In: Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015, pp. 439-444, Michel Verleysen. Bruges, BE, April 22-24, 2015.

# Learning objects from RGB-D sensors using point cloud-based neural networks

Marcelo Borghetti Soares, Pablo Barros, German I. Parisi, Stefan Wermter *

University of Hamburg, Department of Computer Science,
Vogt-Koelln-Strasse, 30, 22527, Hamburg, Germany

**Abstract**. In this paper we present a scene understanding approach for assistive robotics based on learning to recognize different objects from RGB-D devices. Using the depth information it is possible to compute descriptors that capture the geometrical relations among the points that constitute an object or extract features from multiple viewpoints. We developed a framework for testing different neural models that receive this depth information as input. Also, we propose a novel approach using three-dimensional RGB-D information as input to Convolutional Neural Networks. We found F1-scores greater than 0.9 for the majority of the objects tested, showing that the adopted approach is effective as well for classification.

## 1 Introduction

The efficient recognition of parts of the environment is relevant for a variety of scenarios ranging from lower-level grasping and manipulation to higher-level robotic assistance. The way the human brain performs these activities is an attractive source of inspiration, due to its capacity to deal with noise, to operate in cluttered scenarios and to generalize from few examples. Concerning object recognition, some researchers have focused on viewpoint-independent recognition, assuming that this process can be accomplished using invariant features that must match three-dimensional representations of objects stored in the memory [1]. Others state that the brain stores multiple viewpoint representations of objects in the brain and uses strategies to interpolate and to generalize the results for viewpoints that are not initially presented [2].

Additionally, depth information can play an important role in recognition [3] and many approaches for scene understanding and object recognition have been developed in the last years taking into account the rich information provided by depth sensors. The impact of depth and RGB data sources was explored in [4], where the improvement obtained for object recognition tasks performed by a mobile robot was shown. Similar approaches have been tested in which different neural models were employed to treat different representations such as color, shape or depth [5],[6]. This paper differs from these previous approaches since we developed and evaluated different neural models that take into account the geometry of the point cloud. We consider that this three-dimensional information processed by the brain is particularly useful to distinguish parts, faces, etc, of an object under different environment conditions. This is potentially important to recognize categories

*This work is partially supported by CAPES Brazilian Federal Agency for the Support and Evaluation of Graduate Education (Process number 10441-13-1).

In: Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015, pp. 439-444, Michel Verleysen. Bruges, BE, April 22-24, 2015.

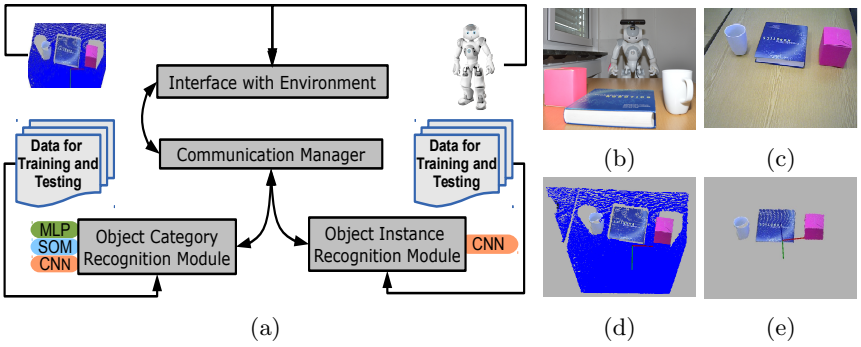



Fig. 1: (a) Scheme of the proposed framework divided in modules. (b) Scenario with a NAO robot and three objects. (c) RGB image from the RGB-D device. (d) Point cloud from RGB-D device. (e) Segmented objects.

and instances (different books, different cups, etc) of objects. Our contribution can be summarized as follows: i) a novel approach using three-dimensional RGB-D information as input to Convolutional Neural Networks and ii) an implementation that encompasses both invariant geometric features from the objects and features extracted from multiple viewpoints and that allows testing different neural models.

