A biologically motivated system for unconstrained online learning of visual objects

Heiko Wersing[1], Stephan Kirstein[1], Michael Götting[2], Holger Brandl[1], 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] Bielefeld University - Neuroinformatics Group, Faculty of Technology
PO Box 100131, D-33501 Bielefeld, Germany

Abstract. We present a biologically motivated system for object recognition that is capable of online learning of several objects based on interaction with a human teacher. The training is unconstrained in the sense that arbitrary objects can be freely presented in front of a stereo camera system and labeled by speech input. The architecture unites biological principles such as appearance-based representation in topographical feature detection hierarchies and context-driven transfer between different levels of object memory. The learning is fully online and thus avoids an artificial separation of the interaction into training and test phases.

1 Introduction

The capacity for learning and robust recognition of numerous objects makes the human visual system superior to all currently existing technical object recognition approaches. One aspect of this is the capability of quickly analyzing and remembering completely unknown new objects. In this contribution we refer to this ability as *online learning*, which is of high relevance for cognitive robotics and computer vision. A typical application domain we are heading for is to increase the knowledge of an assistive robot in a changing and unpredictable environment [1]. 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 instanteneous and active learning process that reduces the amount of necessary training data and allows an iterative error correction based on user feedback.

To realize such learning, we present a system that combines 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 of a context-dependent working or sensory memory and a persistent object memory that can also be trained online.

The learning is unconstrained in the sense that we 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 [2]. This results in a particularly robust appearance-based representation of objects using a consistent library of typical local shape elements.

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, show results on the performance and learning behaviour in Section 5 and give a short final discussion in Section 6.

2 Related Work

Compared to the large body of work on offline training of model-free object recognition architectures, only few work has been done 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 [3], 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. [4], 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 [5] 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 [6]. The learning proceeds online by using a short-term memory for identifying reoccur-

ring pairs of acoustic and visual sensory data, that are then passed to a long-term representation of extracted audiovisual objects.

Arsenio [7] 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 using a geometric hashing representation.

An interesting approach to supervised online learning for object recognition was proposed by Bekel et al. [8]. Their VPL classifier consists of three major stages. The two feature extraction stages are based on vector quantization and a local PCA measurement. The final stage is a supervised classifier using a local linear map architecture. The image acquisition of new object views is triggered by pointing gestures on a table, and is followed by a short training phase, which take some minutes. The main drawback is the lack of an incremental learning mechanism to avoid the complete retraining of the architecture.

Kirstein et al. [9] 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.

3 System Overview

The visual input is a left and right image pair, obtained from a stereo camera head mounted on a pan-tilt unit. 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. We use the concept of peripersonal space [10] 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. If nothing is present within this space, the cues of motion and color/intensity determine the gaze selection of the system (see [10] for more details).

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. By estimating 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 normalized size of 144x144 pixels.

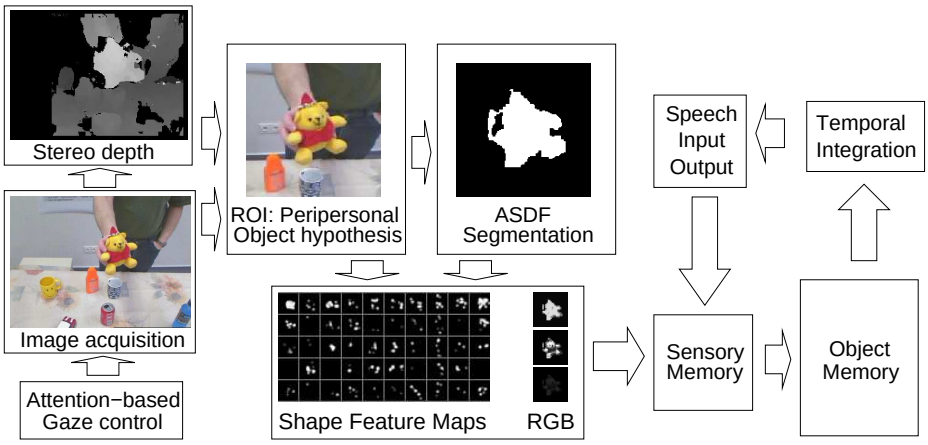

Fig. 1. Overview over the visual online learning architecture. See text for explanation.

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) [11]. 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 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 and 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 [11] for more details). The method obtains an intrinsic stability by continuously iterating the vector quantization based on state on the previous frame. We additionally use skin color detection 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 visual pathway of Wersing & Körner [2] to obtain a complex feature map representation that is based on 50 shape and 3 color feature maps. The color channels are just downsampled images in the three RGB channels. The output is a high-dimensional view-based representation of the input object, that is then passed to the further 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 confirm 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 makes a separation into training and testing phases unnecessary. 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 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 [10].

