, consisting of three stages of vector quantization, local PCA, and a local linear map classifier. The image acquisition of new object views is triggered by pointing gestures on a table, and is followed by a short training phase, which takes some minutes. The main drawback is the lack of an incremental learning mechanism to avoid the complete retraining of the architecture. The approach has been integrated in a larger architecture for cognitive vision [13].

Li et al. have presented a system for interactive object learning on a mobile robot that features an elaborated multi-modal dialogue system to enable context-dependent attention selection using speech references made by the user [2]. Pointing gestures can be used in combination with speech to perform a color-based segmentation of objects to be learned. The integration of a classifier for actually performing object learning was, however, not yet accomplished.

Roth et al. developed an online learning system for the task of person detection on surveillance camera images [14]. The system employs a reconstructive model using incremental principal component analysis for autonomously selecting positive examples for an online AdaBoost classifier. The same incremental online AdaBoost was also combined with an adaptive tracking model for the incremental learning of hand-held objects with limited pose variation [15]. In both settings a static background was assumed and used for object segmentation.

Kirstein et al. [4] have presented an online learning architecture that is operated in a more constrained scenario with defined black background to ease the figure-ground segmentation. Their focus was the transfer from a short-term to more condensed long-term memory representation using incremental vector quantization methods.

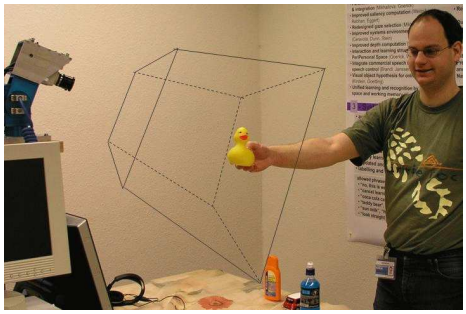

Figure 1: Typical training situation. An object is presented within the peripersonal space and can be trained or recognized.

3 System Overview

We first describe an overview of the system (see Figure 2) and its key components, before we give more details in Section 4.

We use a stereo camera head mounted on a pan-tilt unit, which delivers left and right image pairs as the visual input. The gaze control of the head is driven by an independent circuit that combines the cues of motion, color, and depth for attention-driven selection of the gaze direction. The concept of peripersonal space [16] is used to establish shared attention on a presented object during learning. This means that the system will focus its attention on an object that is presented within a particular short-distance range interval that roughly corresponds to the biological concept of the manipulation space around the body (see Figure 1). If nothing is present within this space, the cues of motion and color/intensity determine the gaze selection of the system. All cues are based on retinotopic activation maps, and we induce a higher priority for motion detection with a higher weight of the corresponding map. (see [16] for more details). A typical sequence of interaction thus consists of first catching the system attention by waving, which centers the gaze direction towards the interacting person. In the second step an object can be brought sufficiently close to the camera to induce learning or recognition of the attended object in the peripersonal space.

The online learning system is working with the camera output that is generated according to the gaze selection of the independent attention system. Based on the current stereo view pair, a depth map is computed that is aligned with the left camera image. The left camera image and the depth map are passed to the peripersonal blob detection stage that generates a square region of interest (ROI), based on the estimated distance of the current object hypothesis. Using the distance, the apparent size of objects within the ROI can be normalized with remaining uncertainties due to the limited precision of the depth computation. The square ROI with distance-dependent size in the original image is scaled to a size of 144x144 pixels. The gaze selection and size normalization remove largely the translation and scale variance inherent to the unconstrained recognition task.

The normalized ROI around the object hypothesis together with the corresponding part of the depth map is passed to the figure-ground segmentation stage of processing, the adaptive scene-dependent filters (ASDF) [17]. The ASDF method makes no strong assumptions on the objects like e.g. being single-colored. Based on the depth map, a relevance map is obtained that covers the

