Proc. Intern. Conf. on Computer Vision Systems 2008. Santorini, Greece.

Online Learning for Bootstrapping of Object Recognition and Localization in a Biologically Motivated Architecture

Heiko Wersing[1], Stephan Kirstein[3], Bernd Schneiders[2],
Ute Bauer-Wersing[4], Edgar Körner[1]

[1] Honda Research Institute Europe GmbH,
Carl-Legien-Str. 30, 63073 Offenbach/Main, Germany
[2] University of Applied Sciences Trier, PO Box 1380, 55761 Birkenfeld, Germany
[3] Technical University of Ilmenau, 98693 Ilmenau, Germany
[4] Univ. of Applied Sciences Frankfurt, Nibelungenplatz 1, 60318 Frankfurt, Germany

Abstract. We present a modular architecture for recognition and localization of objects in a scene that is motivated from coupling the ventral ("what") and dorsal ("where") pathways of human visual processing. Our main target is to demonstrate how online learning can be used to bootstrap the representation from nonspecific cues like stereo depth towards object-specific representations for recognition and detection. We show the realization of the system learning objects in a complex real-world environment and investigate its performance.

1 Introduction

The human visual system enables us to easily perform tasks like navigation, collision avoidance or searching. Understanding the internal processes that lead to these perceptual powers is a major goal of cognitive neuroscience and has high relevance for computer vision. One interesting question is the degree of modularity observable in the human visual system. It has been argued that the visual system consists of a number of interacting but still autonomously functioning subsystems processing different cues like shape, color and motion [1]. Another example are the ventral and dorsal pathways of visual perception, also called "what" and "where" streams due to their role in recognition and localization of objects. This viewpoint is challenged by findings that emphasize the rich interactions between these pathways [2]. Within such a network, subsystems can serve as mutual partners for learning to bootstrap their representations, combining sensory input and output of other modules. In this contribution we investigate an online learning model for this bootstrapping process.

Approaches to online learning have recently gained interest due to their importance for intelligent cognitive systems interacting with dynamically changing environments [3, 4]. Nevertheless this topic is still in its infancy, compared to the large effort done on object recognition and detection methods using offline learning on large image databases. Along one line of research, online learning of representations for larger object ensembles was investigated using segmentation methods during training and recognition, and employing either dimension reduction methods [5, 6] or high-dimensional sparse feature map representations motivated from the ventral pathway [7]. Another research direction is the focus on cross-modal interactive learning of visual and auditory stimulus concepts [8], where generally only simple visual object attributes like color and global shape are considered. Localization and detection of objects has also frequently been considered for attentional processes in visual search [9]. Here the main focus was so far on attenuating low-level features like color, intensity and orientation contrasts to facilitate top-down attention towards target objects [10, 11].

In this contribution we propose a modular online learning architecture using interactions of the "what" and "where" pathways, based on concurrent adaptation of representations. The target is a visual system that is capable of bootstrapping its visual object representations from unspecific (e.g. depth) towards more object-specific cues like shape. The situation that we consider is the learning of object detection and localization using shape and color cues only, where bootstrapping is based on stereo depth information. In analogy to related response properties of neurons in the ventral pathway, the model uses topographic population code representations of visual object information for recognition and localization and adapts these online using simple linear learning models. We thus extend prior work on online learning using segmentation towards segmentation-free detection of arbitrary objects within a scene.

In Section 2 we first present the biological background of our model and then introduce in Section 3 the system architecture. In Section 4 we give some results for our approach on benchmark data for offline learning and then discuss the system performance for online training and testing scenarios.

2 Biological Background

Although ventral and dorsal pathways have long been considered as dissociate in their processing of "what" and "where" [1], recent biological evidence has emphasized similarities that could ease interactions between the anatomically segregated modules. In the ventral pathway, neurons in higher areas like the inferotemporal (IT) cortex are increasingly selective to particular objects and parts, with larger spatial invariance. It has been shown, that selectivity to position shows a Gaussian tuning curve within the receptive field for many object specific IT neurons [12] and also size-specific tuning can be observed [13]. This provides a population code based representation that could be used to obtain estimates of object position and size. The superior temporal sulcus (STS) in the superior temporal cortex has been considered as such an area combining information from both ventral and dorsal pathways, and is strongly involved in spatial search and attention to objects [2].