This article is organized as follows: Section 2 presents the framework used and the neural approaches developed, Section 3 is devoted to our experiments and Section 4 concludes and presents future directions.

## 2 Object recognition and interaction with the environment

A database was built containing RGB-D data from objects in different poses (defined as the objects $(x, y, z)$ coordinates and (*roll*, *pitch*, *yaw*) orientations in relation to the RGB-D device's viewpoint), since the robots should act in an environment and recognize objects regardless of the pose (no capture was made with robots in movement). Figure 1 shows a NAO humanoid robot (b) looking at examples of these objects on a table and the corresponding RGB image (c) and depth capture (d).

We are considering the objects located on the largest surface captured by the RGB-D device. We use RANSAC (RAndom SAmple Consensus) [7] to identify these planes. In our case, the $z$ axis points to the direction of the objects. Therefore, for segmentation, the scene is reoriented in a way that the $z$ axis becomes orthogonal to the normal vector of the plane. Thus, we identify all objects that have $y$ coordinate values higher than the average value of the $y$ coordinate of the points that compose the plane. Finally, we consider only objects located within a *tolerance distance* from the center of mass of the segmented plane (Figure 1d).

In: Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015, pp. 439-444, Michel Verleysen. Bruges, BE, April 22-24, 2015.

### 2.1 Framework for recognition

Figure 1a shows the system overview. The *Interface with Environment* provides an interface through which it is possible to manually select the objects previously segmented. The features extracted from the selected object are sent to the other modules. The *Communication Manager Module* is responsible for redirecting messages exchanged by different modules. Usually, the *Interface with Environment Module* will send descriptors captured from RGB-D data to recognition modules and these modules will send back the categorization/instantiation.

### 2.2 Recognition based on three-dimensional feature descriptors

In this case, the input to the neural network models is a feature vector obtained using VFH (Viewpoint Feature Histogram) [8] to collect a multidimensional descriptor representing the geometrical relations of the points that compose an object. We also use *Principal Component Analysis* to reduce the input vector. This descriptor is used for category recognition with two neural approaches: i) Feedforward Network (MLP+VFH) and ii) Self-organizing Map (SOM+VFH). The SOM adopts the labelling scheme of the output presented in [9].

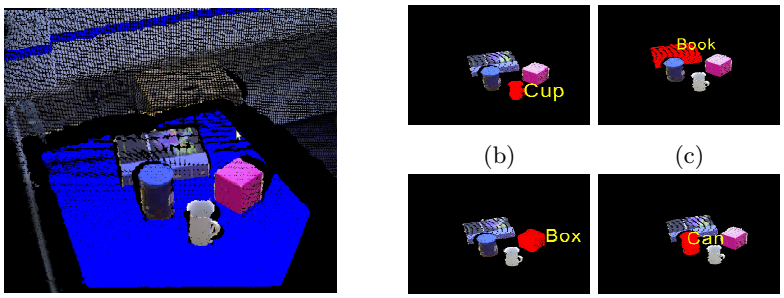
### 2.3 Recognition based on features from multiple viewpoints

We developed an approach that receives multiple viewpoints as input to a Convolutional Neural Networks (CNN) [10]. Generally, CNNs are composed of multiple layers divided in i) a set of $n$ feature maps (every map is the result of a convolution operation) and ii) a set of $n$ subsampling maps obtained from the feature maps. We developed and tested two different CNN approaches: i) 2DCNN: CNN with two-dimensional kernel and ii) S-3DCNN: CNN with three-dimensional kernel, where a three-dimensional kernel convolves with a stack of images [11]. As typical in applications that use RGB images, this stack is populated with similar images. In our case, for each object we generate $n$ sliced planes that are orthogonal to the $z$ axis (imagine a bread sliced by a knife). As the normal vector to the slice is parallel to the $z$ axis, each slice is represented as a projection in $x$-$y$ plane. This sliced object fills the stack of images, each slice occupying one position. The stack is used by the S-3DCNN preserving the sequence, as the geometrical relation between the parts of the object matters. Finally, these slices are then convolved with the three-dimensional kernel and the weights of this kernel are adjusted taking into account the position and sequence in the object.

