Proc. Int. Conf. on Artif. Neur. Netw. ICANN 2005, Warsaw. pp. 487-492

Online Learning for Object Recognition with a Hierarchical Visual Cortex Model

Stephan Kirstein, Heiko Wersing, and Edgar Körner

Honda Research Institute Europe GmbH
Carl Legien Str. 30
63073 Offenbach am Main, Germany
{stephan.kirstein,heiko.wersing,edgar.koerner}@honda-ri.de

Abstract. We present an architecture for the online learning of object representations based on a visual cortex hierarchy developed earlier. We use the output of a topographical feature hierarchy to provide a view-based representation of three-dimensional objects as a form of visual short term memory. Objects are represented in an incremental vector quantization model, that selects and stores representative feature maps of object views together with the object label. New views are added to the representation based on their similarity to already stored views. The realized recognition system is a major step towards shape-based immediate high-performance online recognition capability for arbitrary complex-shaped objects.

1 Introduction

Although object recognition is a long-studied subject in computer vision, the main focus in research has been so far on achieving optimal recognition performance on selected data sets of object images. Since the training of a recognition system is normally done offline, the efficiency of learning with regard to the learning speed has been considered less relevant in most approaches, leading to typical training times from several minutes to hours. Another problem is that most powerful classifier architectures like the multi layer perceptrons or support vector machines do not allow online training with the same performance as for offline batch training. Due to the lack of rapid learning methods for complex shapes, research in man-machine interaction for robotics dealing with online learning of objects has mainly used histogram-based feature representations that offer fast processing [5, 1], but only limited representational and discriminatory capacity. An interesting approach to online learning for object recognition was proposed by Bekel et al. [2]. Their VPL classifier consists of feature extraction based on vector quantisation and PCA and supervised classification using a local linear map architecture. Image acquisition is triggered by pointing gestures on a table, and is followed by a training phase taking some minutes.

We suggest to use a strategy similar to the hierarchical processing in the ventral pathway of the human visual system to speed up the object learning process considerably. The main idea is to use a sufficiently general feature representation

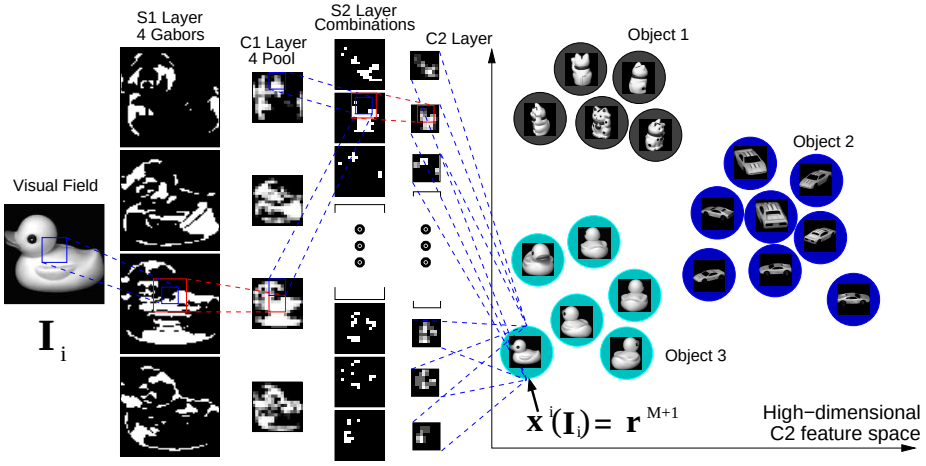

Fig. 1. The visual hierarchical network structure. Based on an image $\mathbf{I}_i$, the first feature-matching stage S1 computes an linear sign-insensitive receptive field summation, a Winner-Take-Most mechanism between features at the same position and a final threshold function. We use Gabor filter receptive fields, to perform a local orientation estimation in this layer. The C1 layer subsamples the S1 features by pooling down to a quarter of the original resolution in both directions using a Gaussian receptive field and a sigmoidal nonlinearity. The features in the intermediate layer S2 are sensitive to local combinations of the features in the planes of the C1 layer, and are thus capable of detecting more complex feature combinations in the input image. We use sparse coding for unsupervised training of these so-called combination feature neurons. A second pooling stage in the layer C2 again performs spatial integration and reduces the resolution by one half in both directions. Object representatives are learnt using an incremental vector quantization approach with attached class labels. Representatives $\mathbf{r}^k$ are computed as the output $\mathbf{x}^i(\mathbf{I}_i)$ of the hierarchy and added based on sufficient Euclidean distance in the C2 feature space to previously stored $\mathbf{r}^k$ of the same object.