There has been an increasing effort in the recent years to provide models of the hierarchical processing in the ventral pathway leading to so-called view-

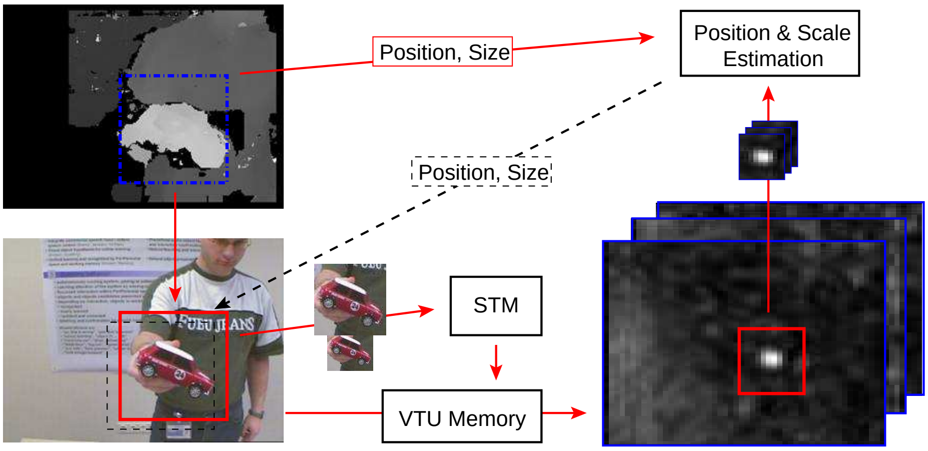


Fig. 1. Architecture overview. A stereo-based position and size estimation is used to bootstrap learning of these properties from a view-tuned unit (VTU) object-specific feature map. Learning of VTUs at multiple scales using attention and an appearance-based short-term memory is done synchronously with position and scale learning.

tuned-units with response properties similar to IT neurons [14–16]. Online learning of object representations using these models was demonstrated by [7] and shown to scale well also for larger object ensembles.

3 System Architecture

The visual input for the system is given by the output of a stereo camera head mounted on a pan-tilt unit and delivering two RGB images. The architecture (see Fig. 1) consists of the following main components:

- A stereo-based attention system determines a near object in the peripersonal space [17], points the camera gaze towards it, and delivers a size and position estimate for the object blob hypothesis as a region of interest (ROI).
- A short-term memory (STM) collects several views of the current object in the focus of attention, using a hierarchical sparse feature map representation. Views are stored at multiple resolutions and online learning of view-tuned units (VTUs) is performed for several object classes.
- Based on the trained VTUs, a map of object selective VTU responses is computed on the scene image and can be used for object localization
- An integration component independently estimates object identity, position and scale, based on maximum selection and the local activation in the VTU map. If an object is in the depth-based attended ROI, the component is trained with the current local VTU response map as input and the ROI parameters as output.

In the following we describe the system components in more detail:

Attention and Gaze Selection. The attention system is based on the gaze control system presented by Goerick et al. [17] for online learning of object representations. Using a stereo-based depth map, connected pixels lying within a defined depth range (also called peripersonal space) are clustered, and the frontal region is taken as an initial object hypothesis. The focus of attention and gaze direction of the system is centered on the center of mass (x, y) of this region. The size s (in pixels) of a square region of interest (ROI) around this point is scaled according to the distance estimate using the depth map. The ROI size s in pixels is computed as $s = 144 \cdot 0.6/d$, i.e. an object at distance $d = 60$ cm gives a ROI of $s = 144$ pixels. The attended ROI is rescaled to a set of input RGB patches $\mathbf{I}^{\{1,2,3\}}$ of fixed sizes 144, 128, and 96, which are passed to the short-term-memory for learning. This induces size normalization in each of the three scales. In addition to setting the focus of attention we use the parameters of the attended ROI $\mathbf{r} = (x, y, s)$ to train the position and scale estimation module that is capable of localizing objects, when no stereo-based attention is available. This is the case, when the objects are not separable within the scene based on stereo alone. Unspecific object size and position within the peripersonal space is an example of an action-related representation that is typically localized in the dorsal pathway of human visual processing.

