2009 21st IEEE International Conference on Tools with Artificial Intelligence

Using Growing Neural Gas Networks to Represent Visual Object Knowledge

Guillermo S. Donatti, Rolf P. Würtz
International Graduate School of Neuroscience and Institut für Neuroinformatik
Ruhr-Universität, 44780 Bochum, Germany
{guillermo.sebastian.donatti, rolf.wuertz}@neuroinformatik.rub.de

Abstract

We present a so-called Neural Map, a novel memory framework for visual object recognition and categorization systems. The properties of its computational theory include self-organization and intelligent matching of the image features that are used to build their object models. Its performance for representing the visual object knowledge comprised by these models and for recognizing unknown objects is measured using three different types of image features, which extract different granularity of information from object views of the ETH-80 image set. The obtained experimental results slightly outperform previous ones using PCA-based methods on the same image set, and they suggest that the medium-sized image features maximize the object models' informativeness and distinctiveness.

1 Introduction

A fundamental aspect of perception is the processing of visual information in relation to accumulated world knowledge. This can be subsumed under the processes of *object recognition* [12] deciding about an object's unique identity, and *object categorization* calculating an object's kind [8]. Although humans can determine easily the correspondence of objects in a natural scene with the ones previously seen, this remains a very challenging task for artificial systems.

The complexity of modeling these tasks comes from the fact that the space of all possible views of all objects is prohibitively large, which results in a high disparity between the known and the newly encountered object views. This variability can be grounded on the fact that objects in natural scenes are observed from different viewing positions. Additionally, the objects' shape can vary considerably both inter-category and intra-category. Objects in natural scenes are also not isolated, but normally seen against different backgrounds, interacting with more objects, and sometimes partially occluded by some of them. Furthermore, objects are subjected to photometric effects including the position and distribution of light sources in the scene, their wavelengths, the effects of mutual illumination with other objects, and the distribution of shadows and specularities [11]. In existing artificial systems, each one of these possible variations applied to a known object view generates a different object view and discerning their conceptual equivalence is not a trivial task to accomplish.

Consequently, a large variety of artificial object recognition and categorization systems motivated in biology have been proposed. These systems use either a feature-based [7, 2] or a correspondence-based approach [14, 5, 15]. In both, the processing of an object view relies on the extraction of image features together with the use of stored object models derived from training object views. The first ones classify object views by detecting which features are present disregarding spatial relations. These models usually fail when confronted with complex backgrounds, multiple objects, or occlusion. The latter ones store object models as ordered arrays of local features, which are matched with object views by solving the correspondence problem. Those models perform better on realistic images, but require more processing time than feature-based ones.

A major limitation of these approaches is that the inherent structure of their proposed object models is derived from the bottom-up process of visual information, with little or no use of top-down knowledge. The resulting artificial systems are either limited to represent a narrow range of object types, or they are in conflict with neuropsychological and neuroanatomical observations [4], as well as with the results of psychophysical experiments [9] and computational studies [11]. A recently proposed artificial system dynamically generates its object models using a more balanced approach [14]. This method yields high recognition rates, still it performs rather poorly on categorization and further research on its visual object knowledge representation is needed to improve its performance and robustness.

The present work proposes combining a Growing Neural Gas (GNG) network [3] with a classifier motivated by the population coding and decoding processes of cortical neurons [10] in a so-called *Neural Map*, which is capable

1082-3409/09 $26.00 © 2009 IEEE
DOI 10.1109/ICTAI.2009.81

of generating a structural association for the elementary image features that serve as components of dynamically generated object models. During training, the relationships between a given set of image features extracted from training object views are automatically established through an unsupervised learning process according to their similarity. Throughout recall, these relationships are exploited by the classifier to retrieve the best matching model image features for a given set of image features extracted a from a test object view.

The detailed description of the above mentioned processes is organized as follows: initially, Section 2 introduces object views and the types of image features extracted from them. Section 3 describes the training and recall processes of the proposed Neural Map together with its components and main characteristics. Section 4 defines two experiments to evaluate the performance of the Neural Map for representing the visual object knowledge derived from the training object views. Finally, Section 5 discusses the experimental results and outlines future research steps.

