Proc. IEEE/RSJ Int. Conf. on Humanoid Robotics, Tsukuba, 2005

Peripersonal space and object recognition for humanoids

Christian Goerick, Heiko Wersing, Inna Mikhailova and Mark Dunn
Honda Research Institute Europe GmbH
Carl-Legien-Strasse 30
63073 Offenbach / Main
Germany
Email: christian.goerick@honda-ri.de

Abstract— **This work is concerned with a framework for visual object recognition in real world tasks. Our approach is motivated by biological findings of the representation of space around the body, the so-called peripersonal space. We show that the principles behind those findings can lead to a natural structuring of object recognition tasks in artificial systems. We demonstrate this by the supervised learning and recognition of 20 complex-shaped objects from unsegmented visual input.**

Keywords: **Peripersonal space; visual object recognition; humanoid vision system.**

I. Introduction

For many years the research concerning intelligent systems was focused on isolated cognitive tasks. It was ignored that the interplay between the parts of the system as well as interaction with the environment can considerably ease the solution [1]. Nowadays it is widely accepted that scaling out (integrating a task into a complete system) is as important as scaling up [2]. In the spirit of this new philosophy several attempts to facilitate the hard task of recognition via integration into acting systems were done [3], [4], [5], [6]. One source of facilitation is the ability of an acting system to change its perceptions through its actions [7]. Another source of recognition facilitation is interaction and communication between humans and humanoids. The main concept put forward within this framework is shared attention with the current implementations of pointing and gaze following. The core point of shared attention is the cognitive concept of understanding the intention or goal of the communicating partner [8]. The realization of the full concept is far from being complete since many necessary sub-concepts are far from being understood. Shared attention is a psychological concept focusing on the observable behavior of man and artifacts. We would like to put forward a concept rooting in physiology, i.e. concerned with the internal mechanisms and representations of biological organisms and the caused effects.

Here we propose to use the peripersonal space for structuring the human-robot interaction. We follow the definition of peripersonal space given in [9]: Peripersonal space is defined as the space wherein individuals manipulate objects, whereas extrapersonal space, which extends beyond the peripersonal space, is defined as the portion of space relevant for locomotion and orienting. The benefit of using this biologically inspired concept is that it leads to robust interaction and recognition and allows to build up the higher level abilities on the base of a stable complete closed-loop system.

The full biological concept of peripersonal space includes sensory perception as well as manipulation. We consider it to be very valuable for humanoid robots like ASIMO [10] because of the natural integration of perception and actions. Here we focus on object recognition within the framework of peripersonal space only: how the object recognition problem is phrased within this framework and what the underlying object hypotheses are. Next steps towards integrating actions with object recognition would be using the affordances concept [11]. An affordance is a property of an object, or a feature of the immediate environment, that indicates how that object or feature can be interfaced with. In our case affordances link the visual appearance or shape of an object with the manipulative capabilities of the robot.

This paper is structured as follows:

- section II focuses on the concept of peripersonal space in biology,
- section III shows how this concept can be applied to a technical system,
- section IV describes the details of the object recognition model,
- section V reports on and discusses the performed experiments,
- and finally section VI concludes how object recognition can benefit from the usage of peripersonal space.

II. Peripersonal Space: Biological Foundations

The research in neurobiology provides large evidence that the brain uses representations which are both perception and action oriented. The most famous example are so-called “Mirror Neurons”. These neurons fire in the case of e.g. a monkey performing an action or watching somebody else performing the same action (for review see [12]). Less known in the robotics community are the neurons which change their visual receptive fields according to the position of the limbs. For example the bimodal neurons in the monkey premotor cortex respond to the touch of the skin as well as to the visual input in the proximity of the skin. The visual receptive fields of these neurons are not in retinotopic coordinates and not in

