Improving Online Learning of Visual Categories by Deep Features

Lydia Fischer
Bielefeld University
Honda Research Institute Europe

Stephan Hasler
Honda Research Institute Europe

Sebastian Schrom
Technical University of Darmstadt

Heiko Wersing
Honda Research Institute Europe

Abstract

Recognition of visual categories is a key component of any human-robot interaction dealing with objects. The interactive learning of new categories based on human demonstration provides great opportunities for more flexible capabilities of robots that can extend their pre-trained knowledge towards novel requirements in a new operation environment. In this article we first present an earlier developed system capable of real-time interactive visual category learning on a humanoid robot. We then show that learning speed and performance can be strongly enhanced by employing a visual feature computation hierarchy based on deep learning. The main result is that an architecture pre-trained for object identification provides very good generalization to more complex categorization tasks, which can be motivated from concepts like chorus-of-prototypes models (Edelman, 1995).

1 Introduction

Interactive machine learning aims at tasks where a deeper cooperation of humans and machines is beneficial. Amershi et al. [1] review a wide range of approaches to incorporate human interaction for accelerating learning and improving performance of machine learning models. We can identify two important challenges: (i) the machine should solve its task satisfactorily and securely, and (ii) communication should make the behaviour of the cooperating participants understandable and informative. Areas of application are online learning and life-long learning which are nowadays especially interesting for personalisation [13] and big data analysis.

Another highly relevant area is interactive learning for robots to learn new concepts about their environment and human partners [11, 3]. Affordances have been considered for defining acquired categories relevant for robot actions [8]. Kirstein et al. [6] have developed an interactive visual category learning architecture which combines an intuitive speech-based user interaction with a biologically motivated short-term and long-term memory system for categorizing objects presented in hand in arbitrary 3D poses. The architecture was integrated in the autonomous learning and interaction system (ALIS) [4] for the Honda humanoid robot Asimo[1].

Categories acquired through interactive learning can provide efficient means of communication and task cooperation between an intelligent robot and a human. This leads not only to a better performance but also to a higher acceptance and trustworthiness of the system which is an important ingredient when humans and machines should work together. The invariant[2] recognition of visual categories is a

[1] An internet search for "learning asimo" leads to a BBC feature about this system.

[2] Here invariant refers to object transformations like 3D rotation and scaling.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

complex problem, which no current online system can directly learn from scratch. A hierarchical solution can be to separate the learning architecture into lower feature-representing stages and higher view-dependent representations, resembling the architecture of the human visual cortex [14]. In this article we first review the architecture of the interactive learning system from Kirstein et al. [6] and demonstrate a considerable improvement of this system by employing feature representations derived from a deep learning architecture.

2 An interactive category learning system

The system [6] performs real-time interactive learning of visual categories. During training the user rotates an object in hand in front of the stereo camera and tells the system a set of corresponding visual category labels, e. g. 'yellow green box'. Based on extracted color and texture features the system tries to learn the different concepts simultaneously, without any a priori assignment of features to categories (i. e. assigning color features to color descriptions has to be learned from the input statistics). In the next training step the user can present a previously seen or new object. The system categorizes the new visual input and tells the user the detected categories or responds with 'unknown category'. The user can confirm or correct the labels and the system will update its category representation using the new information.

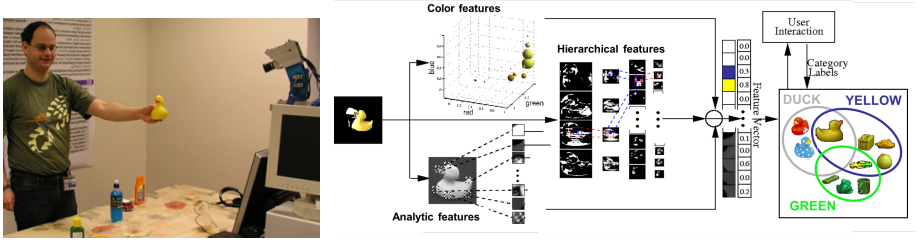

Figure 1: Category training setup and feature extraction for interactive online category learning [6].