2 Feature Extraction

2.1 Object views

The object views used for the present work are extracted from the ETH-80 image set [6], a subset of the COGVIS database particularly designed as a basis for both psychophysical and computational studies concerning object categorization. It contains views and segmentation masks of 80 objects within a taxonomy composed of 8 basic level categories (i.e., cows, dogs, horses, apples, pears, tomatoes, cars, and cups) from 4 superordinate areas (i.e., animals, fruits and vegetables, human made big, and human made small). Each category contains 10 different individuals that are represented by 41 images from view-points spaced equally over the upper viewing hemisphere. The experiments described in Section 4 employ the images of the ETH-80 cropped-close version — some examples are depicted in Figure 1.

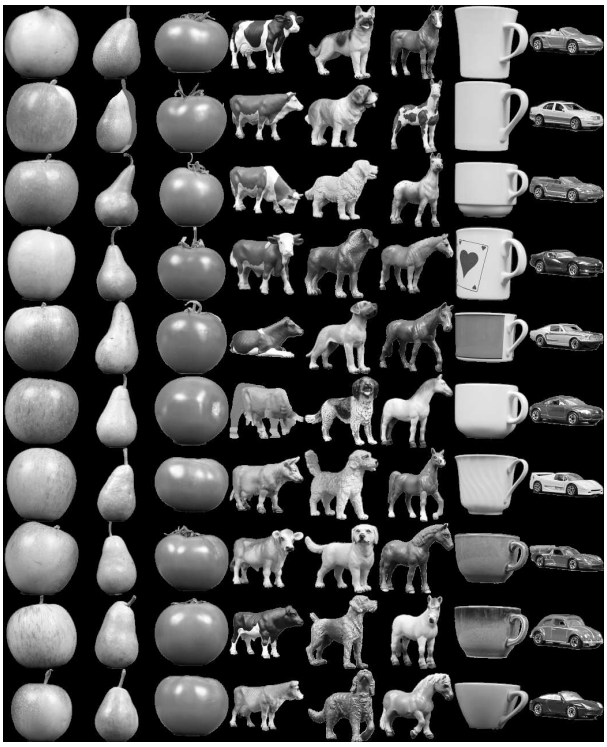


Figure 1. Samples of the ETH-80 cropped-close version extracted at 90° vertically and 45° horizontally. This version contains gray value images of single objects scaled to 128×128 pixels.

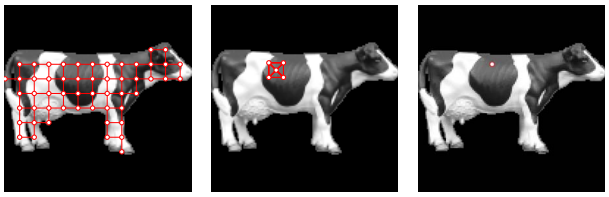
2.2 Image features

The present work uses three different local feature detectors, which represent patches of information derived from the object view at different granularity. They consist of the complex responses of a set of Gabor filters from one or more jets [5]. In the feature extraction process, each detector initially places a 10×10 pixels grid over the image, and the (known) segmentation mask is used to discard the background pixels. The remaining ones are used as building blocks for creating *Grid*, *Square*, and *Node* image features. In case of encountering different numbers of building blocks, the *Grid* feature detector invalidates randomly selected ones, in order to ensure equally sized *Grid* image features. Gabor jets are extracted from the object view at each pixel position given by the feature building blocks (see Figure 2). Finally, the image features are created by concatenating the Gabor jets $\mathcal{J}(x_p)$, $0 \leq p < P$ extracted from their respective pixel positions x_p in the object view, with P indicating the number of Gabor jets in the image feature. A generic image feature is given by

$$\mathcal{F}(\mathbf{x}) = \left(a_{p,m,l} \cdot \exp(i \cdot \phi_{p,m,l})\right)_{p,m,l} , \qquad (1)$$

where each complex filter response is expressed in terms of amplitudes $a_{p,m,l}$ and phases $\phi_{p,m,l}$, with $0 \leq m < M$ denoting the spatial frequency and $0 \leq l < L$ the orientation of the Gabor filter, and $\mathbf{x}$ represents the concatenated pixel positions x_p. For simplicity, we drop $\mathbf{x}$ and use $\mathcal{F}$ to denote the image features.