that remains unchanged, while object-specific learning is accomplished only in the highest levels of the hierarchy. We perform online learning of objects using a short-term memory and similarity-based incremental collection of templates using the intermediate level feature representation of the proposed visual hierarchy from [6]. After a short introduction to our processing and memory model in Sect. 2, we demonstrate its effectiveness for an implementation of real-time online object learning in Sect. 3, and give our conclusions in Sect.4.

2 Hierarchical Visual Processing Model

Model Architecture. The visual hierarchical model proposed in [6] is based on a feed-forward architecture with weight-sharing [3] and a succession of feature-sensitive and pooling stages (see Fig.1). For a comparision to other recent feed-forward models of recognition see [6]. The output of the sparse feature repre-

sentation of the complex feature layer (C2) is used to incrementally build up the appearance-based object representation with an incremental vector quantisation model. These extracted C2 features are sensitive to coarse local edge combinations like e.g. t-junctions and corners. Given a set of N input images $\mathbf{I}_i$, $i = 1, \ldots, N$, the feature map outputs of the C2 layer of the hierarchy are computed as $\mathbf{x}^i(\mathbf{I}_i)$. The labeled object information is stored in a set of M representatives $\mathbf{r}^k$, $k = 1, \ldots, M$, that are incrementally collected. We define R_l as a set of representatives $\mathbf{r}^k$ that belong to object l. The acquisition of templates is based on a similarity threshold S_T. New views of an object are only collected into the object representation if their similarity to the previously stored templates in R_l is less than S_T. The parameter S_T is critical, characterizing a compromise between the object representation accuracy and the computation time. We denote the similarity between view $\mathbf{x}^i$ and representative $\mathbf{r}^k$ by A_{ik} and compute it based on the quadratic Euclidean distance in feature space by $A_{ik} = \exp\left(-(\mathbf{x}^i - \mathbf{r}^k)^2/\sigma\right)$. Here, σ is chosen for convenience such that the average similarity in a generic recognition setup is approximately 0.5.

Online Training. For one learning step the similarity A_{ik} between the current training vector $\mathbf{x}^i$, labeled as object l and all representatives $\mathbf{r}^k \in R_l$ of the same object l must be calculated and the maximum value is computed as $A_i^{\max} = \max_{k \in R_l} A_{ik}$. The training vector $\mathbf{x}^i$ and the corresponding class label will be added to the object representation if $A_i^{max} < S_T$. Assuming that M representatives were present before, we then choose $\mathbf{r}^{M+1} = \mathbf{x}^i$. Otherwise we assume that the vector $\mathbf{x}^i$ is already sufficiently represented by one $\mathbf{r}^k$, and do not add it to the representation.

Online Recognition. Recognition of a test view $\mathbf{I}_j$ is done with a nearest neighbour search of the hierarchy output $\mathbf{x}^j(\mathbf{I}_j)$ to the set of representatives. In contrast to the training, the similarity A_{jk} must be calculated between the current C2 feature vector $\mathbf{x}^j$ and all representatives $\mathbf{r}^k$. The class label of the winning representative $\mathbf{r}^{k^j_{\max}}$ with $k^j_{\max} = \arg\max_k(A_{jk})$ is then assigned to the current validation vector $\mathbf{x}^j$. Due to the non-destructive incremental learning process, online learning and recognition can be done at the same time, without a separation into training and testing phases.

Rejection. For a real application of the online learning system e.g. in a robot interaction scenario it is crucial to reach good classification results, but also unknown objects and clutter should be rejected. This can be done based on the similarity of a test view to the winning representative. Due to the different structural complexity of the appearance variation of different objects, the rejection can be largely improved by choosing the detection threshold similarity dependent on an estimate of the object complexity. This can be estimated by the average number of non-zero elements of the C2 feature vectors.

