Proc. International Conference on Computer Vision Systems ICVS, Bielefeld, Germany 2007

Online Learning of Objects and Faces in an Integrated Biologically Motivated Architecture

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

Abstract. We present a biologically motivated integrated vision system that is capable of online learning of several objects and faces in a unified representation. 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. We combine 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 driven by interactively sharing attention between user and system. It is fully online and avoids an artificial separation of the interaction into training and test phases.

1 Introduction

There is a growing amount of research in assistive and cognitive robotics that stresses the importance of online learning as a key ability of intelligent systems acting in a changing and unpredictable environment [1–3]. We define online learning as the ability of a learning system to work in real-time and provide immediate feedback about the current learning state. This induces 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 this contribution we present a biologically motivated architecture for the online learning of objects and people in direct interaction with a human teacher. Our system combines a flexible neural object recognition architecture with an attention system for gaze control, and a speech understanding and synthesis system for intuitive interaction. 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 of the system is guided by a focus of attention that is communicated to the user by the gaze direction of a stereo camera head, in order to establish attention sharing. 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 using an initial stereo-based coarse object hypothesis. Alternatively, the attention can also be focused on regions of skin-colour, that are detected as being sufficiently face-like. The general 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 [4]. This results in a particularly robust appearance-based representation of objects using a consistent library of typical local shape elements, that can also be applied to face recognition.

The plan of the manuscript is as follows: We review related work in Section 2 and give an overview on our system in Section 3. In Section 4 we explain the components of the visual memory in more detail, show results on the performance and learning behaviour in Section 5 and give a discussion in Section 6.

2 Related Work

The ability of online learning for object recognition has recently gained increasing research interest, although it is still sparsely studied compared to the established field of offline training of model-free recognition architectures. 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.

Bekel et al. [5] proposed an approach to supervised online learning for object recognition, consisting of three stages of vector quantization, local PCA, and 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 [6].

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 [3]. 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 [7]. 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[2]. In both settings a static background was assumed and used for object segmentation.

Kirstein et al. [8] 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. The approach was recently extended to cope with an unconstrained scenario, where an adaptive segmentation method enables training within arbitrary environments [9].

3 System Architecture

We use a stereo camera head mounted on a pan-tilt unit that delivers a left and right image pair for visual input (see Fig.1). The gaze is controlled by an attention system using bottom-up cues like edge/color/intensity contrast, motion, and depth, presented in more detail in [10]. Additionally we provide top-down information on face targets to be followed with a peaked map at the detected face position. Each cue is represented as a retinotopic activation or saliency map. We use a simple addition of the different cues, where we induce clear priorities by weighting the cues in the following sequence: contrast $<$ motion $<$ depth $<$ face. This simple model enables a quite complex interaction with the system to guide the attention appropriately.

The default state of the gaze selection system is an exploratory gazing around that will focus on strong color and intensity contrasts. Moving objects will attract more attention. An even stronger cue is generated by bringing an object into the peripersonal space, that is the near-range space in front of the camera that corresponds to the manipulation space of a humanoid robot [10]. However, also the weaker cues of contrast give a contribution and stabilize the attention. The strongest cue is the presence of a detected face, generating a strong task-specific attention peak at the detected position.

To trigger the online learning and recognition, two parallelly computed object hypotheses are used. Firstly, objects will be learned and recognized, if they are presented within the peripersonal space. The object is attended, as long as it resides within this interaction space. Secondly, using skin color segmentation [11], candidate region segments are classified according to their face similarity. An accepted face region is then selected and processed using the same online learning and recognition pathway as for objects. The attention is retracted from the face, if no valid face-like segment was detected near the image center for two input frames. We describe the two alternative attention selection methods in the following in more detail.