These local feature detectors can extract several *Square* and *Node* image features from each object view, but only one *Grid* image feature. Overall, this feature extraction process generates 3280 *Grid*, 154166 *Square*, and 265937 *Node* image features from the ETH-80 image set.



(a) (b) (c)

Figure 2. The image features: *Grid* features (a) cover the whole object view, *Square* features (b) consist of 5 Gabor jets, and single jets constitute *Node* features (c).

3 Visual Object Memory Model

In order to represent visual object memory we use a *Neural Map*, a biologically inspired memory organization framework that provides a structural association to image features derived from training object views. It comprises a GNG network [3] and a classifier motivated in the cortical neurons population coding and decoding processes [10]. These components equip the data structure of the Neural Map with self-organized feature structuring and intelligent feature matching.

3.1 Self-Organized Feature Structuring

A set of image features F_T extracted from training object views is automatically integrated into the Neural Map's associative structure of model features through an unsupervised two-stage learning procedure.

In the first stage, the amplitude values of the image features $\mathcal{F} \in F_T$ are employed to generate sample signals from a high-dimensional data distribution. The GNG [3] uses these sample signals to incrementally develop a neural network. During this process, each neuron is associated with a point in the feature distribution. Neurons are connected with synapses, which are locally adapted according to the feature similarity:

$$\mathcal{S}\left(\mathcal{F}, \tilde{\mathcal{F}}\right) = \frac{\sum_{p,m,l} a_{p,m,l} \cdot \tilde{a}_{p,m,l}}{\sqrt{\sum_{p,m,l} a^2_{p,m,l}} \cdot \sqrt{\sum_{p,m,l} \tilde{a}^2_{p,m,l}}} \,. \quad (2)$$

This measure allows for smooth similarity potentials with fairly wide maxima [14].

In comparison with previous and similar approaches (e.g., Kohonen Feature Map, and Growing Cell Structure), GNG is more flexible since no dimensionality assumptions need to be made, and it allows continuous learning by adding neurons and synapses until a performance criterion is met. The algorithm's resulting network has a topological structure composed of a set of N neurons connected by synapses closely reflecting the topology of the feature distribution.

During the second stage, the developed neural network is utilized to establish a map $\mathcal{C} : F_T \rightarrow \mathbb{R}^N$ by calculating a distance measure for each feature's amplitude values with all its neurons:

$$\mathcal{C}(\mathcal{F}) = (c_i)_i = (1 - \mathcal{S}(\mathcal{F}, \mathcal{F}_i))_i \,, \quad (3)$$

where $\mathcal{F}_i$ represents the i^{th} neuron's point in the feature distribution with $c_i \in \mathbb{R}$, $0 \leq i < N$.

The vector of values resulting from this map constitutes a feature code and represents the neuronal population code for each of the image features extracted from the training object views. Population coding has many advantages in information processing (e.g., its resulting codes are robust against a potential loss of neurons). It is also compatible with neurophysiological findings about distributed representations the brain's object recognition system.

3.2 Intelligent Feature Matching

Finding the best matching model features for a set of novel features F_R extracted from test object views is at the core of the object recognition system [14] and the categorization system [13]. In our work, they are matched through a recall procedure, where the previously learned information inferred from the self-organized Neural Map favors the selection of some model features over others until the best matching one is found.

This recall procedure is accomplished by a classifier, which decodes the responses of the neurons from the self-organized Neural Map for each novel feature $\mathcal{F}' \in F_R$ as follows. Initially, it calculates the novel feature code according to Equation 3. Subsequently, it obtains the set of candidate model features F_C closest to the novel feature:

$$F_C = \left\{\mathcal{F} \in F_T : |\mathcal{C}(\mathcal{F}) - \mathcal{C}(\mathcal{F}')| \leq \mu + 10^{-5}\right\} \quad (4)$$

$$\mu = \min_{\bar{\mathcal{F}} \in F_T} \left(\left|\mathcal{C}\left(\bar{\mathcal{F}}\right) - \mathcal{C}(\mathcal{F}')\right|\right) \quad (5)$$

Then, it chooses the model feature from the set F_C with the highest similarity to the novel feature as the best matching one model feature $\mathcal{F}''$:

$$\mathcal{F}'' = \underset{\mathcal{F} \in F_C}{\operatorname{argmax}} \left(\mathcal{S}(\mathcal{F}, \mathcal{F}')\right) \,. \quad (6)$$

Finally, the classifier produces a set of matched image features F_M, which represent the decoding process of the neuronal responses for the novel features set F_R.