3 Experimental Results

For our experiments we use a setup, where we show objects, held in hand with a black glove, in front of a black background. Images are taken with a cam-

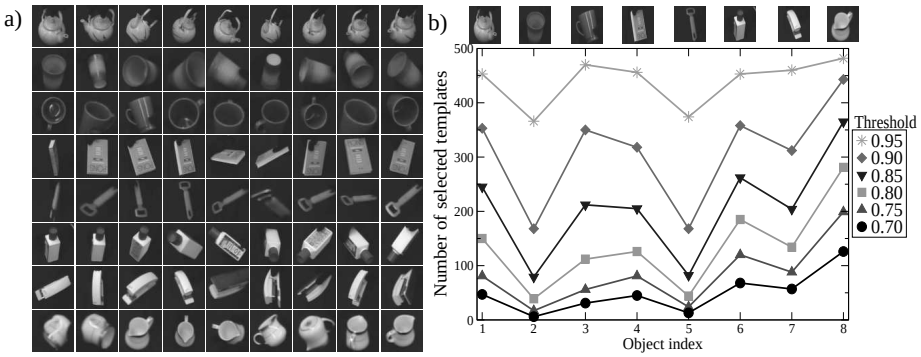

Fig. 2. Test images and the number of selected representatives for different similarity thresholds. (a) Some example images of the eight freely rotated objects, taken in front of a dark background and using a black glove for holding, causing also some minor occlusion effects. The difficulty of this database is the rotation of objects around three axes. Additionally some object views are only partially segmented. (b) Number of selected representative vectors for changing similarity thresholds. The selected number strongly depends on the shape-dependent appearance variation of the objects.

era, segmented using local entropy-thresholding, normalized in size (each view is 64x64 pixel large) and converted to grey scale. We show each object by rotating it freely by hand for a few ten seconds, which results in 500 input images $\mathbf{I}_i$ per object. Another set of 500 images for each object is recorded for validation. Some rotation examples are shown in Fig.2a. The difficulty of this training ensemble is the high variation of objects during rotation around three axes and the sometimes only partially segmented object views (e.g. mug and cup).
Figure 2b shows how the similarity threshold S_T influences the number of selected representatives. It can be seen that this number strongly depends on the complexity of the object, i.e. shape-dependent appearance variation.
The first investigation of training time should demonstrate how long it takes to incrementally train one object using a real camera. The training speed is limited by the frame rate of the used camera (12,5 Hz) and the computation time needed for the entropy segmentation, the extraction of the corresponding sparse C2 feature vector $\mathbf{x}^i$ with 3200 dimensions and the calculation of similarities A_{ik} (see Sect.2). For the shown curves of the teapot and the cup we trained all other seven objects and incrementally trained the teapot or cup as the eighth object. Figure 3a shows how long it takes until the newly added object can be robustly separated from all other objects.
We also investigated how fast our model performs on a saved image ensemble, without the limitation of the camera's frame rate and the segmentation. For this exploration all eight objects are trained in parallel. We started with a training ensemble of 25 training views for each object and determined the needed training time and the classification rate. Afterwards we increased the number of training views for each object in 25 view steps until all 500 training views for each object

are reached. Figure 3a (labeled with "database") shows that the training phase takes less than 1 minute for a training ensemble of $8 \cdot 500 = 4000$ object views. We compared the classification results (see Fig.3b) of our model with a normal nearest neighbour classifier (NNC), where *every* training vector $\mathbf{x}^i$ is directly used as a representative and a one-layered sigmoidal network trained by a gradient-based supervised learning on the C2 feature vectors $\mathbf{x}^i$. The sigmoidal network consists of an input and output layer, without hidden layers. For every object we used one output node, whereas each node has a linear scalar product activation and a sigmoidal transfer function. We trained these networks with all training vectors or with the selected representatives of our model. Figure 3b shows that the exhaustive NNC and our model using C2 activations perform quite equal, which means that our model can reduce the number of relevant representatives (2906 representatives are selected from 4000 training views (72,65%), for 64x64 pixel images and $S_T = 0.92$) without losing classification performance. The sigmoidal networks perform in comparison to our model slightly worse, but the representational resources are strongly reduced. We used only one linear discriminating weight vector per object, i.e. 8 weight vectors. We also trained these networks with the selected representatives of our model, which speeds up the training process but also slightly reduces the classification rates. Further we show that the usage of C2 features considerably increases the classification performance of our model compared to the original grey value images (see Fig.3b). The rejection capability of our model was tested with all 8 trained objects and 16000 different grey value clutter images. These images were randomly cut out of large scenes and contain parts of different objects, portions of buildings, faces and so on. The reached false negative rate for the 8 trained objects was 7.9%, whereas 8.3% of all clutter images are classified as an object.

