Honda Research Institute Europe GmbH
https://www.honda-ri.de/

An Integrated System for Incremental Learning of Multiple Visual Categories

Stephan Kirstein, Heiko Wersing, Horst-Michael Groß, Edgar Körner

2008

Preprint:

This is an accepted article published in Proceedings of the International Conference on Neural Information Processing (ICONIP). The final authenticated version is available online at: https://doi.org/[DOI not available]

An Integrated System for Incremental Learning of Multiple Visual Categories

Stephan Kirstein[1,2], Heiko Wersing[2], Horst-Michael Gross[1] and Edgar Körner[2]

[1] Ilmenau University of Technology
Neuroinformatics and Cognitive Robotics Lab
P.O.B. 100565, 98684 Ilmenau, Germany
{stephan.kirstein,horst-michael.gross}@tu-ilmenau.de
[2] Honda Research Institute Europe GmbH
Carl-Legien-Str. 30 63073 Offenbach am Main, Germany
{heiko.wersing,edgar.koerner}@honda-ri.de

Abstract. We present a biologically inspired vision system able to incrementally learn multiple visual categories by interactively presenting several hand-held objects. The overall system is composed of a foreground-background separation part, several feature extraction methods and a life-long learning approach combining incremental learning with category specific feature selection. In contrast to most visual categorization approaches where typically each view is assigned to a single category we allow labeling with an arbitrary number of shape and color categories and also impose no restrictions to the viewing angle of presented objects.

1 Introduction

An amazing capability of the human visual system is the ability to learn an enormous repertoire of visual categories. This large amount of categories is acquired incrementally during our life and requires at least at the beginning the direct interaction with a tutor. Inspired by child-like learning we propose an architecture for learning several visual categories in an incremental and interactive fashion. The architecture is composed of several building blocks including segmentation, feature extraction, a category learning module and user interaction, which allow training of categories based on natural hand-held objects.

The development of online and life-long learning systems became more and more popular in the recent years e.g. [9], [11] or [14]. The work of [9] allows online learning and detection of hand-held objects in cluttered scenes based on a combination of a background model and a tracking method but is restricted to static camera settings like security cameras. Of particular interest is the work proposed by [11], because it targets for a similar interactive category learning task as investigated in this paper but relies on a simple learning method which can only be applied to categories with little appearance changes.

The learning system proposed in this paper is related to earlier work that enables object recognition of complex shaped objects presented by hand in cluttered scenes [14]. Especially the preprocessing steps are therefore similar, but several modifications to the feature extraction, the learning method and the user interaction were necessary to allow the same functionality for learning categories based on natural objects. Natural objects typically belong to several different categories (e.g. red-white car), therefore a decoupled representation for each category (for category red, white and car) is required, which can not be handled by typical incremental learning systems dealing with classification tasks. This decoupling leads to a more condensed representation and higher generalization performance compared to classification architectures. Additionally several modifications with respect to the figure-ground segregation were applied.

Since several years many architectures dealing with object detection and categorization tasks have been proposed in the computer vision community. Interestingly most of these approaches are only based on local parts-based features, which are extracted around some defined interest points (e.g. implicit shape models (ISM) [8]) to build up object models for categories like faces or cars. The advantages of such models are their robustness against partial occlusion and scale changes, but also the ability to deal with clutter. The main drawback of these architectures is the restriction to the canonical view of a certain category, while objects in a natural environment usually occur in many different orientations. Typically such architectures also require long training phases to generate the object models, which make them unsuitable for interactive training. A recent work of [2] tries to overcome this speed limitation and proposes a combined approach of ISM [8] and a semi-supervised clustering algorithm which enables to incrementally build up object categories based on dialog interactions with a human teacher. Although the general approach is interesting because it minimizes the necessary interaction with a tutor, it still does not allow learning of categories from arbitrary viewpoints.

The manuscript is structured as follows: in Section 2 we describe the building blocks of our category learning system and show its ability to incrementally learn several visual categories in Section 3. Finally we discuss the results and related work in Section 4.