The framework is based on the concept of shared attention in time and space. The system uses a stereo-based depth proximity cue to determine the relevant region containing the object for each input image. As long as the user holds an object close to the camera, the acquired views are assumed to belong to the same object and to have the same category labels. The size-normalized relevant region is shown to the user as feedback and is the input for the further processing steps.

The potential category labels are predefined in the grammar of a commercial speech recognition module. A speech-based confirmation process is used to deal with wrongly understood words. At any time during the presentation of an object the user can tell labels to the system. The system will repeat the understood labels and ask for confirmation ('Did you say: yellow green box?') expecting the user to respond with 'Yes, this is correct.' or 'No, this is wrong.'. If one of these phrases is understood by the system it will respond with 'Okay', otherwise it will ask the user to repeat the confirmation.

To deal with the resource constraints of an interactive online learning process, different memory concepts are combined. The feature vectors of each view of an object are first stored in the quick but size-constrained short term memory (STM) and are associated with a label vector, containing a 1 for each present category and a -1 for absent categories, i. e. labels the user did not assign during the presentation of the object. The long term memory (LTM) tries to efficiently model the categories by a minimal set of relevant prototypes. These prototypes are selected from the STM, and are later updated using learning vector quantization [7]. Objects are deleted from the STM if they are sufficiently well modelled in the LTM.

For each view of a presented object, the system detects the presence or absence of a category based on the closest prototype in the LTM. Hence for each view of the sequence category memberships are determined. The categories that are detected in most views are communicated to the user. The sequence is determined from a dynamic tracking of the object, also indicating a removal of the object from the scene. The user can either confirm the system output, add missing category labels, or provide a completely new set of labels, where the latter one is necessary whenever the system has a false positive. In each of the three cases, the system uses the new data together with the confirmed or corrected labels to update the category representation.

3 Experiments using deep features

Kirstein et al. [6] used different features for the real-time build-up of the visual category representation: Hierarchical features learned from unsupervised sparse coding [14], analytical features learned from supervised greedy optimization of object discrimination capability [5] and color histograms. Due to the recent great success of deep neural networks, we were curious if deep neural network features can provide an improvement for our online learning architecture. We present the result of this investigation in the following. To make a controlled experimental comparison to previous results of Kirstein et al. [6] we use the same recorded data ensemble and offline category experiment definition.

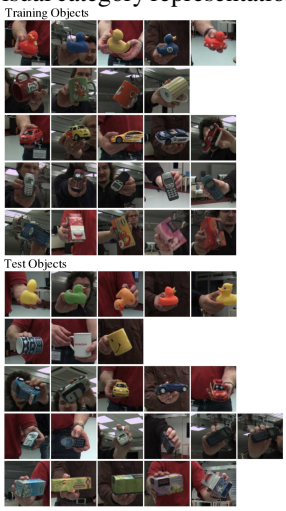

Figure 2: The data [6] split into training/testing.

Data set [6]: We use the so-called unconstrained database [6] containing 48 objects (e. g. red car) which belong to different shape (rubber duck, cup, car, cell phone, box) and color (yellow, green, blue, white, red) categories. For each object there exist 1200 different views in a cluttered office environment. Each object is freely rotated by hand covering almost the full viewing sphere (Fig. 2) with hand-caused occlusions.

Training of deep features: As deep feature hierarchy we use the *AlexNet* [9], a convolutional neural network with 8 layers; the first 5 layers are convolutional and the remaining 3 are fully-connected. The original network was trained to distinguish 1000 classes. Initialized from this weight setting, we fine-tuned the network to discriminate those 126 objects[3] which were also used by Hasler et al. [5] for the supervised training of analytic features. The training of the feature representation is therefore optimized to identify individual objects, i. e. generate a sparse output vector which only should be active for the corresponding object and silent for all other objects. Here, we analyse if this representation optimized for identification will naturally lead to good categorization performance, which is a more difficult problem for categories with high variability (e. g. chairs).