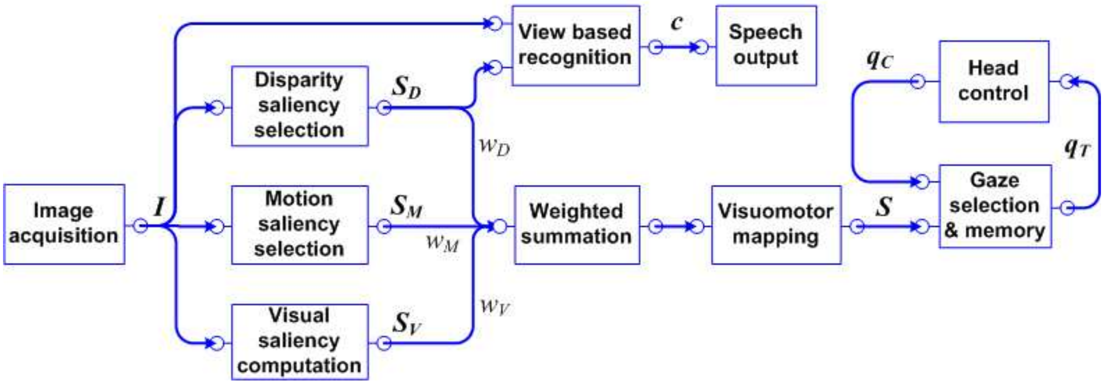


Fig. 1. Systems schematics. See section III for a description.

some "world" coordinates, but in body coordinates: they move when the correspondent part of the body moves, and not when the eyes move. The brain features a mechanism for the transformation from one coordinate system to another depending on the task to perform. Some of the space representations are very flexible. For example during the tool usage the visual receptive field previously dedicated to the space around the hand can extend to the space occupied by the hand and the tool [13]. The adaptation of the visual receptive field can only be observed if the tool is actively used.

In the brain depth information is also organized in behaviorally relevant representations. Experiments show that different functional loops are activated while executing the same task in near and far space [14]. The near or peripersonal space is defined as the space wherein individuals manipulate objects, whereas far or extrapersonal space is defined as the portion of space relevant for locomotion and orienting. Of particular interest for us is the fact that attentional mechanisms make use of these different spatial representations [9].

A further suggestion from the same literature on the possible space representation can be made by the observation of bimodal (triggered by both touch and visual stimuli) neurons' firing. Some of these neurons are firing stronger the closer visual stimuli come to the body. It shows again that in the brain there exist other depth representations than a homogeneous and full depth information which is standard in robotic applications. In the following section we propose how to integrate depth into the attentional mechanism of a technical system by using the concept of peripersonal space and with a body and task-related representation.

III. Peripersonal Space: Technical Interpretation

The technical system used for experiments is depicted in Figure 1. It consists of image acquisition, visual saliency computation, disparity saliency selection, motion saliency selection, saliency weighting and summation, visuomotor mapping, gaze selection including memory, head control as well as object recognition and speech output. As pointed out in the previous section, we are especially interested in the relation between peripersonal space and attention shifting and the consequences for object recognition.

The functional elements of the system can be described as follows: The visual saliency computation is implemented as suggested in [15]. It produces a retinotopic map $\mathbf{S}_V$ where positions corresponding to visually interesting locations are activated, i.e. have a value between zero and one according to their degree of interestingness. The motion saliency selection produces a map $\mathbf{S}_M$ with an activation corresponding to the largest connected area of motion within a defined time-span.

The disparity saliency selection performs a disparity computation using SRI's Small Vision System [16] and selection of the closest connected region within a specific distance range and angle of view. The position of this region in image coordinates is represented as an activation blob within the map $\mathbf{S}_D$. If there is no stimulus within the specified range and angle of view, the activation of the map is all zero. This simple mechanism represents a first approximation to the concept of the peripersonal space put forward in the previous section. It establishes a body-centered zone in front of the system that directly influences the behavior of the overall system as we will show at the end of this paragraph. The selection and propagation of the closest region only corresponds to a hard weighting of the presented stimuli with respect to the "closeness" to the system