Objects. Based on the current stereo view pair for the selected gaze, 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

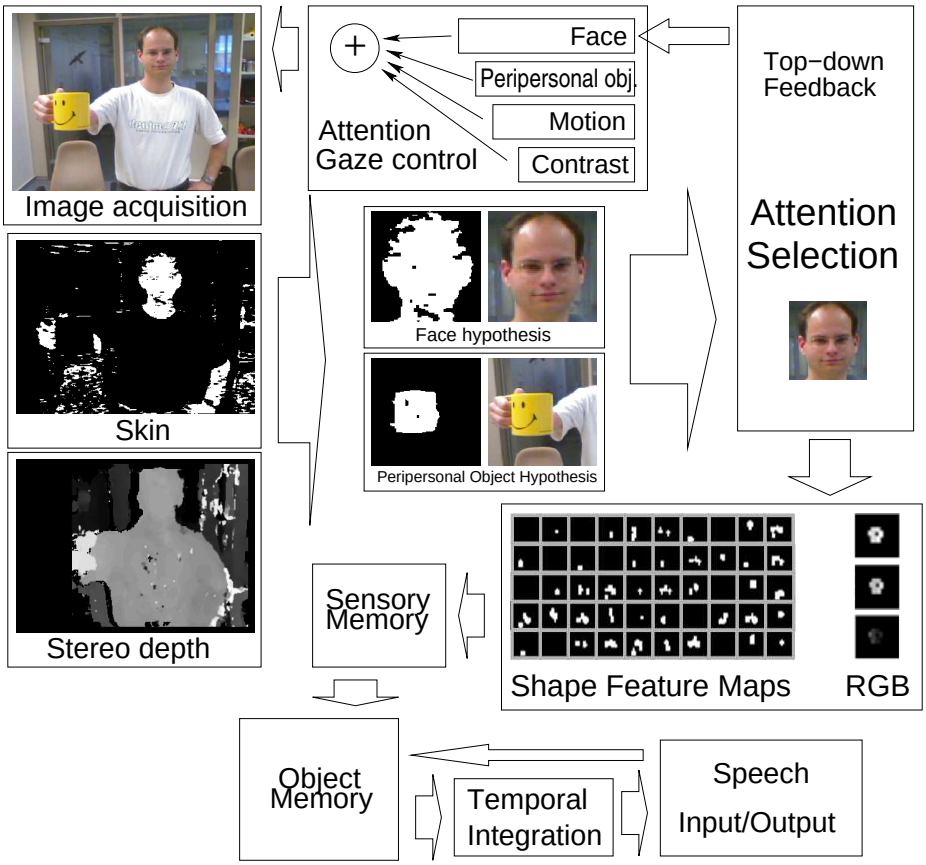

Fig. 1. Overview over the visual online learning architecture. See text for explanation.

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 normalized to 144x144 pixels. This 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) [12]. 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 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 covered by the relevance map (see [12] for more details). The method obtains an intrinsic stability by continuously iterating the vector quantization based on the state of the previous frame. We additionally use a skin color detection model [11] 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.

Faces. First a pixelwise classification of the input is done using the skin-color model and segments are formed using connected components analysis. Of these segments we select the segment, that is closest to the center of the current gaze, i.e. the fovea. Since we assume that the head is at the top, we center the ROI at the top area of the skin segment. To determine the appropriate ROI size, we estimate the distance of the segment using the distance values of the skin pixels according to the depth map. We thus obtain a size normalization, which keeps the head size variation reasonably among $\pm 20\%$. This ROI is normalized to a size of 144x144 and then passed to a classifier that has been trained with an ensemble of face images obtained using the same attention model versus skin-colored non-face segments. This step is necessary to reject skin color segments from hands or similar-colored parts of the environment. The classification is based on the same feature map representation that is used for the online learning of objects and individual faces. The decision is based on a linear decision unit, trained by supervised learning [4]. If the confidence exceeds a threshold, then the region is passed to the online learning model, together with the corresponding binary skin map for segmentation. If both a peripersonal object and a face hypothesis are detected in the current view, then only the face region is selected.

The attended ROI and segmentation mask are fed into the hierarchical feature detection model that delivers a high-dimensional view-based representation of the input object/face 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 attended 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 attended. 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 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 and faces we also use wildcard labels such as "object one", "object two", and "person one" etc.

4 Object Memory Representation

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 a hierarchical model of the ventral visual pathway [4]. 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 (e.g. bottles). If a face is delivered at the input, then the corresponding mask is given by the skin map.

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 [4].

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 add color information, the 3 RGB channels are used as a downsampled ROI at the same resolution of 18x18 as the shape features. Although the dimensionality of a single view representation is thus (50+3)x18x18=17172, the effective dimensionality is much smaller, due to the sparsity of the feature 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 other online learning models [5]. 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, which can be very efficiently implemented due to sparsity [8].