Short-term Memory for Online Training of VTUs. The short-term memory module collects object views, when an object is in the focus of attention. Using a concept of sensory memory as proposed in [7], views are buffered until a label is given by speech input, after which the views in the sensory buffer are assigned to the object in the current context. Representation in the STM is based on the output of a hierarchical feature detection model of the ventral visual pathway as described in [16]. For each input RGB patch $\mathbf{I}^{\{1,2,3\}}$, the output of the feature hierarchy is computed as $\mathbf{x}^k(\mathbf{I}^k) = (\mathbf{c}_1^k, \ldots, \mathbf{c}_{50}^k, \hat{\mathbf{I}}_R^k, \hat{\mathbf{I}}_B^k, \hat{\mathbf{I}}_G^k)$, where $\mathbf{c}_i^k$ is the output of the topographic combination feature detection map of feature i, and $\hat{\mathbf{I}}_R^k, \hat{\mathbf{I}}_B^k, \hat{\mathbf{I}}_G^k$ are downsampled coarse images of the RGB image channels. Due to the spatial convergence the size of each map is 8 times reduced compared to the input patch size, (i.e. 18x18, 16x16, 12x12). The collection is incremental, if a new view is sufficiently dissimilar to existing view representatives based on Euclidean distance, it is added to the STM. From this STM, training views are concurrently randomly drawn and used to train VTUs as linear discriminating units with a one-against-all classification output vector [16]. Here we perform online gradient descent in the quadratic error between output and target value from supervised learning. To increase the rejection capability we also add a set of clutter views, based on a collection of arbitrary images from the internet, which are trained as a rejection class. We also performed experiments using clutter from our real scene setting, but observed no consistent improvement. Note that unlike previous work by [7] we do not perform segmentation during learning, since this is also not available for object detection in the scene.

Object-specific VTU Maps for Detection and Localization. The VTUs are trained for each object class in three scales using the STM. Formally,

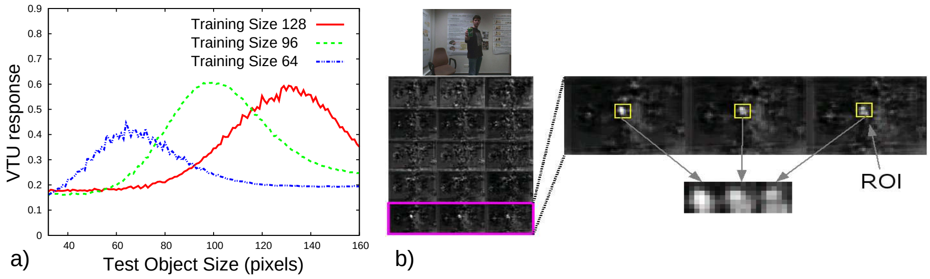


Fig. 2. a)VTU size tuning. Average response of VTUs trained at particular object sizes, using the COIL17 object data. b) Selection of training data from VTU maps for localization. For training, the map activity of the current object is taken at different scales around the maximum within the attended ROI.

a VTU is a linear discriminator, which is trained to respond on a receptive field of shape feature and coarse color input. As a result of the spatial integration in the feature hierarchy, these VTUs exhibit a rather regular Gaussian tuning with respect to spatial displacement and size change of the object within the receptive field (see Fig. 2a). This tuning behaviour is similar to the sensitivities in the IT cortex [12, 13]. Due to the topographic arrangement of the feature map input the VTU was trained for, we can setup a complete map of VTUs for each object and scale, covering the complete input scene (600x450 pixels) at 8 times reduced resolution. From the population code of the map response with displaced receptive fields and multiple scales we can train a model to read out the precise localization and scale information.

Position and Scale Estimation. This component learns the mapping from the local VTU map population response to the local position and size for objects, without being specific to single objects. Functionally, this could be realized in the STS area of the human visual pathway, combining convergent inputs from ventral and dorsal pathways, and important for object-based search and spatial attention processes [2]. If an object is present in the depth-based focus of attention, the ROI parameters $\mathbf{r}$ are computed based on the depth blob and used as a training target. The training input pattern is obtained from combining the local VTU map responses in a neighbourhood of the maximal response within the attended region. Due to the topographic representation, the ROI from the original image can be easily remapped to the VTU maps by division by 8.