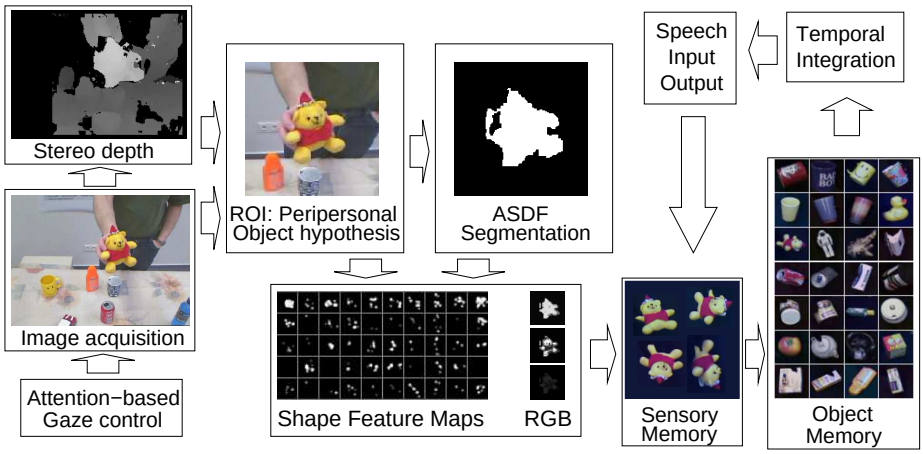

Figure 2: Overview of the visual online learning architecture. Based on the depth estimation, an object is selected and segmented. The segment is passed through the visual feature hierarchy and subsequent views of the current context are stored in the sensory memory. Transfer to the object memory is guided by speech-based feedback.

object only coarsely with considerable overlap to the background. For each pixel location in the ROI, a local feature vector is computed based on RGB color channels, depth, and pixel position. Using a dynamic vector quantization model, first an unsupervised segmentation is computed using the local feature vectors in the ROI as input ensemble. Then the input image is segmented according to the mapping to the Voronoi cells of the found vector quantization centers. Due to a sufficient number of centers, we obtain an oversegmentation and can then select object segments as those that are sufficiently contained within the relevance map (see [17] for more details). The method obtains an intrinsic stability by continuously iterating the vector quantization based on state of the previous frame. We additionally use skin color detection [18] to remove parts of the hand that hold the object. The output of the ASDF stage is a mask describing the current figure-ground hypothesis on the ROI.

The selected ROI and the segmentation mask from the ASDF stage are fed into the model of the ventral visual pathway of Wersing & Körner [5] to obtain a complex feature map representation that is based on 50 shape and 3 color feature maps. The color channels are downsampled images in the three RGB channels. The output is a high-dimensional view-based representation of the input object which is passed to the higher object memory representation stages for learning and recognition.

To allow a particularly interactive online learning, we use a memory concept that is separated into a sensory memory carrying the currently attended object and a persistent memory that carries consolidated and consistently labeled object view representations. As long as an object is presented within the peripersonal space and has not been labeled or confirmed, the obtained feature map representations of views are stored incrementally within the sensory memory. At the same time, all newly appearing views are being classified using the persistent object memory. If the human teacher remains silent, then the system will either generate a class hypothesis or reject the presented object as unknown and verbalize this using the speech output module. The human teacher can con-

firm the hypothesis or make a new suggestion on the correct object label. As soon as feedback by the teacher is available, the learning architecture starts the concurrent transfer from the sensory memory buffer into the consolidated object memory. This extends over the whole history of collected views during the presentation phase and also proceeds with all future views, as long as the object is still present in the peripersonal space. The labeling of the current object can be done by the teacher at any time during the dialogue and is not restricted to being a reaction on a class hypothesis of the recognition system. The concept of a context-dependent memory buffer avoids a separation into training and testing phases. The transfer from the sensory to the object memory is sufficiently fast to remain unnoticed to the human trainer and the learning success can be immediately tested, allowing for a real online learning interaction.

The speech input and output is very important for the intuitive training interaction with the system. We use a system [19] with a headset, which is the current state-of-the-art for speaker-independent recognition. The vocabulary of object classes is specified beforehand. To be able to label arbitrary objects we also use wildcard labels such as "object one", "object two" etc.

4 Object Memory Representation

In the following we describe in more detail the main components of the object memory and recognition system. For a more detailed description of the attention, gaze selection and stereo processing system we refer the reader to [16].

4.1 Hierarchical Feature Processing

The output of the ASDF figure-ground segmentation stage is a binary mask signal $\mathbf{m}^{\text{seg}}$ that is combined with the candidate ROI image $\mathbf{I}$ (of size 144x144 pixels) and fed into the hierarchical model of the ventral visual pathway developed by Wersing & Körner [5]. To obtain invariance against rotations in the image plane, which is normally a problem for appearance-based recognition, we determine the principal axes of the figure-ground mask and rotate the ROI and mask aligned with the horizontal direction. This normalization introduces much better robustness for the recognition of elongated objects like e.g. bottles.