4 Conclusion

We have shown that the hierarchical feature representation is well suited for online learning using an incremental vector quantization model approach. Of particular relevance is the technical realization of the appearance-based online learning of complex shapes for the context of man-machine interaction and humanoid robotics. This capability introduces many new possibilities for interaction scenarios and can incrementally increase the visual knowledge of a robot.
The simple template-based representation of objects in our approach allows a simple incremental buildup of the representation during online-learning. Nevertheless, due to the exhaustive storage of high-dimensional feature map information a similar approach seems prohibitive given the, arguably large, but finite neural resources in the brain. We have already investigated that a later offline refinement of the representation like supervised gradient-based training can be used to reduce the representational effort considerably. Such a process could be used to guide the transfer from a simple photographic short-term memory presented here to a more optimized long-term memory representation. The in-

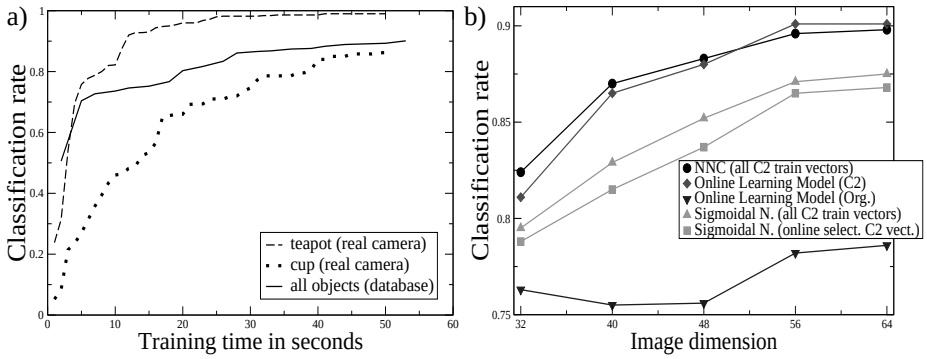

Fig. 3. Classification rate over training time and a comparison between nearest neighbour classifier (NNC) (all training vectors are used), our online learning model and sigmoidal networks for different image dimensions. (a) Classification rate dependent on needed computation time for the whole training set (labeled with "database") and the training of one object with a real camera. The training speed using a camera is also limited by the frame rate of the camera and the segmentation. Good recognition performance can be achieved within 20-30 seconds of online training. (b) Comparison of classification rates between a simple NNC (all training vectors are used), our model (C2 activations, original images) and sigmoidal networks for different image sizes. It can be seen that the classification rates of the NNC and our online learning model using the C2 activation are more or less equal and that the one-layered sigmoidal networks perform slightly worse. The classification results of our model using the C2 features are distinctly better than the results using the original grey value images.

vestigation of appropriate models that are related to the representation of visual representation in the inferotemporal cortex [4] will be the subject of future study.
Acknowledgments: We thank C. Goerick, M. Dunn, J. Eggert and A. Ceravola for providing the image acquisition and processing system infrastructure.

References

1. Arsenio, A.: Developmental learning on a humanoid robot. Proc. IJCNN 2004, Budapest.
2. Bekel, H., Bax I., Heidemann G., Ritter H.: Adaptive Computer Vision: Online Learning for Object Recognition. Proc. DAGM 2004 Springer C. E. Rasmussen et al. (2004) 447–454
3. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36 (4)** (1980) 193–202
4. Tanaka, K.: Inferotemporal cortex and object vision: stimulus selectivity and columnar organization. Annual Review of Neuroscience, **vol. 19** (1996) 109–139
5. Steels, L., Kaplan, F.: AIBO's first words. the social learning of language and meaning. Evolution of Communication, **vol. 4, no. 1** (2001) 3–32
6. Wersing, H., Körner, E.: Learning Optimized Features for Hierarchical Models of Invariant Object Recognition. Neural Computation **15 (7)** (2003) 1559–1588