Around the maximum we cut out within each scale map of the current object a local patch $\mathbf{p}^{\{1,2,3\}}$ of 7x7 VTU output responses (see Fig. 2b). It turns out that it is recommendable to restrict the training to appropriate patches, i.e. for which a proper localization is principally possible. Therefore we reject all cases, when the global maximum in the VTU maps for the currently attended object is not inside the attended ROI. Consequently, the learning begins only after the VTU learning has reached a certain object selectivity. The patches are written into a vector $\mathbf{v}$ and we train a simple linear estimation model $\mathbf{r}^* = M\mathbf{v} + \mathbf{b}$ with a matrix M and bias $\mathbf{b}$. The error is minimized by online stochastic gradient descent in the summed quadratic differences between $\mathbf{r}$ and $\mathbf{r}^*$. The training is based on a local memory history of the past 500 valid input patches $\mathbf{p}$.

After training, the module delivers new target ROIs based on globally determining the maximum in the global VTU map, and then using its locally learned model for obtaining an object position (relative to the maximum) and size estimate from the local population activity. This ROI can be used to generate a new focus of attention, even if no stereo-driven hypothesis is available. In a preliminary study [18] we also investigated alternative models for the position and scale estimation using only offline learning and artificial data and compared

- just taking the maximum position and scale in the map and compute corresponding position by multiplying by 8
- computing the center of gravity around the local maximum in the map
- training a radial-basis-function (RBF) network with the VTU map input and the target outputs using different numbers of hidden nodes
- using a linear model mapping the local VTU activity pattern to the target values like in this contribution

The first two simple models exhibited roughly a double position and scale error compared to the trained models, due to the lack of robust interpolation from the population code output. The RBF networks delivered best performance, but the linear model was only slightly worse. Since the linear approach is best suited for an online setting we therefore chose this one for our combined architecture.

System Implementation. The whole system runs in a component environment for the large-scale real-time simulation of intelligent systems [19]. The STM learning runs at a frame rate of about 5 Hz, while the computation of the VTU maps on the whole image runs at 1-2 Hz. Computation hardware is a 2.4 GHz Quadcore Intel processor.

4 Results

We first state some results for comparison to other computer vision detection approaches using offline learning on a benchmark problem. We then show the temporal learning curves for the online learning system in interaction with a user and evaluate the performance for different scenarios.

4.1 Benchmark Comparison for Offline Learning Detection Task

The task of object detection in complex cluttered scenes has been intensively studied using several different computer vision approaches. To assess the detection and discrimination capabilities of the VTU map model, we performed a benchmark comparison on the UIUC single scale car detection task [20]. This is a single-class detection problem for side views of cars, based on a training ensemble of 500 car and 500 clutter views. The test ensemble consists of 170 test images with varying difficulty, where one or several cars have to be detected.

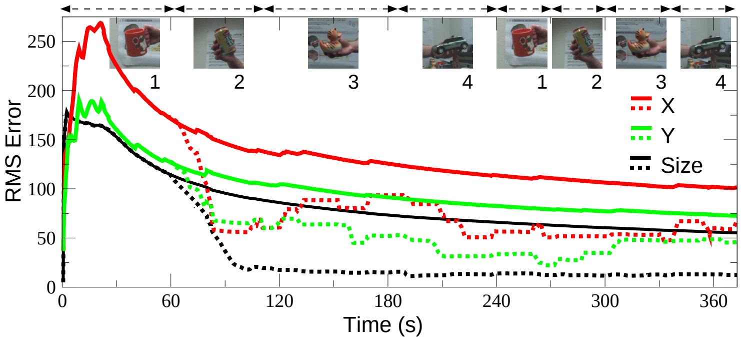


Fig. 3. Temporal dynamics of online learning RMS error for x, y, and size of object ROI. The solid lines give the total average error, while the dashed lines are local averages over the last 70 steps. At the top the shown object is visualized, we first train 4 objects and then test them from 240s on. Learning begins to successfully reduce errors, when the VTU response is getting more tuned. New objects first increase the local errors, and then require some time until error converges again.

With the introduction of the database, Agarwal et al. achieved 76.5% equal error detection rate using a parts-based approach. This was improved by Leibe et al. [21] to 97.5% using a parts-based implicit shape model, and recently Mutch & Lowe [15] achieved 99.9% using a biologically motivated feature hierarchy similar to our approach. We train a single VTU sensitive to the car training views and then take the local maxima in the VTU map output as target detection. Using this, we obtain an equal error rate of 97%. One advantage of our model is that it is based on only 50 shape features obtained from sparse coding [16], compared to large libraries of up to 1000 local features as in [21, 15]. This causes a $50-100$ times faster implementation and makes online learning possible.