## 3 Experiments

For our database, it was important to have objects observed from different viewpoints on tables and on the floor to comprise different situations in which the robots could act, since we aimed also testing how the robots generalize the results for different scenarios. We captured 5 different categories with 5 different objects for each category (instances) in 6 different viewpoints. In addition, this database also contains one instance of each category per object (used only in experiments of

In: Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015, pp. 439-444, Michel Verleysen. Bruges, BE, April 22-24, 2015.




(a) (b) (c) (d) (e)

Fig. 2: (a) Online experiment with four objects: (b) Cup, (c) Book, (d) Box and (e) Can.

category recognition): 5 different categories in 8 different viewpoints, 5 times for each viewpoint (to take into account noisy captures). The total number of images acquired was 350. The samples used for training were composed by single objects, but online tests with multiple objects were also conducted.

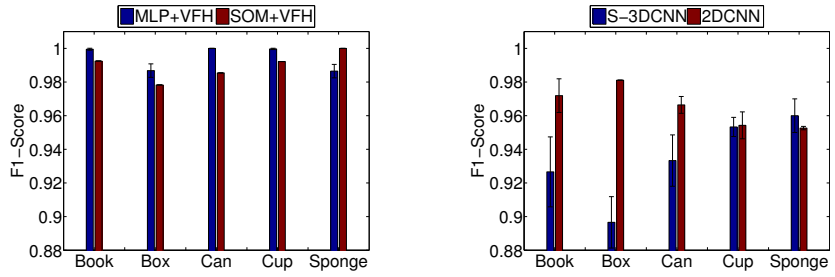
The number of units for the MLP, CNN and SOM was 100, 250 and 360 respectively. In the CNN, the number of feature maps in the first and second layer was 20 and 30 respectively. We used a kernel of $5 \times 5$ pixels. For all neural networks, the learning rate was 0.01. The number of slices in the S-3DCNN was 10. The training and testing sets were divided in 60% and 40%. We performed a systematic search in the parameter space and defined the values that performed better. Each experimental result presented below was obtained from an average over 5 simulations. This number of runs was chosen due to the computational time required and since the methods presented very stable results.

Figures 2(a)-(e) show an example of online recognition with four objects on a table using MLP+VFH. It is important to note that the robot can select one object each time, which is particularly interesting in cluttered environments. Also, the object partially occluded can be recognized.

Figure 3a shows the recognition results based on three-dimensional feature descriptors. We can note that both methods have good accuracy with F1-scores greater than 0.96. To test the performance under different viewpoints, we applied to each sample rotation in roll, pitch and yaw ($3 \times 3 \times 3$). Thus, we have 28 samples per point cloud (27 generated and 1 original). Considering the 350 samples of our database the total amount is 9800 ($28 \times 350$). For each sample generated, we also added noise in 10% of each of the clouds.

Figure 3b shows the recognition based on features from multiple viewpoints. The results have good accuracy with F1-scores greater than 0.89. In the case of the CNN, we also applied rotation (roll, pitch and yaw) to create a dataset of 9800 samples. The results obtained from Figure 3a and Figure 3b represent different methodologies (and theories) and should be analysed separately. The CNN approaches receive images (size $50 \times 50$ pixels) from different viewpoints that provide

In: Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015, pp. 439-444, Michel Verleysen. Bruges, BE, April 22-24, 2015.




(a) (b)

Fig. 3: F1-scores for each category: (a) MLP+VFH and SOM+VFH and (b) 2DCNN and S-3DCNN.