The maps of the visual saliency, the motion saliency selection and the disparity saliency selection are weighted, summed up and transformed into motor space, yielding the integrated saliency map

$$\mathbf{S} = visuomotor(w_V \mathbf{S}_V + w_D \mathbf{S}_D + w_M \mathbf{S}_M) \ , \qquad (1)$$

where w_V, w_D and w_M are the respective weight factors. The motor space is spanned by the pan and tilt angles of the head. The visuomotor mapping can be constructed analytically or learned online [17]. The saliency map S is the first input to the gaze selection. The gaze selection is a dynamic neural field (DNF), an integro-differential equation modeling the dynamics of activations on the cortex. The dynamics can roughly be described as maximum selection with hysteresis and local interaction. For the theoretic fundamentals of the gaze selection see [18], for a comprehensive biomimetic head control using DNF see [19]. The DNF is parameterized to yield a unique activation, which is interpreted as the target gaze direction $\mathbf{q}_T$. This target gaze direction is propagated to the head control unit, which delivers the current gaze direction $\mathbf{q}_C$, the second input to the gaze selection.

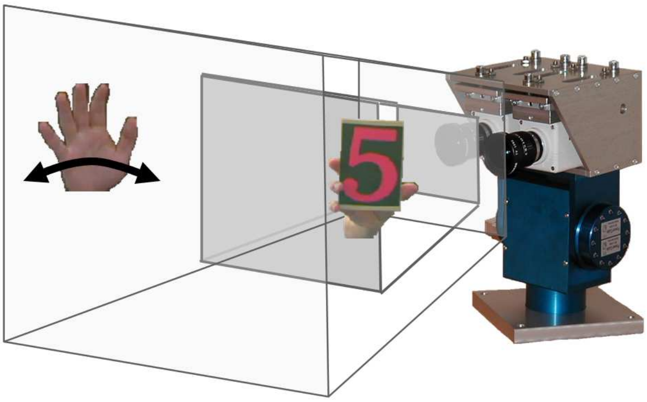


Fig. 2. Schematic visualization of the approximation of the peripersonal space. The inner volume represents the peripersonal space, the outer volume the complete field of view with the sensitivity to visual and motion stimuli.

For the work presented here we parameterize the system with $w_V = 1.0, w_M = 3.0$ and $w_D = 4.0$. This corresponds to prioritizing the disparity information over the motion and visual saliency and the motion information over the visual saliency. With those weights the system shows the following behavior. Without any interaction the gaze selection is autonomously driven by the visual saliency and the memory of the gaze selection. A natural way for humans is to raise the attention by stepping into the field of view and waving at the system. This kind of visual motion cue works in humans from the earliest days of infancy [20]. Due to the chosen weights the system will immediately gaze in the direction of the detected motion. The motion cue can continuously be used in order to keep the gaze direction of the system oriented towards the waving hand. Continued waving while reducing the distance to the system finally leads to a hand position within the peripersonal space of the system defined by the disparity saliency selection. Again, due to the chosen weights the signal from the peripersonal space will dominate the behavior of the system. Practically this means that the system will continuously fixate the hand and what is in the hand of the user. This kind of behavior can be used in order to perform various tasks. For a schematic visualization of the space in front of the system see Figure 2.

Defining the peripersonal space as a body centered volume in the space in front of the system corresponds to the biological findings. Inducing attention shifts by objects within the peripersonal space also corresponds to biological data. As already discussed in section I, the full concept of peripersonal space includes action centered parts of the representation, but here we focus on the consequences for object recognition only. We are convinced that concepts like the peripersonal space ease the problem of object recognition or at least the learning and bootstrapping thereof.

The main problems for the recognition of rigid objects are translation, scale and 3D rotation invariance as well as invariance with respect to illumination changes and occlusion. If we perform the classification only within the peripersonal space, those invariance requirements are reduced to a large extent. Translation invariance is established by the gaze control fixating the 3D blob in the peripersonal space, while the depth information is used for improving scale invariance. Since the depth region is limited to a specific range, the occurring size variations are bound to a certain interval. The main invariances that have to be established by the classifier itself are 3D rotation, illumination changes, occlusion and the remaining position and size fluctuations that occur due to inherent fluctuations in the disparity signal.

IV. Object Recognition

We use a view-based approach to object recognition, where we perform the classification only on objects that enter the peripersonal space. The underlying object hypothesis is an isolated 3D blob within the disparity map that is segmented and used to compute a region of interest (ROI) centered around the blob. The size of the ROI is dependent on the estimated distance, computed from the average disparity of the blob to obtain a coarse size normalization of objects. Using the disparity blob simplifies the invariance requirements for the recognition system as pointed out in the previous section. The current output of the classifier is the identity of the recognized object with a confidence level. The classification is entirely learned by presenting the set of objects to be recognized. It represents an example of tuning a general system to solving a specific task by learning.