4.2 Online Learning Scenario and Evaluation

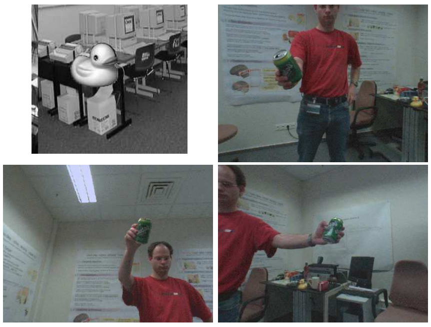
We present the objects to be learned in an office environment, where the gaze control system is responsible for keeping the objects within camera view (compare Fig. 4b). Starting from an empty object memory we can train a number of objects and interactively watch the progress both in learning of object representations and the position and size estimation. Figure 3 shows an example for a learning curve recorded during a typical training session for 4 objects.

For a more comprehensive evaluation of the architecture we selected one artificial scenario with perfect ground truth, and three online scenarios of varying difficulty. For the latter we considered the training of the position estimation after convergence of VTU training, to leave out the transient phase. The test error is computed on data from a disjoint sequence, performed by another subject. The four scenarios are:

1. We select 17 objects from the COIL20 [22] database, where we removed similar objects (2 cars and one pillbox) from the same category, since we are

		X	Y	Size
COIL17	Training	4.92	4.59	6.76
	Test ROI	5.75	5.53	8.44
	Test scene	52.83	56.12	12.58
Single Pose	Training	8.02	6.27	12.04
	Test ROI	11.60	8.68	14.64
	Test Scene	69.37	58.58	16.69
Multi Pose	Training	17.16	14.77	16.69
	Test ROI	20.92	17.23	17.24
	Test scene	158.32	96.65	22.11
Single Pose ++Size Var	Training	15.94	15.13	19.74
	Test ROI	21.08	19.30	22.33
	Test scene	130.71	92.45	24.50

a) Position and size errors (pixels)



b) Artifical and real views

Fig. 4. a) Results of position and size estimation RMS error for 4 scenarios. b) One artificial image and three images showing view-point variation during online interaction.

here not interested in detailed identification of single objects. Objects are segmented and placed at random positions and scale variations from 64-128 pixels visible size onto a set of clutter images of size 320x320, cropped from images collected from the internet. All the images are greyscale only.

2. We train 10 objects in an interactive fashion, where the pose variance is limited to a single pose with a variation of about 10-30 degree rotation around all axes. Object distance varies from 40-80 cm, while position in space is strongly varied (compare Fig. 4b).
3. Like 2. but allowing full rotation with multiple poses
4. Like 2. but distance varying from 40-120 cm

The table in Fig. 4a summarizes the performance of the position and scale estimation component. The training error is the root mean square (RMS) error of the position x, y and ROI size s. The test error is computed by either restricting the maximum search in the VTU maps to the attended target ROI or allowing free search in the whole scene. The first error allows to asses the performance of the local population code estimation from the VTU map, while the second is also heavily influenced by globally wrong responses of VTUs to clutter.

The results show that the COIL17 scenario is easiest, although it contains full rotation in depth along a single axis and no color. Within the test ROI, test position and scale error is only slightly larger than training error. The test position error in the complete scene corresponds to about half the maximal object size. For the online learning scenario, the single pose is easiest, with the increased depth interval raising difficulty and the multi-pose setting as the hardest task. This is consistent across all errors. As is evident from the large discrepancy between errors in ROI and scene settings, false maxima determine here most of the position error. The larger errors in x than in y position are induced by the larger x dimension of the scene images.

We also performed an evaluation of the detection performance regardless of position, independently for each object. For each test frame containing a test

Single Pose	0.4	4.7	2.4	0.9	1.2	3.2	1.5	2.1	4.0	1.2
Multi Pose	15.0	26.9	25.2	47.1	22.5	27.2	21.8	39.9	29.5	7.4
S. Pos+Dpth	6.6	23.1	18.4	12.4	15.8	28.3	17.9	30.9	25.0	19.0

Table 1. Equal error rates (%) for object detection in the test scenarios.

object, the maximum activity is computed within each object set of VTU maps. Based on the vector of confidences, we can compute an ROC curve, where each frame is counted as a true positive for the object contained in the frame and a possible false positive for other objects. The results are listed for each object in Table 1. Across all objects, the detection of a single pose with limited size variation is substantially easier with much less false positive detections. Especially for the multi-pose setting the integration of changing object appearance information over the whole viewing sphere reduces the VTU selectivity considerably.