4.1 Hierarchical Feature Processing

The output of the ASDF figure-ground segmentation stage is a mask signal that is combined with the candidate ROI (of size 144x144 pixels) and fed into the hierarchical model of the ventral visual pathway developed by Wersing & Körner [2]. To obtain invariance against rotations in the image plane, which is normally quite a challenge 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 Fig. 1). In the system that we consider here, we use 50 shape

features, that are sensitive to particular local structural elements in the image, and the three RGB channels. The 50 shape feature maps are represented at a resolution of 18x18, due to the spatial convergence in the hierarchy. As was shown before, the output of the feature representation of the complex 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 [2].

The efficiency of the representation is achieved by sparse coding that ensures 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 at the same resolution of 18x18 as the shape features. Although the complete dimensionality of a single view representation is thus (50+3)x18x18=17172, the effective dimensionality is much smaller, due to the sparsity of the representation vector and the restriction of activity around 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 [3–8]. The key element is a transformation of the input into a sparse robust feature map representation that captures locally invariant relevant 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, where new representatives are added, when they are sufficiently dissimilar to all current entries in the sensory memory. The similarity is measured based on Euclidean distance in the feature map vector space. Due to the sparsity of the feature map vectors this similarity computation can be very efficiently implemented [9].

When a labeling signal arrives, because the human teacher has labeled 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. If there are already some views available in the object memory, the comparison is performed against the already existing representation. 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

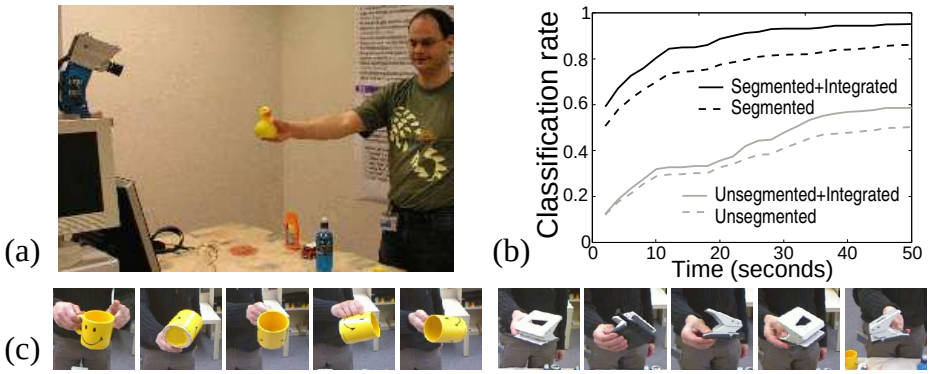

Fig. 2. Presentation scenario for our online learning architecture (a), and average recognition performance versus training time (b) for training the 10th object after 9 were already trained, with and without segmentation and temporal integration.(c) demonstrates the typical rotation variation that is applied during all experiments.

condensed and consolidated using additional learning mechanisms that operate on a slower time scale [9].

Every arriving view is being classified based on the information in the object memory using a nearest-neighbour classifier for the labeled representatives. 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-neighbour 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 achieves 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 Fig. 2a, with typical ROI views of objects that are being processed. During all experiments the objects were freely rotated by hand to obtain a strong appearance variation.

In Fig.2b we show plots of the recognition performance versus training time during online learning. For this evaluation we train nine objects from a training set of 10 objects (upper row in Fig. 3) that was generated by storing 300 views per object from a typical training session. Then the tenth object is trained in steps of 10 images (1.67 sec in Fig. 2c) and a testing step is performed. The test is done by classifying a completely disjoint test set of 300 views per object that was collected using a different training person. Test performance is measured over all 300 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 shown in Fig. 2b show the resulting classification rate, averaged over an ensemble of experiments, where each of the 10 objects was one time the final object.

We compare in Fig. 2b the conditions of either using ASDF segmentation or omitting it (and thus also rotation normalization), and with or 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 significantly higher recognition rate. Using the temporal integration can additionally reduce the error from 15% after 50 seconds of training to 4% error. If we remove the color features and use only the shape representation in combination with ASDF and temporal integration we obtain a residual error of 10%, underlining the independent quality of the shape representation.