The object recognition module is based on the biologically motivated processing architecture proposed by Wersing & Körner [21], using a strategy similar to the hierarchical processing in the ventral pathway of the human visual system. Within this model, unsupervised learning is used to determine general hierarchical features that are suitable for representing arbitrary objects robustly with regard to local invariance transformations like local shift and small rotations. Object-specific learning is only carried out at the highest level of the hierarchy. This allows a strong speedup of learning, compared to other general purpose statistical classifiers, that need large amounts of training data for achieving robustness. As was shown in [21] this architecture is highly competitive with other current recognition methods, and can also be applied to the difficult case of segmentation-free recognition.

The input of the hierarchy is the region of interest (ROI), that is obtained from the left camera image using the disparity blob within the peripersonal space. This ROI is scaled to a defined size and provides the color input image for the following computation stages. The processing hierarchy is implemented as a feed-forward architecture with weight-sharing [22]. It is composed of a succession of feature-sensitive stages, denoted as S1 and S2 layer, and pooling stages, denoted as C1 and C2 layer (see Fig.3 and [21] for details). The output of the feature maps of the complex feature layer (C2) provides a general

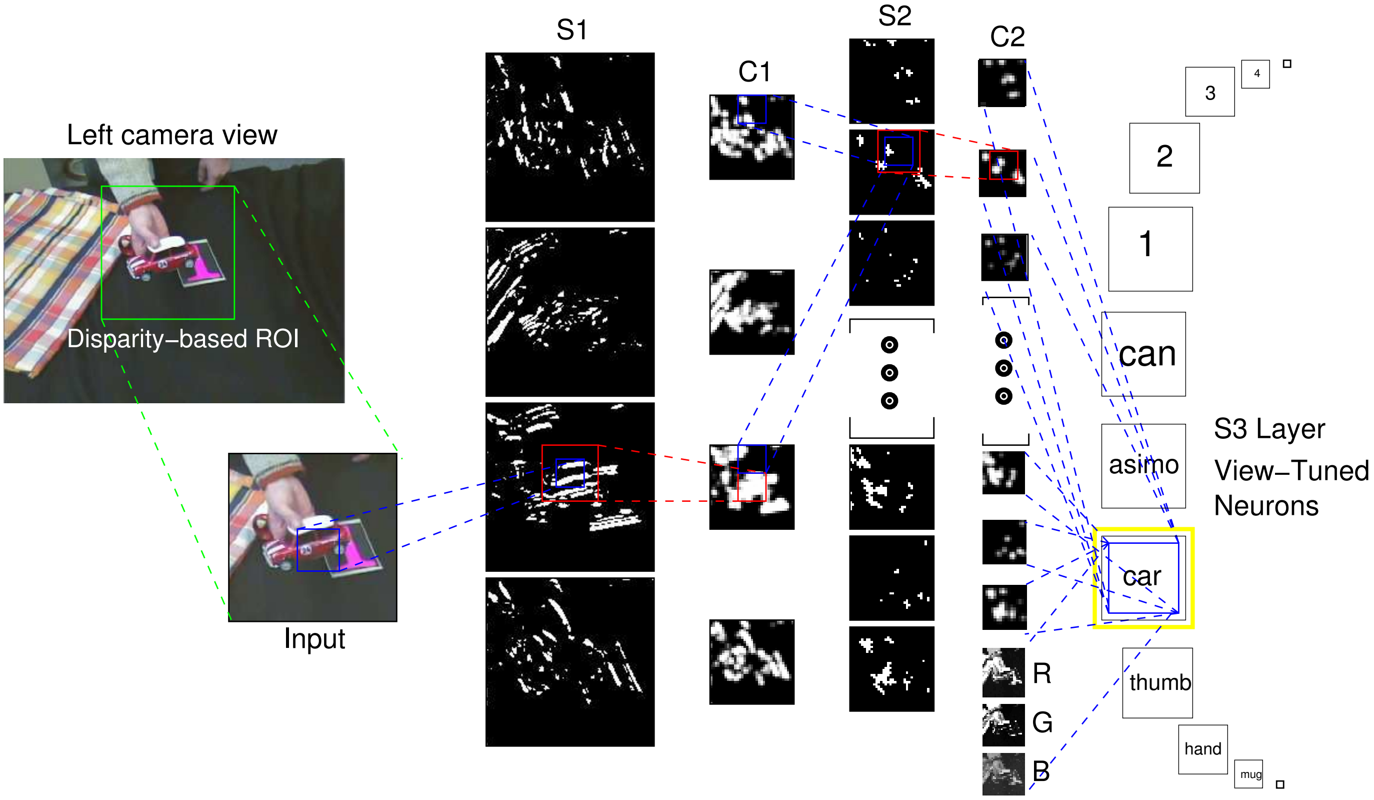