2 Incremental Category Learning System

In the following we describe the building blocks of our learning system (see Fig.1) composed of preprocessing, figure-ground segregation, and several feature extraction methods providing shape and color information. The core part of our category learning system is a life-long learning vector quantization method which is trained in direct interaction with a human tutor.

2.1 Preprocessing and Figure-ground Segregation

The input to our category learning system is a stream of image pairs taken by a camera head with pan-tilt unit and parallel aligned cameras. Depth information is calculated after the rectification of both camera images to correct lens distortion. This depth information is used to generate an object hypothesis in

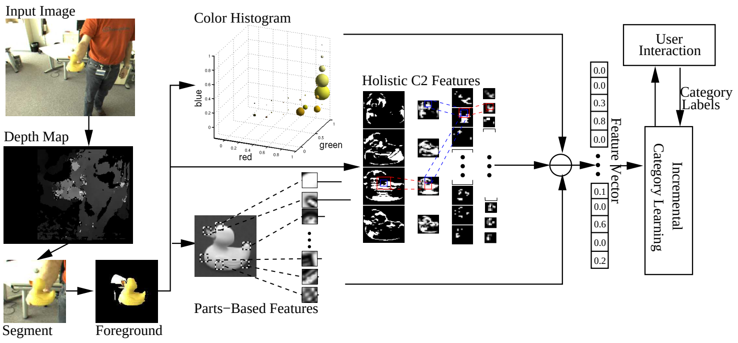

Fig. 1. Category Learning System. Based on an object hypothesis extracted from the depth map a figure-ground segregation is performed. The detected foreground is used to extract color and shape features. Color features are extracted as histogram bins in RGB space. In contrast to most other categorization approaches we combine holistic features obtained with a detection hierarchy and parts-based features based on SIFT-descriptors. All extracted features are concatenated into a single structureless vector which, together with the category labels provided by a human tutor, is the input of the incremental category learning module.

cluttered scenes, where we assume that things in the near range of the camera system are of particular interest for the system. The result of the preprocessing is a region of interest (ROI) used to localize and track segments in the scene.

This extracted segment still contains a substantial amount of background clutter, but for the incremental build up of category representations it is beneficial if such clutter is suppressed. Therefore we apply an additional figure-ground segregation as proposed by [1] to reduce this influence. The basic idea of this segregation method is to use learning vector quantization (LVQ) to learn a predefined number of distinct prototypes for foreground and background where the noisy depth information is used as initial hypothesis. The learning of these prototypes is based on different feature maps consisting of RGB-color features as well as the pixel positions. Instead of the standard Euclidean metrics, for the distance computation additional prototype specific relevance factors are calculated based on generalized matrix LVQ [10]. These relevance factors are adapted online and dynamically weight the different maps to maximize the margin between foreground and background. The output of this segregation step is a binary mask defining the foreground. In the following processing steps only foreground pixels are used to extract category specific features.

2.2 Feature Extraction

Color Features. For the representation of color information we use the common histogram binning method which combines robustness against view and scale changes with computational efficiency [12]. Overall 6x6x6=216 histogram bins within the RGB space are used, where typically a small amount of features are specific for a complete color category.

Shape Features. The shape features are obtained by a hierarchical feed-forward architecture and parts-based feature detectors. The feature detectors of the hierarchical feed-forward architecture are obtained by unsupervised learning, providing a set of general but less category-specific features, while the parts-based features are trained supervised with respect to category specificity. We combine these different shape features to show the ability of the category learning method to select appropriate features out of a large amount of possible candidates. Such feature combinations are rare because most categorization methods rely on parts-features only, but in offline experiments we recognized an increase in categorization performance, when both extraction methods are combined compared to using them individually.

The hierarchical feed-forward architecture is based on weight-sharing and a succession of feature detection and pooling stages (see [13] for details). The first feature-matching layer S1 is composed of four orientation sensitive Gabor filters. Additionally a Winner-Take-Most mechanism between features at the same position and a final threshold function is applied. The following 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 50 features in the intermediate layer S2 are obtained by sparse coding and are sensitive to local combinations of the features from the C1 layer. The layer C2 again performs spatial integration and reduces the resolution to a half in each direction, resulting in 50 18x18 sparsely activated feature maps with a total dimensionality of 16200, where typically less than 10% of all features are non-zero.