The rotation-normalized ROI is processed using a hierarchy of feature detection and pooling stages that achieves a robust appearance-based representation of an object view as a collection of several sparsely activated feature map representations (see Figure 3). Starting from an RGB input color image $\mathbf{I}_i = (\mathbf{I}_i^R, \mathbf{I}_i^G, \mathbf{I}_i^B)$, we compute an intensity image $\mathbf{I}'_i = 1/3\ \mathbf{I}_i^R + 1/3\ \mathbf{I}_i^G + 1/3\ \mathbf{I}_i^B$. The first feature-matching stage S1 consists of four orientation-sensitive odd Gabor filters, a Winner-Takes-Most competition between features at the same position and a final threshold function. We adopt the notation that vector indices run over the set of neurons within a particular feature plane of a particular layer. To compute the response $q_1^l(x, y)$ of a simple cell in the first layer S1, responsive to Gabor type l at position (x, y), first the image vector $\mathbf{I}'$ is multiplied with a Gabor filter $\mathbf{w}_1^l(x, y)$, and pointwise multiplied with the binary segmentation mask $m^{\text{seg}}(x, y) \in \{0, 1\}$:

$$q_1^l(x, y) = |\mathbf{w}_1^l(x, y) * \mathbf{I}'| \cdot m^{\text{seg}}(x, y). \tag{1}$$

The inner product is denoted by $*$, i.e. for a 144×144 pixel image $\mathbf{I}$ and $\mathbf{w}_1^l(x, y)$ are 20736-dimensional vectors. We apply the masking after the edge detection,

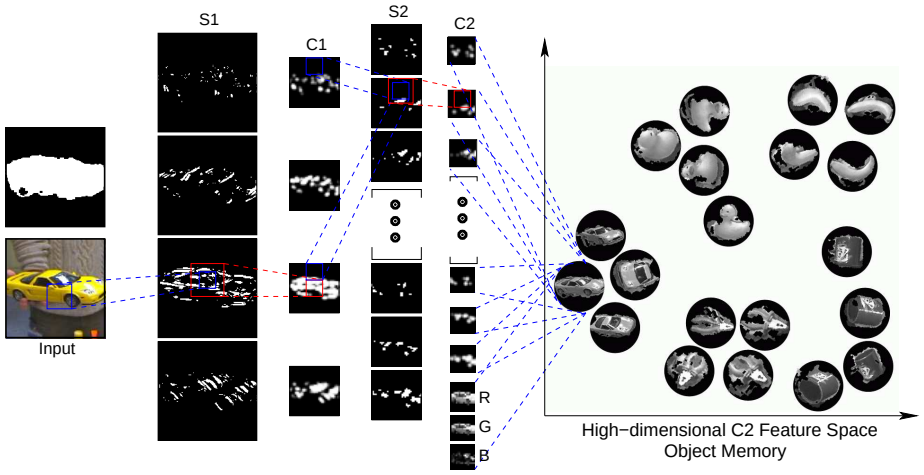

Figure 3: Hierarchical object representation and object memory. Based on a ROI with additional segmentation mask, the input is processed in a sequence of topographically organized feature detection (S1,S2) and pooling stages (C1,C2). The object memory provides an exemplar-based representation of views embedded in the high-dimensional C2-feature space.

to avoid the occurence of spurious edges at wrong segmentation borders. In a second step, a soft Winner-Takes-Most (WTM) mechanism is performed with

$$r_1^l(x,y) = \begin{cases} 0 & \text{if } \frac{q_1^l(x,y)}{M} < \gamma_1 \text{ or } M = 0, \\ \frac{q_1^l(x,y) - M\gamma_1}{1-\gamma_1} & \text{else,} \end{cases} \tag{2}$$

where $M = \max_k q_1^k(x,y)$ and $r_1^l(x,y)$ is the response after the WTM mechanism which suppresses sub-maximal responses. The parameter $0 < \gamma_1 < 1$ controls the strength of the competition. The activity is then passed through a simple threshold function with a common threshold θ_1 for all cells in layer S1:

$$h_1^l(x,y) = \mathrm{H}\big(r_1^l(x,y) - \theta_1\big), \tag{3}$$

where $\mathrm{H}(x) = 1$ if $x \geq 0$ and $\mathrm{H}(x) = 0$ else and $h_1^l(x,y)$ is the final activity of the neuron sensitive to feature l at position (x,y) in the S1 layer. The activities of the first layer of pooling C1-cells are given by

$$c_1^l(x,y) = \tanh\big(\mathbf{g}_1(x,y) * \mathbf{s}_1^l\big), \tag{4}$$

where $\mathbf{g}_1(x,y)$ is a normalized Gaussian pooling kernel with width σ_1, identical for all features l, and tanh is the hyperbolic tangent function. From S1 to C1 we perform a four-fold resolution reduction in x and y directions.