Fig. 3. The object recognition model. Based on the disparity computation, a region of interest (ROI) is extracted around an object within the peripersonal space and normalized in size to provide an input color image with size 144x144 pixels. Shape and color processing is first separated in the feature hierarchy and then fused in the view-based object representation. In the color-insensitive shape pathway the first feature-matching stage S1 computes an initial linear sign-insensitive Gabor-filter orientation estimation, a Winner-Take-Most mechanism between features at the same position and a final threshold function. The connected frames between the processing layers visualize the receptive fields of the local feature detectors. The C1 layer subsamples the S1 features by pooling down to a 36x36 resolution using a Gaussian receptive field and a sigmoidal nonlinearity. The 50 features in the intermediate layer S2 are trained by sparse coding and are sensitive to local combinations of the features in the planes of the previous layer. A second pooling stage in the layer C2 again performs spatial integration and reduces the resolution to 18x18. The color pathway consists of three downsampled 18x18 maps of the individual RGB channels that are added to the C2 feature maps. Classification is based on view-tuned neurons in the S3 layer, sensitive to the high-dimensional C2 activation patterns of a particular object and trained by supervised learning.

high-dimensional object representation that achieves a stronger view-based abstraction with higher robustness than the original pixel image [21]. Classification of an input image with a resulting C2 output is done in the final S3 layer by so-called view-tuned neurons that are obtained by supervised gradient-based training of a linear discriminator for each object, based on the C2 activation vectors of a training ensemble.