As parts-based feature detectors a set of preselected SIFT-descriptors is used as proposed by [4]. For each new object view the response of those detectors is calculated at each location in the image using the dot product as similarity measure. The maximum response per feature detector is kept and stored in an activity vector, which neglects all spatial information. The offline feature selection scheme follows the approach described in [4], where all SIFT-descriptors of each training image are clustered into 100 components. Out of the large number of resulting clusters an iterative scheme selects at each step a SIFT-descriptor as new detector until a given number is reached (e.g. 100 in our case). The choice of detectors is based on the highest additional gain for a certain shape category.

Combined Feature Representation. All extracted features of an image are combined into a single structureless and sparsely activated feature vector $\mathbf{x}^i = (x_1^i, ..., x_F^i)$, with resulting feature space dimensionality of $F = 16516$. The task of the category learning method, described in the next section, is to automatically select a category specific feature subset, which best represents the category without the given knowledge which features contain color or shape information.

2.3 Incremental and Interactive Category Learning

The learning of visual categories is based on a limited and changing set of labeled training vectors, which are stored into a short term memory (STM). The category learning method must be able to extract the category informations from the STM and conserve this information in the long term memory (LTM) representation. To achieve this transfer an incremental exemplar-based network is combined with a forward feature selection method. This allows life-long learning and enables a separation of cooccuring visual categories based on selected category-specific feature sets, which most exemplar-based networks can not handle.

Short Term Memory. This memory type is similar to the online learning vector quantization method developed earlier [5]. In contrast to the naive approach, where each view is stored in the STM a similarity calculation is performed based on all vectors with identical category label list. New views are only added to the STM if the similarity to such vectors is below a specified insertion threshold S_T. Based on this simple selection schema it could be shown [5], that the number of training views can be reduced by about 30% without losing generalization performance, and reducing the LTM training time. Additionally we assume a limited memory size of the STM, which requires a deletion heuristic of feature vectors if the capacity limit is reached. Therefore STM vectors are deleted which belong to the same category label list and for which almost no categorization errors occur. Such vectors are already successfully transferred to the LTM and can be deleted without information loss.

Long Term Memory. The general idea of the category LVQ (cLVQ) method is to iteratively make small changes to the representation of erroneous categories. After a change the performance gain is calculated, based on all available training vectors. If the increase is larger than a threshold ϵ the feature or LVQ node is permanently added to the representation and otherwise removed.

For guiding the feature testing we use a statistical feature scoring method (see [3] for an introduction to feature selection methods) as proposed in [6], where a single scalar r_{fc} for each category c and feature f is calculated to estimate the category specificity. Similar to the development of children, the feature scoring is only based on previously seen exemplars of a certain category and can strongly change if further information is encountered. Therefore we continuously update the feature scoring values to follow this change and also allow in rare cases the deletion of selected features. The feature testing itself selects predominately those features which occur almost exclusively for a certain category and also are often present in the most categorization errors. In parallel to the incremental selection of features also new LVQ nodes are inserted and tested. For the node insertion we propose to insert new LVQ nodes based on the training vectors with most categorization errors. This typically leads to an improvement for several categories and generates a compact representation, which is a major requirement for fast and interactive learning.