The features in the intermediate layer S2 are sensitive to local combinations of the features in the planes of the previous layer, and are thus called *combination cells* in the following. We use 50 features that were trained using a sparse coding unsupervised learning approach (see [5]), and which provide an efficient representation of the combined local edge feature responses. We introduce the layer activation vectors as $\bar{\mathbf{c}}_1 = (\mathbf{c}_1^1, \ldots, \mathbf{c}_1^K)$, $\bar{\mathbf{w}}_2^l = (\mathbf{w}_2^{l1}, \ldots, \mathbf{w}_2^{lK})$ with K=4. Here $\mathbf{w}_2^{lk}(x,y)$ is the receptive field vector of the S2 cell of feature l at position (x,y), describing connections to the plane k of the previous $C1$ cells. The combined linear summation over previous planes is then given by

$q_2^l(x,y) = \bar{\mathbf{w}}_2^l(x,y) * \bar{\mathbf{c}}_1$. After the same WTM procedure with strength γ_2 as in (2), the activity in the S2 layer is given by $h_2^l(x,y) = H(r_2^l(x,y) - \theta_2)$ after thresholding with a common threshold θ_2. The step from S2 to C2 is analogous to (4) and given by $c_2^l(x,y) = \tanh(\mathbf{g}_2(x,y) * \mathbf{s}_2^l)$, with Gaussian spatial pooling kernel $\mathbf{g}_2(x,y)$ with range σ_2 and two-fold reduction in x and y dimension. The final resolution is 18x18 for each C2 feature map. As was shown before, the output of the feature representation of the C2 feature layer can be used for robust object recognition that is competitive with other state-of-the-art models, when offline training is being used [5]. The free parameters are chosen as $\gamma_1 = 0.9, \theta_1 = 0.3, \sigma_1 = 4, \gamma_2 = 0.9, \theta_2 = 0.75, \sigma_2 = 2$, according to the optimized choice evaluated in [5].

The efficiency of the representation is achieved by sparse coding ensuring that object views are represented using only sparse activation in the high-dimensional feature space. To represent also coarse color information, the 3 RGB channels are used as a downsampled ROI $\hat{\mathbf{I}}_i = (\hat{\mathbf{I}}_i^R, \hat{\mathbf{I}}_i^G, \hat{\mathbf{I}}_i^B)$ at the same resolution of 18x18 as the shape features. We denote the combined color and shape feature map output as $\mathbf{x}_i(\mathbf{I}_i) = (\mathbf{c}_2^1, \ldots, \mathbf{c}_2^{50}, \hat{\mathbf{I}}_i^R, \hat{\mathbf{I}}_i^G, \hat{\mathbf{I}}_i^B)$. Although the complete dimensionality of a single view representation $\mathbf{x}_i$ is thus (50+3)x18x18=17172, the effective dimensionality is much smaller, due to the sparsity of the representation vector and the confinement of activation to the figure-ground mask. Nevertheless it is a key feature of our biologically motivated visual processing model that robustness, generalization and speed of learning is not achieved by a dimension reduction as in most other current online learning models [8, 9, 3, 10, 11, 12, 14]. The key element is a transformation of the input into a sparse robust feature map representation that captures relevant locally invariant structures of the objects.

4.2 Sensory and Object Memory

The object representation system for online learning and recognition is separated into two subsystems: A sensory memory for temporarily remembering the currently attend object within focus and a persistent object memory that integrates all object knowledge incrementally over time.

The high-dimensional output vectors of the feature hierarchy are continuously stored within the sensory memory. The task of this memory is to capture all current views of an object to be able to use them for transfer to the object memory when a speech label has been given. This means that also those views can be used for training that were recorded before a labeling of the object was obtained from the human trainer, relaxing the constraints on the training dialogue. The sensory memory is realized as an incremental vector quantization model S, which consists of K representative vectors $\mathbf{s}_k \in S, k = 1, \ldots, K$. A new representative $\mathbf{s}_{K+1} = \mathbf{x}_i$ is added if the feature map output $\mathbf{x}_i(\mathbf{I}_i)$ of the current input image is sufficiently dissimilar to all current entries in the sensory memory: $||\mathbf{x}_i - \mathbf{s}_k|| > T_S$ for all k, where T_S is a similarity threshold. The similarity is measured based on Euclidean distance in the feature map vector space. Due to the sparsity of the feature map vectors the distance computation can be very efficiently implemented [4]. If the focus of attention is lost, because the object is retracted from the peripersonal space, the sensory memory is cleared.