We visualize the actual time course of the different memory types during a training session of 18 objects in Figure 3. 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 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 blueishly 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

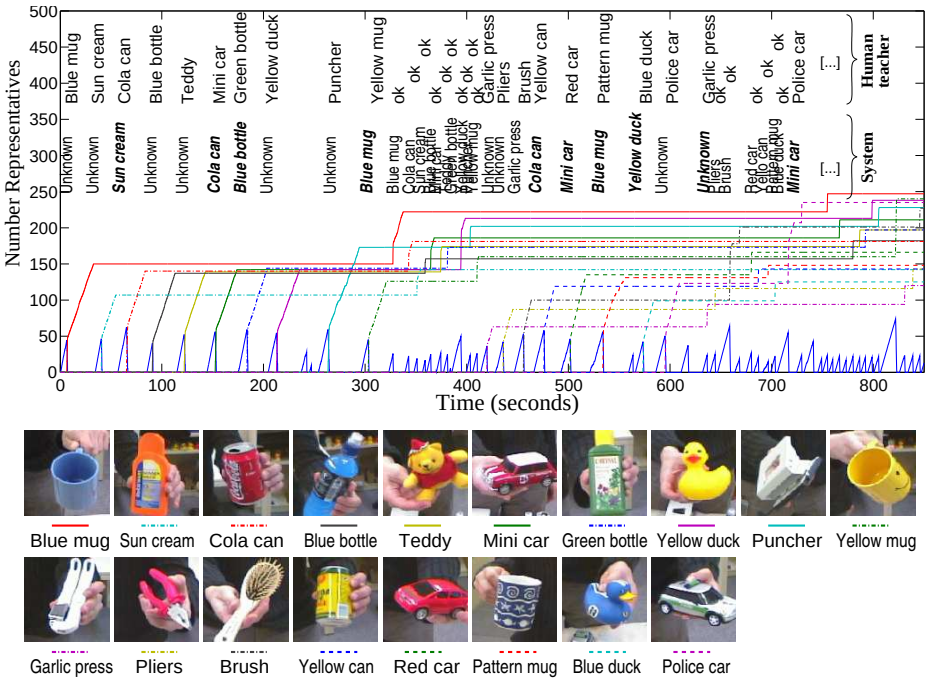

Fig. 3. Temporal learning dynamics during a training session for 18 objects. The plot shows the number of representatives for the sensory memory ("sawtooth" at bottom of plot) and representatives for each object in the object memory over time. The corresponding training dialogue is stated synchronously 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.

be seen from the time course in Fig. 3, 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.

6 Discussion

We have presented an architecture for online learning of arbitrary objects that uses aspects of biologically motivated visual processing in a very 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 scenarios. 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. Steil, J.J., Wersing, H.: Recent trends in online learning for cognitive robotics. In: Proc. ESANN, Springer (2006)
2. Wersing, H., Körner, E.: Learning optimized features for hierarchical models of invariant recognition. Neural Computation **15**(7) (2003) 1559–1588
3. Jebara, T., Pentland, A.: Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In: Int. Conf. Computer Vision Systems. (1999)
4. Garcia, L.M., Oliveira, A.A.F., Grupen, R.A., Wheeler, D.S., Fagg, A.H.: Tracing patterns and attention: Humanoid robot cognition. IEEE Intell. Sys. **15**(4) (2000) 70–77
5. Steels, L., Kaplan, F.: AIBO's first words. the social learning of language and meaning. Evolution of Communication **4**(1) (2001) 3–32
6. Roy, D., Pentland, A.: Learning words from sights and sounds: a computational model. Cognitive Science **26**(1) (2002) 113–146
7. Arsenio, A.: Developmental learning on a humanoid robot. In: Proc. Int. Joint Conf. Neur. Netw. 2004, Budapest. (2004) 3167–3172
8. Bekel, H., Bax, I., Heidemann, G., Ritter, H.: Adaptive computer vision: Online learning for object recognition. In: Proc. DAGM, Tuebingen. (2004) 447–454
9. Kirstein, S., Wersing, H., Körner, E.: Rapid online learning of objects in a biologically motivated recognition architecture. In: 27th Pattern Recognition Symposium DAGM, Springer (2005) 301–308
10. Goerick, C., Wersing, H., Mikhailova, I., Dunn, M.: Peripersonal space and object recognition for humanoids. In: Proc. Humanoids Conf., Tsukuba. (2005)
11. Götting, M., Steil, J., Wersing, H., Körner, E., Ritter, H.: Adaptive scene-dependent filters in online learning environments. In: Proceedings Eur. Symp. Neur. Netw. ESANN, Bruges. (2006) accepted.