4 Object Recognition and Categorization

We now assess the performance of the procedures from Section 3 for representing the visual object knowledge derived from the image feature types described in Section 2.

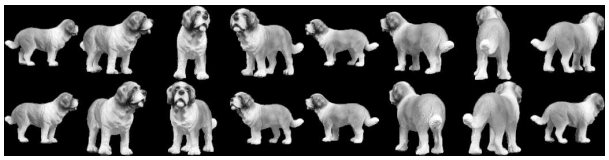
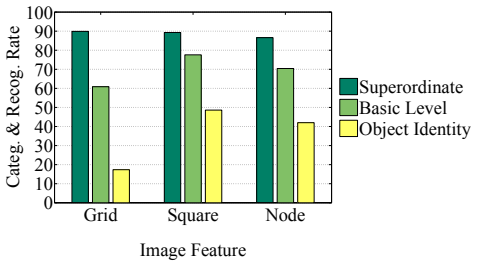


Figure 3. Partitioning for the view point invariance test: the first row depicts a subset of 8 views used for learning; the second row shows the corresponding testing subset.



	Grid	*Square*	*Node*
Animals	94.11%	76.3%	76.37%
Fruits and vegetables	89.22%	100%	99.83%
Human made big	90.33%	89.56%	75.5%
Human made small	78.83%	96%	88.63%
Cows	44.67%	54%	43%
Dogs	55%	46.67%	36%
Horses	45%	41%	28.38%
Apples	32%	98.5%	97.5%
Pears	83.67%	88.56%	82.38%
Tomatoes	57.5%	99.5%	99%
Cars	90.33%	94.94%	82.75%
Cups	78.83%	97.56%	94.38%

Figure 4. Summarized (top) and detailed (bottom) categorization and recognition rates for the levels of the object taxonomy using *Grid*, *Square*, and *Node* image features, respectively.

During these experiments the training parameters of the GNG algorithm are empirically defined based on the ones proposed in [1] together with the ones resulting from GNG learning tests. All results are obtained by averaging several trials with the same experimental set-up under different starting conditions.

4.1 View Point Invariance Test

The first experiment tests the view-point invariance of the Neural Map. All possible objects are available during learning and recall procedures, their respective training and test sets are produced from an intercalated view-point based partition of all object views. Accordingly, test object views from 21 view-points are selected by rotating horizontally the training object views from 20 view-points by 22.5° plus the ones from one extra view-point; Figure 3 shows an example of a subset of this partitioning.

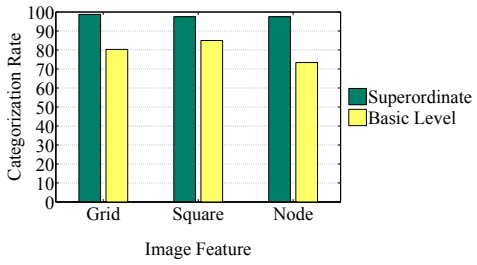
The image features from the training object views are used to shape a self-organized Neural Map as described in Section 3.1. After this learning stage, the image features extracted from a test object view are matched against Neural Map's model features as described in Section 3.2. The matching responses are recorded in the three different levels of the object taxonomy detailed in Section 2.1. Then they are subjected to a winner-take-all voting scheme, which determines from coarse to fine the matched object labels based on the majority of labels found in the object taxonomy. Initially, the superordinate label is established, then the basic level one is selected, and finally the object identity is recognized, depending on the taxonomy sub-tree. When the vote counts of the winner and the second best candidate differ by less than 30%, further voting is calculated over the registered responses in both taxonomy sub-trees. This diminishes the propagation of early made incorrect decisions. Categorization is considered successful in the superordinate and basic levels when their respective matched object labels coincide with the ones from the test object view. Recognition is successful when the identity labels of test and matched object are equal. Categorization and recognition are repeated for all the test object views. The experimental results for the three different types of image features are shown in Figure 4.