When a labeling signal arrives, because the human teacher has named an object or has confirmed a hypothesis generated from the object memory, the information accumulated in the sensory memory is transferred to the object memory in real time. Here we use the same incremental vector quantization

model. We denote the object memory as a collection of individual object representations O_n for object n with M_n representatives $\mathbf{o}_l^n \in O_n, l = 1, \ldots, M_n$. If there are already some views available in the object memory, the comparison is performed against the already existing representation (see Figure 3). If the current object is labeled as object m, then for all the vectors in the sensory memory $\mathbf{s}_k \in S$, a new object representative $\mathbf{o}_{M_n+1}^m = \mathbf{s}_k$ is added when the sensory memory representative $\mathbf{s}_k$ is sufficiently dissimilar to all current entries in the object memory: $||\mathbf{s}_k - \mathbf{o}_l^m|| > T_O$ for all $l = 1, \ldots, M_m$, where T_O is the object similarity threshold. If the training continues after the labeling signal was received, and the object remains within the focus of attention, all following feature map inputs $\mathbf{x}_i$ are directly passed to the object memory according to the same dissimilarity criterion with threshold T_O.

The main advantage of the template-based representation is that training is fully incremental and non-destructive with regard to previous information. This representation can be later condensed and consolidated using additional learning mechanisms that operate on a slower time scale [4].

Every arriving view is being classified based on the information in the object memory using a nearest-neighbor classifier (NNC) based on the labeled representatives. The corresponding NNC class hypothesis m_i of view $\mathbf{x}_i$ is given by $m_i = \arg\min_n(\min_l ||\mathbf{x}_i - \mathbf{o}_l^n||)$. Since the system is running at a sufficient frame rate, we can use a temporal integration over different views to improve the classification results considerably. Our results have shown that a majority voting scheme is particularly efficient in combination with the nearest-neighbor classification approach in the object memory, since it allows to use more ensemble information of the exemplar-based representation stored in memory. In our experiments we use a history of 10 classifications and assign the output class that received most single classification votes. An object is rejected as unknown if this majority vote is less than 50% or if the mean similarity to the majority representatives, measured in the Euclidean feature space, is below a fixed threshold.

5 Results

The complete system has been realized on a cluster of one dual processor PC for gaze control and image capture, one desktop PC running the speech recognition and synthesis system, and one dual processor PC performing all visual processing and online learning after the gaze selection. The recognition system is running at a frame rate of roughly 6Hz, which enables interaction and online learning with direct feedback on the learning result. A generic training scenario is shown in Figure 1. We have selected a large object set containing 50 objects for our experiments, shown in Figure 4 with typical ROI views that are delivered from the attention system. During all experiments the objects were freely rotated by hand to obtain a strong appearance variation.

5.1 Interactive Training

We visualize the actual time course of the different memory types during a training session of 18 objects in Figure 5. The plot displays the number of used representatives in the sensory and object memories together with the training dialogue (abbreviated, the actual dialogue is a little more elaborate). Starting from a completely empty object memory, we first perform a training of 10 objects. In this first phase the system first consistently matches the cola can to the

Figure 4: Overview over the set of 50 objects used for training and testing. The objects were freely rotated, a changing background is obtained due to the gaze control fixating the objects.

previously trained "sun cream" object, and thus classifies the cola can initially as "sun cream", which is then corrected by the teacher. Due to the similar red-white color and shape composition the "mini car" is also first confused with the cola can, and is corrected. Due to the shape similarity the green bottle is first labeled as blue bottle, which is a reasonable error, as long as no correction signal is given. After the feedback by the teacher, the system has learned to discriminate the first 10 objects after 5 minutes of training from many different viewing angles, which is evaluated directly afterwards. In the second training phase 8 objects are added. The initial confusion occurs quite reasonably between cola can and a yellow can, another red car and the mini car, a new blue mug and the first blue patterned mug, and a new blue rubber duck and the initial yellow one. After the initial training in the second phase, the garlic press and police car object have to be additionally refined. After that second retraining phase, all 18 objects are classified from any reasonable viewing angle without further errors.

An important property of the system is that learning occurs most of the time and is not separated into artificial training and testing phases. This can be seen from the time course in Figure 5, where during the first evaluation of the first 10 objects between 320s and 420s the object memory is still expanding, due to the confirmation signals of the human teacher on the system classifications. The same applies to the second evaluation and error correction phase between 640s and 850s. The complete duration of the session until no further recognition errors are encountered is about 12 minutes. This highlights the gain in learning speed that can be achieved due to the active error correction process during learning. When the object memory is enlarged over time, we encounter a slight slowing down of the system frame rate from 6Hz to approximately 4Hz, since the comparison to the memory takes longer.