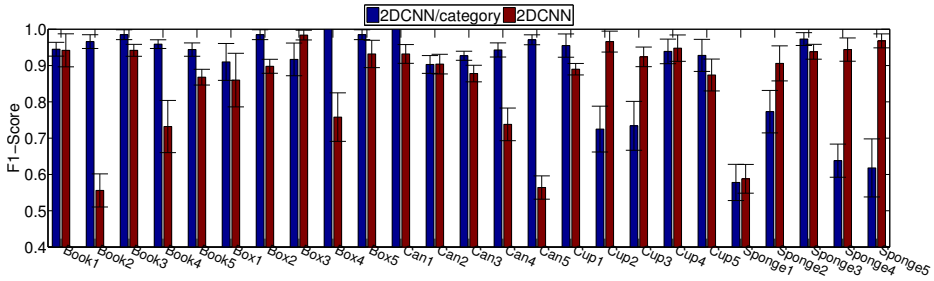



Fig. 4: F1-scores for each instance per category.

information in a slightly different way (projections of the cloud in the $x$-$y$ plane). We demonstrated with this experiment that the recognition based on features from multiple viewpoints offers an efficient alternative approach. The partial knowledge contained in each viewpoint can be used further for a final decision about the classification of one object. It is possible to use multiple "readings" from different viewpoints to recognize an object when the object is not clearly visible due to noise or distance factors, like animals in general do. Similarly, S−3DCNN has also promising results and requires investigation about how to use the partial knowledge contained in each slice, since they can lead to a different classification.

Instance classification results are shown in Figure 4. We choose 2DCNN since it presented better results in the previous experiments. CNNs can also extract features based on textures, which is ideal for instance classification. We used two approaches: i) one 2DCNN to classify all different instances and ii) five different 2DCNN per category. In general, the results were better or similar for the second case, indicating that the division of labour works. But there are 6 cases for which F1-scores were worse (considering the standard deviation): `Box3`, `Cup2`, `Cup3`, `Sponge2`, `Sponge4`, `Sponge5`. We believe that this behaviour was caused by the fact that some of the instances (for example the sponges) do not have strong geometrical and textural features or that they are too similar to be recognized with a limited number of samples.

In: Proceedings of 23th European Symposium on Artifical Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015, pp. 439-444, Michel Verleysen. Bruges, BE, April 22-24, 2015.

## 4 Conclusions and Future Works

The framework presented in this paper was developed keeping in mind that robots should understand the interactions between objects and other parts of the environment, such as supporting surfaces. To learn the objects from its RGB-D devices we developed and tested several neural models. The results enabled us to draw the conclusion that features obtained from multiple viewpoints are a rich source of information to be explored. Each view or slice from S-3DCNN potentially stores valuable information that can be used to improve the recognition. As future steps, we plan to integrate multiple views captured over time and evaluate the improvement in the recognition accuracy.

## References

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[2] H. H. Bülthoff, S. Y. Edelman, and M. J. Tarr. How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5(3):247–260, 5 1995.

[3] D. DeAngelis. Roles of visual area MT in depth perception. In Michael S. Gazzaniga, editor, *The Cognitive Neuroscience*, pages 483–498. MIT press, 2009.

[4] L. C. Caron, Y. Song, D. Filliat, and A. Gepperth. Neural network based 2D/3D fusion for robotic object recognition. In *European Symposium on Articificial Neural Networks (ESANN)*, pages 431–436, 2014.

[5] L. A. Alexandre. 3D Object Recognition using Convolutional Neural Networks with Transfer Learning between Input Channels. In *13th International Conference on Intelligent Autonomous Systems*, volume 301 of *Advances in Intelligent Systems and Computing Series*, Padova, Italy, July 2014. Springer.

[6] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, pages 746–760, Berlin, Heidelberg, 2012. Springer-Verlag.

[7] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[8] R. B. Rusu, N. Blodow, and M. Beetz. Fast Point Feature Histograms (FPFH) for 3d Registration. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, ICRA'09, pages 1848–1853, Piscataway, NJ, USA, 2009. IEEE Press.

[9] G. I. Parisi, P. Barros, and S. Wermter. FINGeR: Framework for interactive neural-based gesture recognition. In *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25*, pages 443–447, 2014.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[11] P. Barros, S. Magg, C. Weber, and S. Wermter. A Multichannel Convolutional Neural Network for Hand Posture Recognition. In *International Conference on Artificial Neural Networks*, pages 403–410, 2014.