For the LTM representation each cLVQ node $\mathbf{w}^k$ is assigned to a vector $\mathbf{u}^k = (u_1^k, ..., u_C^k)$ of arbitrary color and shape categories, where each $u_c \in \{-1, 0, +1\}$ labels a $\mathbf{w}^k$ as positive or negative example of a category. The third state $u_c = 0$ means unknown category membership and is required to incrementally resolve categorization errors, while imposing no representational changes to error free categories. For the calculation of winning node $\mathbf{w}^{k_{\min}(c)}(\mathbf{x}^i)$ for category c the

Euclidean distance computation is combined with dynamic metric adaptation, based on the obtained feature scoring values:

$$||\mathbf{x}^i - \mathbf{w}^k||^2_{\lambda_c} = \sum_{f=1}^{F} \lambda_{cf}(x^i_f - w^k_f)^2, \tag{1}$$

where $\lambda_{cf} = r_{fc}$ for all features f selected for category c, while all other $\lambda_{cf} = 0$ and thus have no influence on the distance computation. For the adaptation of the winning node $\mathbf{w}^{k_{\min}(c)}(\mathbf{x}^i)$ of category c, the standard LVQ learning rule [7] is used, but is restricted to the selected feature dimensions of category c. Additionally the learning rate is dependent on the node age, which allows strong modification of newly inserted nodes, while the representation of well-adapted nodes is conserved.

User Interaction. For interactively providing label information to the STM and LTM we use a simple state-based user interface. This user interface is based on a list of predefined labels, including some wild card labels, to allow the labeling of categories, for which no category label is defined. All labels can be provided to the system in any arbitrary order. In general the user interaction is composed of two operation modes. The default user interaction mode is that the learning system generates a hypothesis list of currently present categories, which afterward can be corrected or confirmed by the user. The other possibility is that the user directly provides category labels, to label previously unknown categories.

3 Experimental Results

For all experiments several complex shaped objects are freely rotated in front of our camera system. Based on the extracted features and the current category representation in the LTM the category decision which categories are currently detected is communicated to the user. This hypothesis generation is repeated until the user confirms or corrects the categorization decisions of the LTM representation, which then triggers the collection of new training vectors into the STM. This means our category learning method has two operation states, where in the one state it produces category hypotheses and in the other state, when confirmed category labels are available, it collects new training vectors in the STM, which later are transferred into the LTM representation. The incremental learning of the LTM representation is performed in both states and is even continued, if currently no new object views are presented, because this knowledge transfer typically takes much longer than the STM training.

Overall our system is distributed on three different 3 GHz CPUs running at a frame rate of 8-10 Hz, which is fast enough to show the desired incremental and life-long learning ability of our system. We consider two different scenarios summarized in Fig. 2. For the first scenario, we start with a blank STM and LTM representation, where we show how efficient the training of categories can be done based on only few representatives of the corresponding category, typically requiring less than 10 min training time. At this state the learning system is able

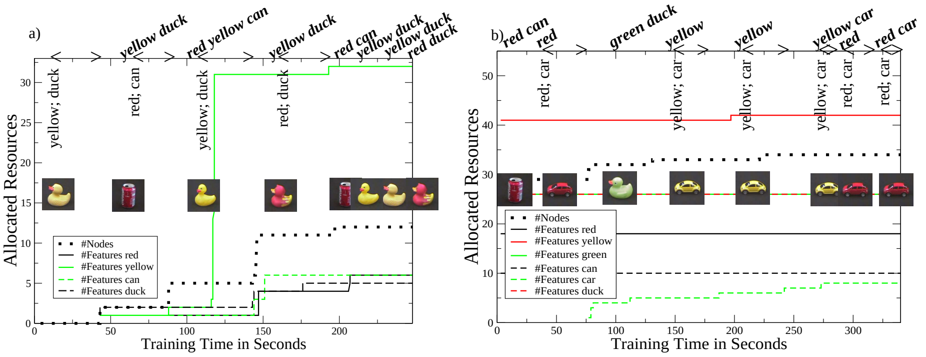

Fig. 2. Incremental learning of visual categories. The formation of selected features and allocated nodes are shown over time, while presenting different objects for starting with a) blank memory and b) at a later learning stage. It can be seen how many cLVQ nodes (dotted lines) are allocated over time, while additionally the change of selected features are shown for each known category (dashed and solid lines). We also added the categorization decisions of the learning system, communicated to the user on top of the figure with sloped text, while the confirmed category labels provided by the user are denoted underneath, separated by semicolons. Additionally the intervals where new training vectors are collected into the STM are marked with <>.

to detect the categories reliably for objects already presented to the system, but typically the generalization to completely new objects is quite poor. For the second scenario we used the representation after about one hour training time, where in the mean time more representatives of the trained categories were shown, which strongly improved the representation and also the generalization performance. Here it should be mentioned that even after more elaborate training the number of allocated features for each category is still small compared to the overall number of extracted features, which is a basic requirement for interactive training. Additionally we show that adding new categories to the representation can be done in a flexible way, without affecting the already known categories.

