Recognizing Rotated and Occluded Shapes Using Binary-Activated Neural Networks and Sparse Feature Representations

Joseph Murray and Kenneth Kreutz-Delgado

Electrical and Computer Engineering, Jacobs School of Engineering
University of California, San Diego
La Jolla, California 92093-0407, USA
jfmurray@ucsd.edu, kreutz@ece.ucsd.edu

Abstract

We present a new object recognition system based on sparse feature extraction and a binary-activated neural network, which is a biologically motivated associative memory. The network [3, 5] uses multiple layers connected with feedforward, feedback and lateral weights. To achieve high capacities, the network relies on highly sparse activation patterns at the input (and other) layers. After activations are propagated through the weight matrix, sparse patterns are generated on higher layers using a k-winners-take-all competition. The number of winners k is chosen by considering capacity constraints of associative memories [5]. For our network with regions of size 9248 and 3000, 25 and 20 winners respectively are allowed. Sparse features are generated from input images using a hierarchical series of filter banks.

The network is trained using a binary Hebbian learning rule, with row and column normalization of the weight matrices to prevent weights from growing without bound. Transformation invariance is achieved by extracting temporal redundancies in the training sequences. During training, an object is presented in a consecutive sequence of transformations, and the algorithm is given the prior knowledge that the underlying object remains present. This allows one representation at a higher layer to be associated with many transformed versions of an input object. Temporal sequence training has been used in other work to learn invariance to other transformations [2, 6, 1].

Three types of experiments are performed: recognizing rotated images, recognizing occluded images and detecting images in cluttered scenes. For images rotated within an $85°$ range, recognition accuracy is 93.3% with 15 pattern classes, and 89.3% with 30 classes (see Table 1). Adding small translations to the images improves generalization performance on the test data.

To test robustness to occlusion, images are presented with 10 to 70% of their pixels removed. Performance is found to degrade smoothly with increasing occlusion, with respectable recognition rates with up to 50% of the image occluded.

The third series of experiments shows the utility of feedback connections in vision systems. In scenes with many objects, the system is given the task of decided whether a target object is present. Information about which object to search for is given to the highest layer of the network, and feedback connections propagate this to lower layers. This process is called *expectation driven detection*. It is important to note that the expectation given to the highest layer is independent of the orientation of the presented objects. Experiments were performed on test images which included from one to three overlapping distractor patterns. This is a difficult recognition problem (as seen in Figure 2) which does not appear often in the literature, yet humans are readily able to solve it. Using one distractor pattern overlapping the target, the detection rate was 79.7% with 4.6% false alarm rate; with two distractors, 80% detection with 21.7% false alarm; and with three distractors, 85.4% detection with 45.1% false alarm. (The false alarm rate can be lowered with corresponding degradation of detection performance.)

While the images used here are fairly simple two-dimensional shapes, we believe that the experiments are of sufficient difficulty to demonstrate the key properties of the system: sparse feature encoding, learned transformation invariance, graceful degradation of performance under occlusion, and expectation driven detection. With the use of more sophisticated sparse feature extraction [4], such a system could be applied to real-world images with multiple transformations. Many state-of-the-art vision systems are sensitive to occlusion and cluttered scenes (such as used in the detection experiments (Figure 2)); the presented system shows (schematically) how these problems could be overcome.

References

[1] Marian Stewart Bartlett and Terrence J. Sejnowski. Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network: Comput. Neural Sys.*, 9:399–417, 1998.

[2] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

[3] Robert Hecht-Nielsen. A theory of the cerebral cortex. In *Proceedings of the 1998 International Conference on Neural Information Processing (ICONIP'98)*, pages 1459–1464, October 1998. Kitakyushu, Japan.

[4] Kenneth Kreutz-Delgado, Joseph F. Murray, Bhaskar D. Rao, Kjersti Engan, Te-Won Lee, and Terrence J. Sejnowski. Dictionary learning algorithms for sparse representation. *preprint, submitted to Neural Computation*, 0:1–38, 2002.

[5] Brian Sagi, Syrus C. Nemat-Nasser, Rex Kerr, Raja Hayek, Christopher Downing, and Robert Hecht-Nielsen. A biologically motivated solution to the cocktail party problem. *Neural Computation*, 13(1575-1602), 2001.

[6] Guy Wallis and R. Baddeley. Optimal unsupervised learning in invariant object recognition. *Neural Computation*, 9(4):959–970, 1997.

Table 1: Results of recognition experiments on rotated images. Training sets with 15 and 30 images were used. Translation refers to +/- 1 pixel translation during training and testing.

			Training			Testing		
Patterns	Translate	Epochs	Images	Correct	Accuracy	Images	Correct	Accuracy
15	No	35	195	195	100.0%	1080	973	90.1%
15	Yes	100	195	192	98.5%	1080	1008	93.3%
30	No	35	390	390	100.0%	2160	1845	85.4%
30	Yes	100	390	375	96.2%	2160	1929	89.3%

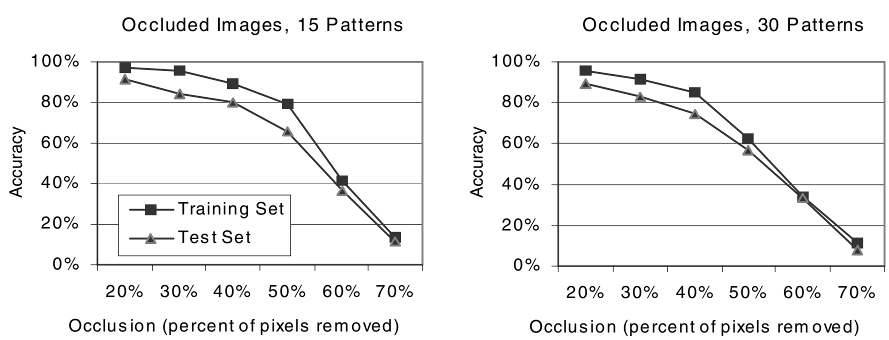

Figure 1: Recognition accuracy when presented with occluded images

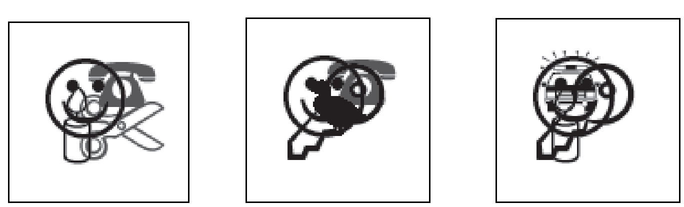

Figure 2: Examples of the images used for the detection experiments summarized in the abstract. Three distractor patterns are superimposed over one target pattern.