5.2 Recognition Performance

In Figure 6 we show plots of the recognition performance versus training time during online learning. For this evaluation we train 49 objects from our training

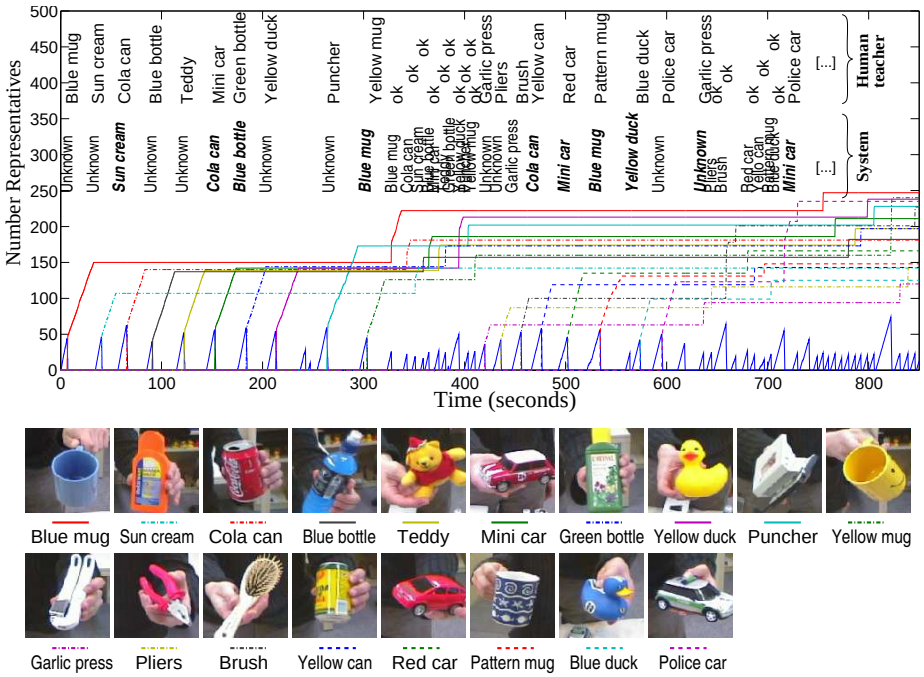

Figure 5: Temporal learning dynamics during a training session for 18 objects. The plot shows the number of representatives for the sensory memory ("saw-tooth" at bottom of plot) and representatives for each object in the object memory over time. The corresponding training dialogue is stated synchronously at the top. The top row states the given labels by the human trainer, while the bottom row gives the classification results of the system, before a human labeling is given. Errors of the system are printed in bold italics. From 0 to 310s the first 10 objects are trained, the recognition of these 10 objects is evaluated from 320s to 420s without any errors. From 420s to 730s another 8 objects are added, and all 18 objects are checked after 730s without errors.

set of 50 objects that was generated by storing 300 views per object from a typical training session. Then the 50th object is trained in steps of 10 images (1.67 sec in Figure 6) and a testing step is performed. The test is done by classifying a completely disjoint test set of 100 views per object that was collected using a different person. Test performance is measured over all 100 test images of the currently trained object giving the classification rate as percentage of correctly recognized objects at this point of online learning. Then training proceeds until all 300 training images are used. The plots in Figure 6 show the resulting classification rate, averaged over an ensemble of experiments, where each of the 50 objects was one time the final object.

We compare in Figure 6 the conditions of either using ASDF segmentation, ASDF segmentation with subsequent rotation normalization, and no segmentation. Each of the three settings is plotted with and without temporal integration with voting over a past history of 10 classifications. The results demonstrate that due to the cluttered background, training with the ASDF speeds up learning considerably and gives a signigicantly higher recognition rate. Performing the rotation normalization gives a further small gain in performance. The contribution of the temporal integration is much more substantial, and reduces the

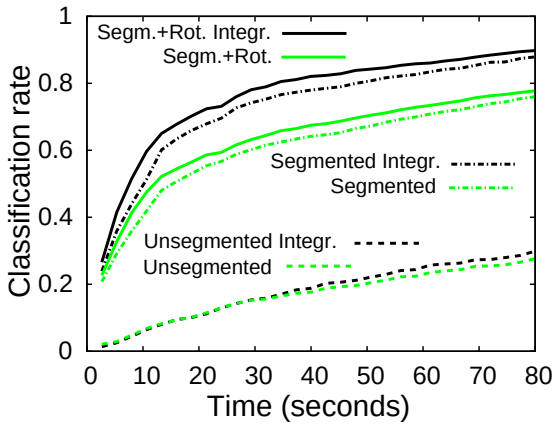

Figure 6: Recognition performance versus training time. The plot shows the average test performance for training the 50th object after 49 objects were already trained. We compare segmentation, rotation normalization, and unsegmented performance with and without temporal integration.

error rate to about one half for training times larger than 30 seconds for both segmented and rotation-normalized cases.