After a 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 condensed and consolidated using additional learning mechanisms that operate on a slower time scale [8].

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 sufficiently high 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 runs on one dual PC for gaze control and image capture, one desktop PC running the speech recognition [13] and synthesis system, and one dual PC performing all visual processing and online learning after the gaze selection. The recognition system runs at a frame rate of roughly 5-6Hz, enabling interaction and online learning with direct feedback on the learning result.

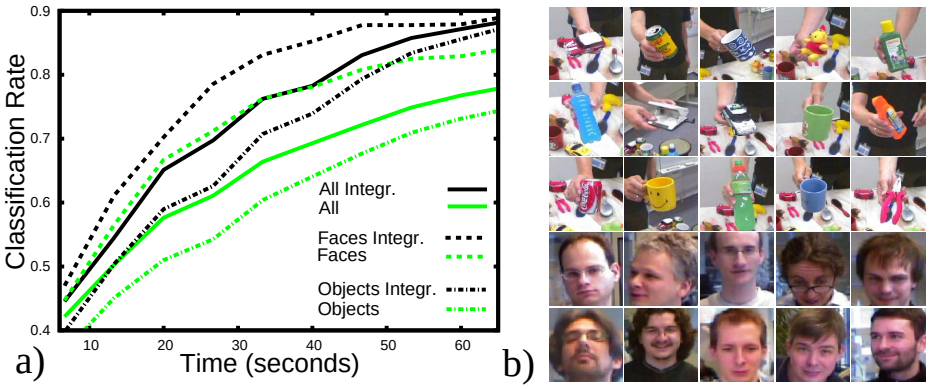

Fig. 2. Performance versus training time. Averaged recognition performance versus training time (a) for training the 25th object/face after 24 were already trained, with and without temporal integration. Additionally a) also shows the results, when only the 15 objects or only 10 faces are used. (b) Shows example train/test images with their typical appearance variations, objects were fully rotated in depth and face direction was allowed to vary. Note the similarities in shape and colors, and the cluttered background.

The performance of the approach is evaluated in Fig. 2a, for the combined online training of 15 objects and 10 faces, shown in Fig. 2b. For this evaluation we train 24 objects/faces from a training set of 25 classes that was generated by storing 100 views per object/face from a typical online training session. Then the 25th object is trained in steps of 10 images (1.67 sec in Fig. 2a) 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 online training session. 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 100 training images are used. The plots shown in Fig. 2a show the resulting classification rate, averaged over an ensemble of experiments, where each of the 25 classes was one time the final class. Additionally 2a) also shows the results, when only the 15 objects or only 10 faces are used. Since there occurs rarely confusion between objects and faces, the complete performance is close to a weighted average between the single settings. The plots show that temporal integration improves the results considerably. Within training time of one minute, classification rates of almost 90% can be achieved. By extending the training time these results could be improved further, the saturation of the learning curve shows, however, that the gain versus necessary training views decreases.

We visualize the actual time course of the different memory types during a training session of 5 objects and 5 people 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). After a training phase of 6 minutes, the system is capable of robustly discrim-