4 Discussion

We have presented a learning system able to interactively learn arbitrary visual categories in a life-long learning fashion. To our knowledge this is the first category learning system which allows category learning based on complex shaped objects held in the hand. Comparable architectures as proposed by [11] or [2] learn categories based on objects placed on a table, which simplifies the ROI detection and figure-ground segregation. Additionally it also strongly reduces the appearance variations of the presented objects and therefore makes the category learning task much easier. We also allow different categories for a single object, while typically the categories are trained independently.

We also could show that our learning system can efficiently perform all necessary processing steps including figure-ground segregation, feature extraction and incremental learning. Especially the ability to handle high dimensional but sparse feature vectors is necessary to allow interactive and incremental learning, where often additional dimension reduction techniques like the PCA are required to allow online learning. This high feature dimensionality is also challenging for the used feature selection method, because of the large amount of possible feature candidates, but still the learning system is able to extract small sets of category specific features out of many possible feature candidates.

Acknowledgments: We thank S. Hasler and A. Denecke for providing the parts-based feature extraction and figure-ground segregation method. Additionally we thank both for the assistance of integrating their work into the system.

References

1. Denecke, A., Wersing, H., Steil, J.J., Körner E.: Robust Object Segmentation by Adaptive Metrics in Generalized LVQ. Proc. ESANN (2008) 319–324
2. Fritz, M., Kruijff, G-J. M, Schiele, B.: Cross-Modal Learning of Visual Categories using Different Levels of Supervision Proc. ICVS Conference (2007)
3. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research **3** (2003) 1157-1182
4. Hasler, H., Wersing, H., Körner, E.: A Comparison of Features in Parts-based Object Recognition Hierarchies. Proc. ICANN (2007) 210–219
5. Kirstein, S., Wersing, H., Körner, E.: A Biologically Motivated Visual Memory Architecture for Online Learning of Objects. Neural Networks **21** (2008) 65-77.
6. Kirstein, S., Wersing, H., Gross, H. M., Körner, E.: A Vector Quantization Approach for Life-Long Learning of Categories. Submitted to ICONIP 2008.
7. Kohonen, T.: Self-Organizing and Associative Memory. Springer Series in Information Sciences, Springer-Verlag, third edition (1989)
8. Leibe, B., Leonardis, A., Schiele, B.: Combined Object Categorization and Segmentation with an Implicit Shape Model. ECCV'04 Workshop on Statistical Learning in Computer Vision (2004)
9. Roth, P.M., Donoser, M., Bischof, H.: On-line Learning of Unknown Hand Held Objects via Tracking. Proc. Int. Conf. on Computer Vision Systems (2006)
10. Schneider, P., Biehl, M., Hammer, B.: Relevance Matrices in LVQ. In Similarity-based Clustering and its Applications to Medecine and Biology (2007).
11. Skočaj, D., Berginc, G., Ridge, B., Štimec, A., Jogan, M., Vanek, O., Leonardis, A., Hutter, M., Hawes, N.: A System for Continuous Learning of Visual Concepts. Proc. ICVS (2007)
12. Swain, M. J., Ballard, D. H.: Color Indexing. Int. J. of Computer Vision **7 (1)** (1991) 11-32
13. Wersing, H., Körner, E.: Learning Optimized Features for Hierarchical Models of Invariant Object Recognition. Neural Computation **15 (7)** (2003) 1559–1588
14. Wersing, H., Kirstein, S., Goetting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J.J., Ritter, H., Körner, E.: Online Learning of Objects in a Biologically Motivated Architecture. Int. J. of Neural Systems **17** (2007) 219-230