Feature computation for categorization experiment: For our experiments we extracted for each input view the activations from the last three fully-connected deep layers as feature vectors where we omitted the *ReLU* non-linearity (Tab. 1). Each feature vector is attached with a ten-dimensional category label vector where each element encodes if the view of the object *belongs to* a category or *does not belong to it* or if the membership is *unknown*. The data is divided into disjoint 24 training objects and 24 test objects to investigate category generalization across individual objects.

Table 1: Deep feature layers

name	layer	dimension
fc6	6th	4096
fc7	7th	4096
fc8	8th	126

Experimental protocol: We use the same online learning protocol as [6] that simulates a subsequent training of 24 objects with their respective categories. In each training epoch[4] a random object with its 1200 views is added to a capacity limited memory (containing at most three objects) which is used to update the category representation. In case there are already three objects stored, the new object replaces the oldest training object. After each of the 24 epochs the categorization performance is calculated for the complete test set. We report the average over ten runs.

Experiments with deep features: A baseline categorization performance using the deep feature representation can be obtained with a simple nearest neighbour (NN) classifier. Since all training data points/views are stored as prototypes, in each training epoch the data of a randomly chosen training object is added as prototypes. The winner takes all rule is applied for classification for each category. Prototypes with unknown category membership are omitted. Figure 3 shows the averaged categorization performance per epoch. As can be seen, the different deep features (fc6, fc7, fc8) lead to similar results for shape and color categories. It seems that especially the shape categories are more challenging in the beginning of the training while the end performance of shape and color categories is comparable for fc6, fc7, and fc8 features.

[3]The 126 objects included 39 objects of the online learning data set. This setting resembles the experimental setup of [6] at which the 126 objects where used for feature leaning in an object identification task as well. The learnt features are then used for learning the categorisation task on the 48 objects.

[4]The online system is running at 10 Hz frame rate, resulting in roughly 2 minute training time per object.

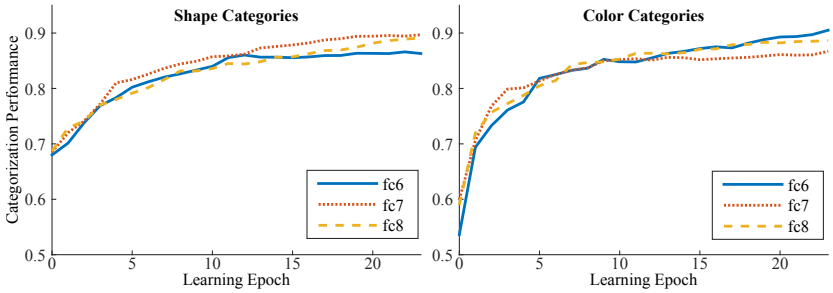

Figure 3: Average categorization performance (over 10 runs) with respect to different deep features (fc6, fc7, fc8 features of the adapted *AlexNet* [9]) using a nearest neighbour classifier.

Comparative evaluation: In this section we compare the results of the prior system [6] with the new results obtained from deep features together with a simple NN classifier (Fig. 3). Because of the similar behaviour of the deep features, we use the low dimensional fc8 features for the comparison only which can be seen in Figure 4. The new deep feature representation achieves a consistent and considerable improvement over the previous feature representation with respect to fast and robust acquisition of categories. We note that the actual categorization representation layer in [6] is different to the simple nearest-neighbour classifier by performing a vector quantization-based compression of representation nodes to speed up the categorization. Future work would thus have to consider a similar memory consolidation for more efficient prototype storage or alternative final categorization architectures[5]. Apart from that the results show that generally the online category learning performance can be considerably enhanced by using pre-trained deep features.