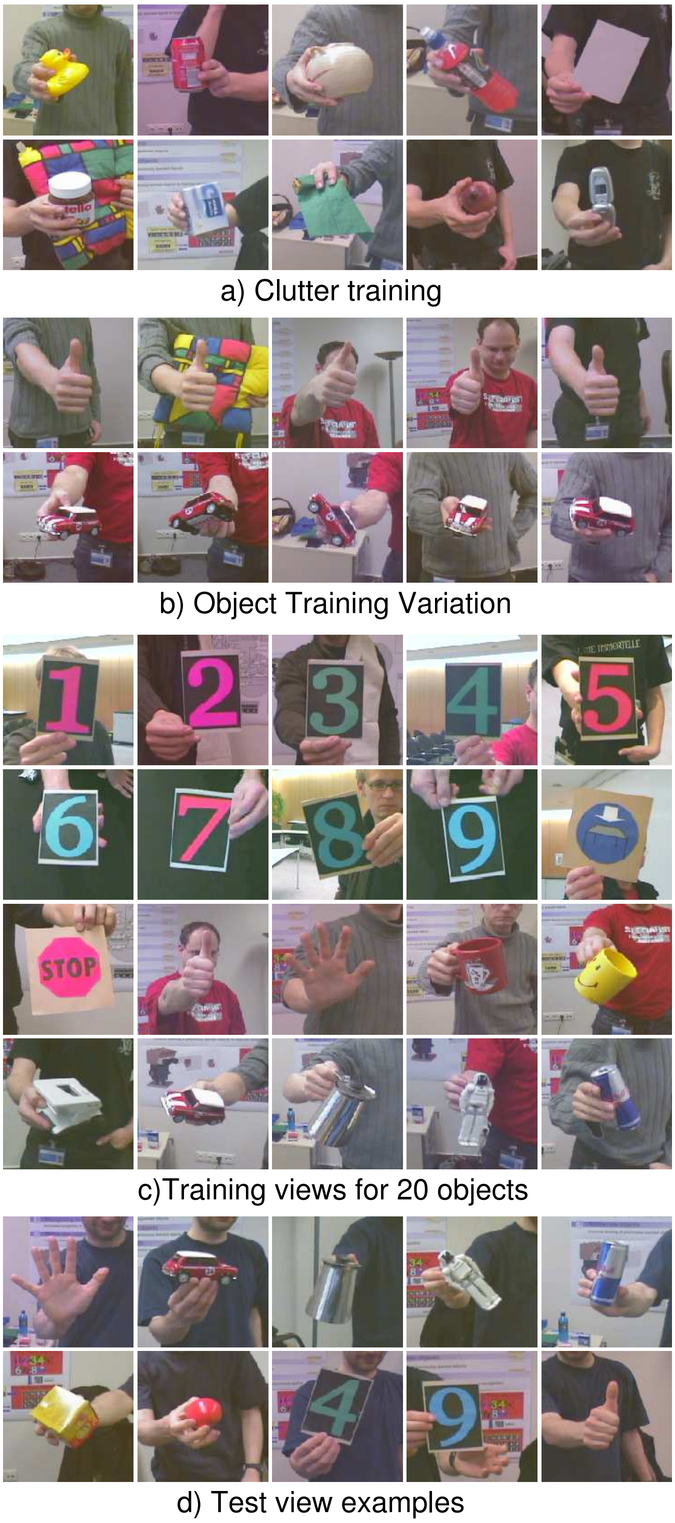
In the setting that we consider here, we perform no additional segmentation of the objects to be recognized. Figure 4 shows typical examples of the ROIs as they are being presented to the classifier. Training is done by showing 20 different objects with changing backgrounds and we expect the learning algorithm to automatically extract the relevant object structure and neglect the clutter in the surround. To demonstrate the generality of the recognition approach we use different types of visual object classes such as number cards, hand gestures, toys, and household objects. The results and details of the training are given in the experiments section.

V. EXPERIMENTS

For our experiments we use a stereo camera head with anthropometric dimensions as shown in Figure 2. It is called TUBBY. It has a pan and a tilt degree of freedom and represents an eyes-on-shoulder construction. This platform is sufficient for our current experiments. The work presented in [23] provides a uniform interface for working with ASIMO and transferring the results to the robot.

All experiments (training and evaluation) are done interactively with TUBBY. The same system is used for both phases. The internal labels of the objects are numbers that are specified whenever a new object is being learned by the system.

We perform the training of the recognition system by showing 20 different objects within the peripersonal space and collect 600 training views for each object. The object ensemble consists of 11 sign cards, (only slight rotation deviation – max. 20 degrees around all axes), 2 hand gestures, and 7 freely rotated objects (see Fig.4). To obtain more variations, training is done by two people. Due to inherent fluctuations of the disparity signal, the objects are only coarsely centered within



a) Clutter training

b) Object Training Variation

c)Training views for 20 objects

d) Test view examples

Fig. 4. Training and test data for the recognition model. The images are the full region of interest that is passed to the recognition model, centered around the disparity blob in peripersonal space. a) Training images for rejection. b) Rotation variation for gestures and objects. c) All 20 different objects. The sign cards were only rotated about 20 degrees around all axes. d) Examples from the test ensemble.

the input image, and size varies about $\pm 20\%$. Note that the objects occupy only about $5-15\%$ of the ROI area, and there is no segmentation information available. The recognition system has to learn to separate objects from background entirely based on the training views. Additionally we collect 1000 views of other objects for the rejection training. For this training ensemble of 13000 views, the corresponding C2 activations are computed (with a dimensionality of 53x18x18=17172), and the S3 view-tuned neurons are trained as linear discriminators for each object within this C2 feature space (see Figure 3). This training takes about 30 minutes. After the training the combined system of attention, disparity computation, and recognition runs at a frame rate of 2 Hz on a 3 GHz Dual Xeon computer. The recognition, based on the ROI input takes approximately 80ms.