To investigate the scaling of the architecture with the number of objects, we show in Figure 7 a plot of the final performance after training for 80 seconds (corresponds to 300 training views), when we vary the number of objects from 5 to 50. Again we compare all the settings that were already described for the plot in Figure 6 and the qualitatively observable gains of segmentation and temporal integration are similar. For the best setup we obtain a slow decrease of classification rate from 100% for 5 objects till about 90% for 20 objects. From 20 to 50 objects the performance stays roughly at about 90% correct, with small fluctuations induced by the different difficulty levels of the objects. This shows that the representational capacity is large enough to capture 50 objects with their natural appearance variations.

6 Discussion

We have performed an extensive investigation of our online learning architecture using a large ensemble of 50 objects of various different shapes, colors and textures. Compared to previous approaches to online learning [11, 12, 15] which were only applied to smaller and limited object ensembles, we could demonstrate that the capacity of our object representation is sufficiently high to accommodate larger numbers of objects. This is caused by the high-dimensional embedding space of our object representation, contrary to other approaches using dimension reduction for generalization.

An interesting question is the degree of generalization over different environment and light conditions that is achieved by our model. We do not impose particular constraints where objects are presented apart from being within the peripersonal space around the camera head. This has the consequence that the overall illumination strength and the direction of light sources is varying in the object view data. From the observed performance we conclude that our principle of hierarchical feature representation like in the human ventral pathway of visual processing can deal with these modest variations in a robust way. We

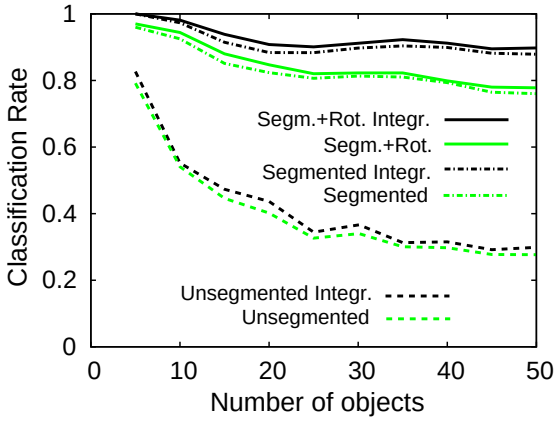

Figure 7: Recognition performance depending on the number of objects. The plot gives the recognition performance after 80 seconds of training for the test error of the n-th object after n-1 objects have already been trained. We again compare segmentation, rotation normalization, and unsegmented performance with and without temporal integration.

also observed that the online learning system can naturally cope with changes of environment that are sufficiently similar for training and testing, because the learned representation is collected consistently according to the present conditions. For the cases of strongly differing light situations between training and test we observed a graceful degradation of results, with strongly cast shadows posing the greatest problems for the appearance-based approach we are using.

The final representation for classification is exemplar-based and its complexity increases linearly with the number of training views seen by the architecture. Due to the sparsity of the representation the amount of memory necessary can be strongly reduced by representing only nonzero feature responses. Nevertheless, if we extend the number of classes by another order of magnitude, such an exhaustive storage becomes infeasible and reaches the limits of current standard computer memory systems. There is evidence that exemplar-based representations play an important role in visual object memory (see [20, 21] for reviews). This poses the question how an appropriate generalization can be obtained based on the available exemplars. Poggio & Bizzi [21] suggest a radial basis function-like tuning as a key mechanism of generalization. Kirstein et al. [4] have proposed an extended memory architecture, that implements a condensation of the representation into long-term memory by shifting the view representatives in the embedding space in order to minimize the classification error. This architecture was, however, not yet implemented for a real-time application.

The ability to perform online learning in direct interaction makes it possible to utilize human feedback during training for higher-level control of behavior. Goerick et al. [22] have integrated the object learning architecture described in this contribution in a system that autonomously learns new visual behaviors in interaction. The learning is governed by an internal needs dynamics that explores new parameterizations of the basic visual interaction loop. The needs dynamics is fed by an unspecific interaction reward and by the specific reward of acquiring new views for the object memory. This is an example of coordinated online learning processes that operate on different time-scales.

7 Conclusion

We have presented an architecture for online learning of arbitrary objects that uses aspects of biologically motivated visual processing in an efficient and robust way. To our knowledge it is the first system that focuses on real online learning of several objects of arbitrary color and shape and their later robust recognition in an unconstrained scenario. The representation is based on a neural model of the ventral pathway and combines a large storage capacity with robustness in difficult real-world environments. Due to the integration of speech dialogue with a context-dependent memory architecture we achieve a high level of interactivity that makes the training procedure simple and intuitive. We consider this as an important step towards cognitive vision systems for robotics and man-machine interfaces that gain considerable flexibility by learning.