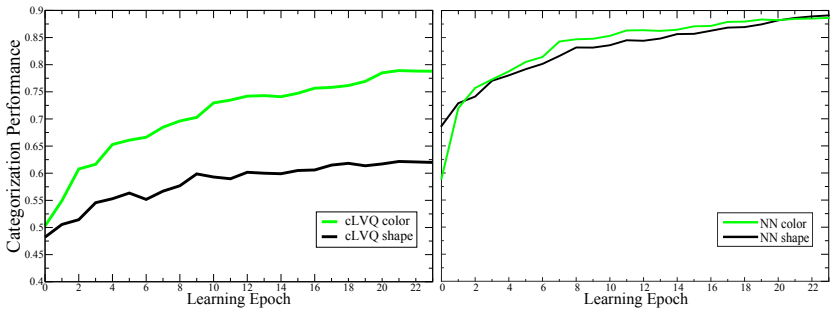

Figure 4: Comparison of categorization performance. Left: The results of Kirstein et al. [6]. Right: A nearest neighbour classifier (NN) using the *fc8* features of the adapted *AlexNet* [9].

4 Conclusion

We demonstrated that using deep feature representations trained in an object identification task generalize well to performing online incremental learning of visual categories. This principle has been long identified as one of the key representation advantages of learning-driven distributed processing systems [12], and our study showed the same effect. Kubilius et al. [10] validate recently that neurons in intermediate layers of a deep network for object identification develop shape sensitivity.

From a visual representation point of view it is interesting to note that even the last layer of the deep network which is trained to respond with a 0/1 mapping for the object identification task is suitable for an essentially different categorization scenario. This can only work, if the residual difference to the target 0/1 responses after the deep training for similar objects (according to categorization) is generalizing well for the later categorization task. Effectively this realizes a representation where categories can be represented by similarity to a number of prototypes, which has been discussed as the chorus-of-prototypes model in vision [2]. This concept is thus also highly applicable to improving interactive incremental category learning.

[5] A non-optimized plain NN classifier Matlab implementation takes 0.3 s for categorizing one frame.

References

[1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[2] Shimon Edelman. Representation, similarity, and the chorus of prototypes. *Minds and Machines*, 5(1): 45–68, 1995.

[3] Christian Goerick, Inna Mikhailova, Heiko Wersing, and Stephan Kirstein. Biologically motivated visual behaviors for humanoids: Learning to interact and learning in interaction. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pages 48–55. IEEE, 2006.

[4] Christian Goerick, Bram Bolder, Herbert Janssen, Michael Gienger, Hisashi Sugiura, Mark Dunn, Inna Mikhailova, Tobias Rodemann, Heiko Wersing, and Stephan Kirstein. Towards incremental hierarchical behavior generation for humanoids. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*, pages 248–255. IEEE, 2007.

[5] Stephan Hasler, Heiko Wersing, Stephan Kirstein, and Edgar Körner. Large-scale real-time object identification based on analytic features. In *International Conference on Artificial Neural Networks*, pages 663–672. Springer, 2009.

[6] Stephan Kirstein, Alexander Denecke, Stephan Hasler, Heiko Wersing, Horst-Michael Gross, and Edgar Körner. A vision architecture for unconstrained and incremental learning of multiple categories. *Memetic Computing*, 1(4):291–304, 2009.

[7] Teuvo Kohonen. *Self-Organization and Associative Memory*. Springer Series in Information Sciences, Springer-Verlag, third edition, 1989.

[8] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012.

[10] Jonas Kubilius, Stefania Bracci, and Hans P. Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput Biol*, 12, 2016.

[11] Andrea Lockerd and Cynthia Breazeal. Tutelage and socially guided robot learning. In *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 4, pages 3475–3480. IEEE, 2004.

[12] James L. McClelland and Timothy T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322, 2003.

[13] Barry Smyth, Maurice Coyle, Peter Briggs, Kevin McNally, and Michael P O'Mahony. Collaboration, reputation and recommender systems in social web search. In *Recommender Systems Handbook*, pages 569–608. Springer, 2015.

[14] Heiko Wersing and Edgar Körner. Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*, 15(7):1559–1588, 2003.