4.2 Leave-one-object-out Cross-validation

In this experiment a Neural Map is combined with a voting scheme to form a novel feature-based categorization system. The overall performance of this system is measured using the leave-one-object-out cross validation. The Neural Map is trained using the image features from all object views of 79 objects, and the ones derived from the views of the remaining object are used for testing. Categorization is considered successful under the same conditions as before. This is repeated for all 80 partitions of the database, and the rates are averaged over all tests. The results are shown in Figure 5. Recognition can not be be performed with this partitioning.

5 Discussion and Further Research

We have introduced a memory framework for object recognition and categorization systems based on dynam-



	Grid	*Square*	*Node*
Animals	100%	96.67%	100%
Fruits and vegetables	99.58%	100%	100%
Human made big	100%	100%	100%
Human made small	91.25%	90%	80%
Cows	60%	80%	60%
Dogs	65%	40%	37.14%
Horses	65%	60%	20%
Apples	75%	100%	100%
Pears	98.75%	100%	90%
Tomatoes	85%	100%	100%
Cars	100%	100%	100%
Cups	93.75%	100%	80%

Figure 5. Summarized (top) and detailed (bottom) categorization rates of novel objects using *Grid*, *Square*, and *Node* image features, respectively.

ically assembled object models [14]. Its properties include the unsupervised structural organization of object components according to their visual resemblance, and the use of this structure for matching novel components. We have tested its performance using image features of different granularity. The results indicate that the medium-sized *Square* image features yield the highest recognition and categorization rates, and slightly outperform the results from [6]. This suggests that these image features maximize the informativeness and distinctiveness derived from the object views.

We also observe a gradual decrease of accuracy from the abstract to the concrete levels of the object taxonomy, reinforcing the expectation that the categories from higher abstraction levels are more easily distinguishable. The mechanism to minimize propagation of early made errors in the coarse to fine voting process proves to be successful.

Generally, the lowest basic level categorization rates are found within the *animals* taxonomy sub-tree, but the false positives are concentrated in the same sub-tree, in agreement with [6]. This reveals difficulties to distinguish similar objects with complex global shapes. We are currently trying to confirm this hypothesis on larger databases.

The categorization experiments presented here are obtained through a best case analysis, because novel object views are processed under the same viewing conditions as during training, with near-perfect object segmentation and known scales. For recognition and categorization in more realistic scenes we will have to use Neural Maps to dynamically create object models, which can be subjected to correspondence-based recognition methods.

Acknowledgments We gratefully acknowledge funding from the EU in the *NovoBrain* project (MEST-CT-2005-020385), from the DFG (WU 314/5-2), and the Ruhr-University Research School funded by Germany's Excellence Initiative (DFG GSC 98/1).

References

[1] B. Bolder. *Sensomotorische Koordination eines Roboterkopfes*. Shaker, Aachen, 2006.

[2] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *European Conference on Computer Vision*, pages 747–760. Springer, 2004.

[3] B. Fritzke. A growing neural gas network learns topologies. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, *Advances in NIPS*, pages 625–632. MIT Press, 1995.

[4] G. Humphreys, M. Riddoch, and C. Price. Top-down processes in object identification: Evidence from experimental psychology, neurophysiology and functional anatomy. *Phil. Trans. Roy. Soc. B*, 352(1358):1275–1282, 1997.

[5] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[6] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *Computer Vision and Pattern Recognition*, pages II–409–415, 2003.

[7] B. W. Mel. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9:777–804, 1997.

[8] T. J. Palmeri and I. Gauthier. Visual object understanding. *Nature Reviews Neuroscience*, 5(4):291–303, April 2004.

[9] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, July 1976.

[10] T. Trappenberg. *Fundamentals of Computational Neuroscience*. Oxford University Press, June 2002.

[11] S. Ullman. *High-Level Vision*. MIT Press, 1996.

[12] J. Ward. *The Student's Guide to Cognitive Neuroscience.* Psychology Press, 2006.

[13] G. Westphal. *Feature-Driven Emergence of Model Graphs for Object Recognition and Categorization*. PhD thesis, University of Lübeck, Germany, 2006.

[14] G. Westphal and R. P. Würtz. Combining feature- and correspondence-based methods for visual object recognition. *Neural Computation*, 21(7):1952–1989, 2009.

[15] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.