To investigate the generalization performance of the recognition model, we recorded an independent set of test views with a third person, that did not participate in the training. For testing we collected 100 images for each object plus additional 1000 clutter images for rejection. The results of the trained recognition system on the test ensemble are shown in the form of an ROC plot that shows the trade-off between false positives (clutter classified as object) and false negatives (objects erroneously rejected as clutter). The plot is obtained for each object by tuning the recognition threshold from low to high values. We achieve less than 5% detection error at the point of equal false-positive and false-negative rate for almost all objects. The only exceptions are the can (20%), toy asimo (8%) and the metallic coffee can (7%). The overall classification error is 7.2 %, when we assign the class of the maximally activated view-tuned neuron as the classifier output.

Ude & Cheng [6] proposed a related approach to foveated object recognition, which is based on a color blob detection with affine warping and a final object detection step using support vector machines (SVM). They use an ellipsoidal region of interest around the object center to ensure that the amount of clutter in the image is small. Using 5 objects with 200 training images each, they report an average false negative rate of 4.5% at a false positive rate of 0.2%. If we compare these numbers to our results we note that we achieve a slightly worse detection rate, however, we use a larger number of 20 objects. We do not perform rotational normalization and no special segmentation during learning or recognition apart from a centering of the object. Note also that we included rather similar objects also in the rejection set, to increase the difficulty of the detection task.

We performed a baseline comparison, using the original RGB images with dimensionality 144x144x3 and utilize a nearest-neighbour classifier for classifying the test images with all labeled 12000 object training images. This can be considered as a control experiment to assess the base similarity in the used image ensemble. Using the plain original image data, the overall nearest-neighbour classification error is very large at 77.5%. This again underlines the advantages of the hierarchical C2 feature representation for representing object appearance in a general and robust way.

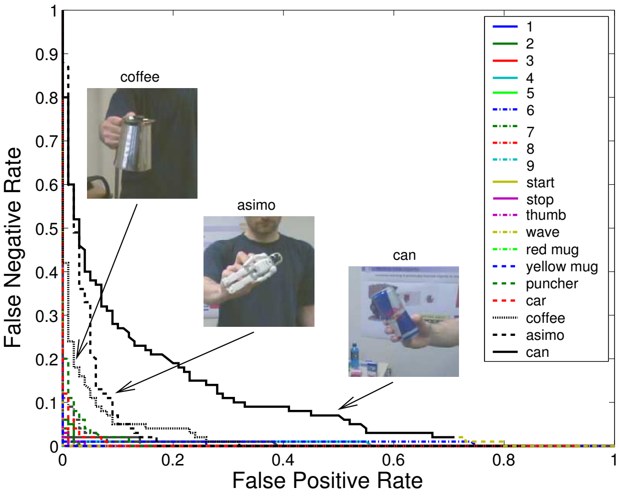


Fig. 5. Object detection performance on the test data. With the exception of three objects, all objects can be recognized with less than 5% error at the point of equal false positive rate (clutter classified as object) and false negative rate (object classified as clutter). The system has also learned to robustly ignore the background around the trained objects.

VI. Conclusion

In this paper we have presented a way to facilitate visual object recognition in real world tasks. Those tasks include among others object recognition on humanoid robots. The facilitation is achieved by structuring the active vision process including attention and recognition as well as the included representations according to the biological concept of the peripersonal space. We have implemented and presented a technical system built according to the presented concepts. Our experiments have shown, that an initial object hypothesis based on a disparity blob within peripersonal space is sufficient for a supervised learning of object appearance. Using stereo-based disparities has the advantage of constructing a spatial object-hypothesis, which is more stable than simple appearance based approaches for segmentation. As a consequence of this we can train such different objects like hands, cups, and metal cans at the same time. This is an important step towards general object learning methodologies for teaching arbitrary objects to a humanoid robot. We compared our results to the state of the art, keeping in mind that the comparison of active autonomous systems is difficult per se. Future work will focus on utilizing the full concept of peripersonal space comprising perception and dexterous manipulation.

Acknowledgment

The authors would like to thank their collaborators at the Honda Research Institute Europe at Offenbach for their constructive feedback concerning the draft of this paper.