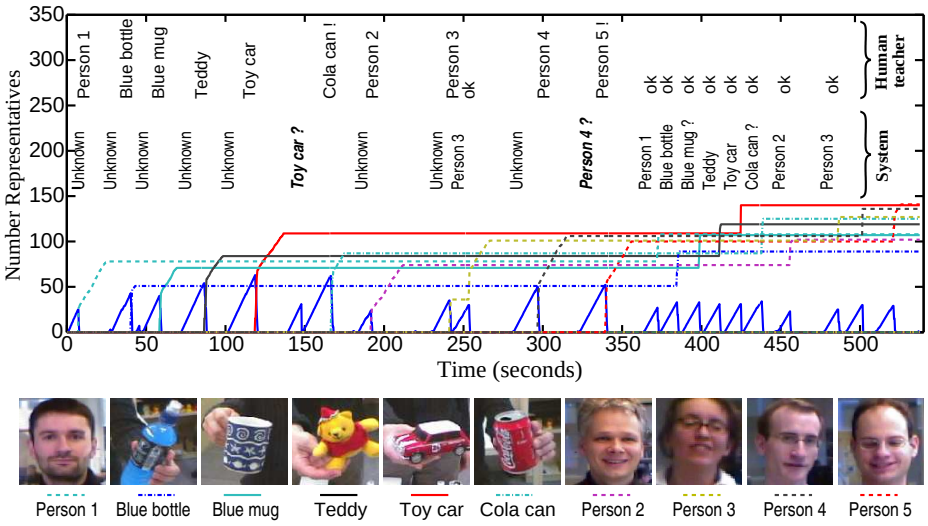

Fig. 3. Temporal learning dynamics during a training session for 5 objects and 5 people. The plot shows the number of representatives for the sensory memory ("sawtooth" at bottom of plot) and representatives for each class in the object memory over time. The corresponding training dialogue is stated 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. A question mark signals a classification with low confidence. First the trainer labels himself as person 1, and then trains 5 objects from 20s to 170s. The first four objects are first rejected as unknown, and then learned after labeling. The cola can is first confused with the toy car, due to similar shape and color. This error is immediately corrected. Four other people are then trained, at 240s the attention focus is shortly lost, and the classification of the system is confirmed after recapturing the face. From 360s on the recognition is correct, and the system continuously learns more views, after confirmation of correct classifications.

inating the 10 classes. 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 Fig. 3, where during the evaluation of between 360s and 560s the object memory is still expanding, due to the confirmation signals of the human teacher on the system classifications.

6 Discussion

We have presented an integrated architecture for online learning of arbitrary objects that uses aspects of biologically motivated visual processing in an efficient and robust way. The object representation is sufficiently general to be extended to the task of face recognition and scales up to recognition problems of realistic difficulty, like full object rotation in depth with cluttered environment. There

exist a large number of carefully optimized methods for face recognition, which would certainly deliver a larger stand-alone performance using our data for offline training. Nevertheless, we consider our realization of real-time online learning in a single representation architecture for objects and faces as unique. The attention system performs gaze control of the system, but is also crucial for ensuring shared attention between learning system and human trainer. 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 view this as a relevant 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) 77–87
2. Roth, P.M., Donoser, M., Bischof, H.: On-line learning of unknown hand held objects via tracking. In: Int. Conf. on Computer Vision Systems, New York. (2006)
3. Li, S., Haasch, A., Wrede, B., Fritsch, J., Sagerer, G.: Human-style interaction with a robot for cooperative learning of scene objects. In: Proc. 7th Int. Conf. Multimodal Interfaces, ICMI 2005, Trento, Italy, ACM (2005) 151–158
4. Wersing, H., Körner, E.: Learning optimized features for hierarchical models of invariant recognition. Neural Computation **15**(7) (2003) 1559–1588
5. Bekel, H., Bax, I., Heidemann, G., Ritter, H.: Adaptive computer vision: Online learning for object recognition. In: Proc. DAGM, Tuebingen. (2004) 447–454
6. Wrede, S., Hanheide, M., Wachsmuth, S., Sagerer, G.: Integration and coordination in a cognitive vision system. In: Int. Conf. on Computer Vision Systems. (2006)
7. Roth, P.M., Grabner, H., Skocaj, D., Bischof, H., Leonardis, A.: Conservative visual learning for object detection with minimal hand labeling effort. In: German Pattern Recognition Symposium, Vienna. (2005) 293–300
8. 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
9. Wersing, H., Kirstein, S., Götting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J., Ritter, H., Körner, E.: A biologically motivated system for unconstrained online learning of visual objects. In: Proc. Int. Conf. Art. Neur. Netw. ICANN. (2006) 508–517
10. Goerick, C., Wersing, H., Mikhailova, I., Dunn, M.: Peripersonal space and object recognition for humanoids. In: Proc. Humanoids Conf., Tsukuba. (2005)
11. Fritsch, J., Lang, S., Kleinehagenbrock, M., Fink, G.A., Sagerer, G.: Improving adaptive skin color segmentation by incorporating results from face detection. In: Proc. IEEE Workshop ROMAN, Berlin, Germany (2002) 337–343
12. 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)
13. Nuance Communications: Nuance vocon 3200 embedded development system, version 2.2, developer's manual. Technical report, Menlo Park, California (2004)