Acknowledgments: We thank J. Eggert, A. Ceravola, and M. Stein for providing the processing system infrastructure. We thank F. Joublin and H. Janssen for their contributions to the setup of the speech recognition and synthesis system.

References

[1] J. J. Steil and H. Wersing. Recent trends in online learning for cognitive robotics. In *Proc. ESANN*, pages 77–87. Springer, 2006.

[2] S. Li, A. Haasch, B. Wrede, J. Fritsch, and G. Sagerer. Human-style interaction with a robot for cooperative learning of scene objects. In G. Lazzari, F. Pianesi, J.L. Crowley, K. Mase, and S.L. Oviatt, editors, *Proc. 7th Int. Conf. on Multimodal Interfaces, ICMI, Trento, Italy*, pages 151–158. ACM, 2005.

[3] L. Steels and F. Kaplan. AIBO's first words. the social learning of language and meaning. *Evolution of Communication*, 4(1):3–32, 2001.

[4] S. Kirstein, H. Wersing, and E. Körner. Rapid online learning of objects in a biologically motivated recognition architecture. In *27th Pattern Recognition Symposium DAGM*, pages 301–308. Springer, 2005.

[5] H. Wersing and E. Körner. Learning optimized features for hierarchical models of invariant recognition. *Neural Computation*, 15(7):1559–1588, 2003.

[6] K. Fukushima. Neocognitron: Hierarchical Neural Network Capable of Visual Pattern Recognition. *Neural Networks*, 1:119–130, 1988.

[7] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2006.

[8] Tony Jebara and Alex Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *Int. Conf. Computer Vision Systems*, 1999.

[9] Luiz-Marcos Garcia, Antonio A. F. Oliveira, Roderic A. Grupen, David S. Wheeler, and Andrew H. Fagg. Tracing patterns and attention: Humanoid robot cognition. *IEEE Intell. Sys.*, 15(4):70–77, 2000.

[10] Deb Roy and Alex Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146, 2002.

[11] A. Arsenio. Developmental learning on a humanoid robot. In *Proc. Int. Joint Conf. Neur. Netw. 2004, Budapest*, pages 3167–3172, 2004.

[12] H. Bekel, I. Bax, G. Heidemann, and H. Ritter. Adaptive computer vision: Online learning for object recognition. In *German Pattern Recognition Symposium*, pages 447–454, 2004.

[13] Sebastian Wrede, Marc Hanheide, Sven Wachsmuth, and Gerhard Sagerer. Integration and coordination in a cognitive vision system. In *Int. Conf. on Computer Vision Systems*, 2006.

[14] P. M. Roth, H. Grabner, D. Skocaj, H. Bischof, and A. Leonardis. Conservative visual learning for object detection with minimal hand labeling effort. In *German Pattern Recognition Symposium, Vienna*, pages 293–300, 2005.

[15] P. M. Roth, M. Donoser, and H. Bischof. On-line learning of unknown hand held objects via tracking. In *Int. Conf. on Computer Vision Systems, New York*, 2006.

[16] C. Goerick, H. Wersing, I. Mikhailova, and M. Dunn. Peripersonal space and object recognition for humanoids. In *Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanoids 2005), Tsukuba, Japan*, 2005.

[17] M. Götting, J.J. Steil, H. Wersing, E. Körner, and H. Ritter. Adaptive scene-dependent filters for segmentation and online learning of visual objects. *Neurocomputing*, 70:1235–1246, 2007.

[18] J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, and G. Sagerer. Improving adaptive skin color segmentation by incorporating results from face detection. In *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, pages 337–343, Berlin, Germany, 2002.

[19] Nuance Communications. Nuance vocon 3200 embedded development system, version 2.2, developer's manual. Technical report, Menlo Park, California, 2004.

[20] T.J. Palmeri and I. Gauthier. Visual object understanding. *Nature Reviews Neuroscience*, 5:291–303, 2004.

[21] T. Poggio and E. Bizzi. Generalization in vision and motor control. *Nature*, 431:768–774, 2004.

[22] C. Goerick, I. Mikhailova, H. Wersing, and S. Kirstein. Biologically motivated visual behaviours for humanoids: Learning to interact and learning in interaction. In *Proc. IEEE/RSJ Int. Conf. on Humanoid Robots, Tsukuba, Japan*, 2006.