References

[1] R. Pfeifer and C. Scheier, *Understanding Intelligence.* MIT Press, 1999.

[2] A. Sloman, "Architectures for human-like machines," Talk at Goldsmiths, January 2005. [Online]. Available: http://www.cs.bham.ac.uk/research/cogaff/talks/goldsmiths.pdf

[3] P. Fitzpatrick, G. Metta, L. Natale, and G. Sandini, "Learning about objects through action - initial steps towards artificial cognition." in *Proceedings of the IEEE International Conference on Robotics and Automation, Tapei, Taiwan*, 2003.

[4] A. M. Arsenio, "Object recognition from multiple percepts," in *Proceedings of IEEE-RAS/RSJ International Conference on Humanoid Robots*, 2004.

[5] S. Nolfi and D. Marocco, "Active perception: A sensorimotor account of object categorization," in *From Animals to Animats 7: Proceeding on the Sixth International Conference on Simulation of Adaptive Behavior*, 2002.

[6] A. Ude and G. Cheng, "Object recognition on humanoids with foveated vision," in *Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanois 2004), Los Angeles*, 2004.

[7] D. Ballard, "Animate vision," *Artificial Intelligence*, vol. 48, no. 1, pp. 1–27, 1991.

[8] F. Kaplan and V. V. Hafner, "The challenges of joint attention," in *Proceedings of the 4th International Workshop on Epigenetic Robotics, Genoa, Italy*, ser. Lund University Cognitive Science Studies 117, 2004, pp. 67–74.

[9] A. Couyoumdjian, F. D. Nocera, and F. Ferlazzo, "Functional representation of 3d space in endogenous attention shifts," *The quaterly Journal of Experimental Psychology*, vol. 56a, no. 1, pp. 155–183, 2003.

[10] Honda Motor Internet Page, http://www.world.honda.com/ASIMO.

[11] J. J. Gibson, *The Ecological Approach to Visual Perception*. Houghton Mifflin Company, Boston, 1979.

[12] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annual Review Neuroscience*, vol. 27, pp. 169–192, 2004.

[13] A. Maravita and A. Iriki, "Tools for the body (schema)," *TRENDS in Cognitive Science*, vol. 8, no. 2, pp. 79–86, 2004.

[14] P. H. Weiss, J. C. Marshall, G. Wunderlich, L. Tellmann, P. W. Halligan, H.-J. Freund, K. Zilles, and G. R. Fink, "Neural consequences of acting in near versus far space: a physiological basis for clinical dissociations," *Brain*, vol. 123, pp. 2531–2541, 2000.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. [Online]. Available: citeseer.nj.nec.com/itti98model.html

[16] SRI International, "Small vision system," http://www.ai.sri.com/software/SVS.

[17] T. Rodemann, F. Joublin, and E. Körner, "Saccade adaptation on a 2 dof camera head," in *Third Workshop on SelfOrganization of AdaptiVE Behavior (SOAVE 2004) Ilmenau*, H.-M. Groß, K. Debes, and H.-J. Böhme, Eds. VDI-Verlag Düsseldorf: Fortschrittsberichte des VDI, 2004, pp. 94–103.

[18] I. Mikhailova and C. Goerick, "Conditions of activity bubble uniqueness in dynamic neural fields," *Biological Cybernetics*, vol. 92, pp. 82–91, 2005.

[19] T. Shibata, S. Vijayakumar, J. Conradt, and S. Schaal, "Biomimetic oculomotor control," *Adaptive Behavior*, vol. 9, no. 3/4, pp. 189–207, 2001.

[20] P. J. Kellman and M. E. Arterberry, *The Cradle of Knowledge: Development of Perception in Infancy*. MIT Press, 1998.

[21] H. Wersing and E. Körner, "Learning optimized features for hierarchical models of invariant recognition," *Neural Computation*, vol. 15, no. 7, pp. 1559–1588, 2003.

[22] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Networks*, vol. 1, pp. 119–130, 1988.

[23] M. Gienger, H. Janßen, and C. Goerick, "Task-oriented whole body motion for humanoid robots," in *Humanoids 2005, Tsukuba, Japan*, 2005.