5 Discussion

We proposed an online learning framework for cross-modal bootstrapping of object-specific representations from unspecific cues like stereo depth. This extends prior work on object online learning from segmentation-based methods to the case of object detection in a scene. An immediate application of this approach could be the search for a recently trained object by a mobile robot, e.g. building on methods as presented in [9]. For the single pose case, detection can be achieved without substantial false positives. For multiple poses, the performance figures show that the model could still serve as an object-specific attention delivering candidate ROIs for objects, that are then inspected by a fovealized more precise recognition method. This allows to combine more selective shape feature channels, than simpler models using only orientation and contrast.

We believe that the concept of modular subsystems that are learning in interaction is of high relevance both for biological vision systems and their computer vision counterparts. Our example illustrated that this can be achieved using the biological principles of sparse and topographic representations and population coding in combination with linear learning methods. We consider the extension of this concept to more complex visual architectures as a promising future research direction.

Acknowledgments: We thank A. Ceravola, B. Bolder, J. Eggert, C. Goerick, M. Dunn, and M. Stein for providing the processing infrastructure.

References

1. Zeki, S.: Localization and globalization in conscious vision. Annual Review Neuroscience **24** (2001) 57–86

2. Karnath, H.O.: New insights into the functions of the superior temporal cortex. Nature Reviews Neuroscience **2** (2001) 568–576
3. Crowley, J.L., Hall, D., Emonet, R.: Autonomic computer vision systems. In: Proc. ICVS, Bielefeld. (2007)
4. Steil, J.J., Wersing, H.: Recent trends in online learning for cognitive robotics. In Verleysen, M., ed.: Proc. European Symp. on Neural Networks. (2006) 77–88
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. Roth, P.M., Donoser, M., Bischof, H.: On-line learning of unknown hand held objects via tracking. In: Proc. Second Int. Cognitive Vision Workshop. (2006)
7. Wersing, H., Kirstein, S., Götting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J., Ritter, H., Körner, E.: Online learning of objects and faces in an integrated biologically motivated architecture. In: Proc. ICVS, Bielefeld. (2007)
8. Skokaj, D., Berginc, G., Ridge, B., Stimec, A., Jogan, M., Vanek, O., Leonardis, A., Hutter, M., Hawes, N.: A system for continuous learning of visual percepts. In: Proc. ICVS, Bielefeld. (2007)
9. Tsotsos, J., Shubina, K.: Attention and visual search: Active robotic vision systems that search. In: Proc. ICVS, Bielefeld. (2007)
10. Hamker, F.H.: The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. Computer Vision and Image Understanding **100**(1-2) (2005) 64–106
11. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: Proc.CVPR. (2006) II: 2049–2056
12. Beeck, H.O.D., Vogels, R.: Spatial sensitivity of macaque inferior temporal neurons. Journal Comparative Neurology **426**(4) (2000) 505–518
13. Ito, M., Tamura, H., Fujita, I., Tanaka, K.: Size and position invariance of neuronal responses in monkey inferotemporal cortex. J Neurophysiol **73**(1) (1995) 218–226
14. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE PAMI **29**(3) (2007) 411–426
15. Mutch, J., Lowe, D.G.: Multiclass object recognition with sparse, localized features. In: CVPR, New York (2006) 11–18
16. Wersing, H., Körner, E.: Learning optimized features for hierarchical models of invariant recognition. Neural Computation **15**(7) (2003) 1559–1588
17. Goerick, C., Wersing, H., Mikhailova, I., Dunn, M.: Peripersonal space and object recognition for humanoids. In: Proc. Humanoids, Tsukuba. (2005)
18. Hegde, A.: Object position and size estimation from output activations of a hierarchical invariant object recognition framework. Master's thesis, Univ. Applied Sciences Frankfurt (2006)
19. Ceravola, A., Joublin, F., Dunn, M., Eggert, J., Goerick, C.: Integrated research and development environment for real-time distributed embodied intelligent systems. In: Proc. IROS, Bejing, IEEE Press (2006)
20. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. IEEE PAMI **26**(11) (2004) 1475–1490
21. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV'04, Prague (2004) 17–32
22. Nayar, S.K., Nene, S.A., Murase, H.: Real-time 100 object recognition system. In: Proc. of ARPA Image Understanding Workshop, Palm Springs (1996)