Human Vision Based Object Recognition

Sye-Min Christina Chan

Abstract

Serre, Wolf, and Poggio introduced an object recognition algorithm that simulates image processing in visual cortex and claimed to get better results than traditional recognition methods. A close examination of their algorithm shows that it is based largely on manually picked parameters. In this project, machine learning methods are applied to modify the algorithm.

Introduction

The way human vison system works has always been an interest to neuroscientists. Recently, computer scientists have become interested in modelling the human vison systems as well. One of the motivations is to improve the understanding of, and possibly to predict, how the brain functions . Olshausen and Field [1] demonstrated that receptive fields of simple cells can be represented as basis functions that are similar to gabor filters, and these basis functions can be learnt from images by sparse coding, a method that maximizes information preservation and sparsity of response. This result is encouraging because it agrees with observations that S1 cells act like gabor filters.

Another motivation is that humans still outperform the best machine vision systems on tasks such as object recognition. It therefore makes sense to model computer vision systems after the human vision system. Serre, Wolf, and Poggio [2] introduced a set of features that resemble the feedforward models of object recognition in V1 and V2 of the visual cortex. The system deveoped using these features gives a higher recognition rate compared to traditional template-based and histogram-based systems. However, the algorithm relies heavily on parameters manually picked to fit biological cells. In addition, S2 and C2 features are computed based on randomly selected C1 prototypes. This method is essentially template matching and does not seem to have a biological justification. The objective of this project is to apply machine learning to the algorithm for feature generation, and to develop a system that simulates the human vision system.

V1 Layer

In [2], S1 receptive fields are modeled as gabor filters. Parameters of these filters are picked so that they agree with biological findings, and filters are grouped into eight bands according to these parameters. S1 cell responses are then obtained by applying these filters to images and C1 cell reponses are resulted from max pooling over scales and positions.

Although it has been proven that methods such as Independent Component Analysis [3] and Sparse Coding [1] can learn basis functions that resemble the gabor-like S1 receptive fields, in order to generate C1 cell responses, which are shift and scale invariant, groupings of S1 cells are necessary. Independent Subspace Analysis (ISA) [4] is an extension of ICA that allows cells belonging to the same subspace to be dependent. More specifically, cells within the same subspace have similar orientations and frequencies, but may have slight variations in locations and very different phases. Therefore, replacing the the bands described in [2] by these subspaces and pooling the responses of cells within same subspaces give phase and shift invariance, which are features of complex cells. In addition, to get scale invariance, basis of different sizes can be grouped together according to the similarity of their frequencies and spatial locations, which can be estimated by fitting them to gabor filters. Figure 1a below shows the the bands of ISA filters formed using 8 by 8 and 12 by 12 basis, and figure 1b shows the the basis fitted to gabor filters.

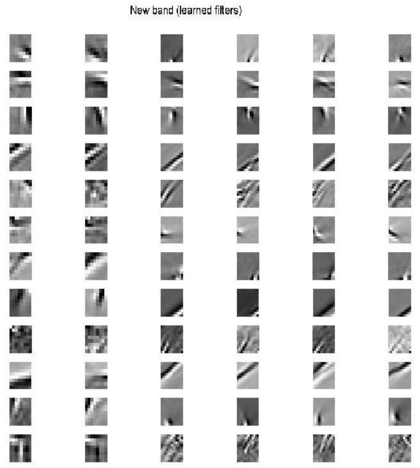

Figure 1a--bands of ISA filters formed by 8x8 filters with subspace size of 2, and 12x12 filters of subspace size 4

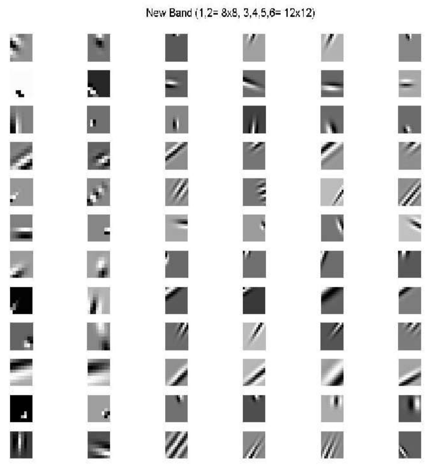

Figure 1b--this shows the same bases on the left fitted to gabor filters

Topographical Independent Component Analysis (TICA) [5] is another extension of ICA that allows some degree of dependence among components. In this implementation, components are allowed to be dependent with their neighbors. Similar to ISA, these neighbhorhoods can be used to replace the bands in [2] to obtain C1 responses.

V2 Layer

In [2], S2 response of an image is the distance between its C1 response and the prototype C1 responses, which are chosen randomly at the beginning of the algorithm. A more biologically related approach is to apply learning algorithms on C1 responses to get S2 responses. Figure 3 shows the V2 filters obtained from applying sparse coding on ISA-generated C1 responses. This is a modification of the contour net algorithm in [6], replacing the first layer with ISA responses.

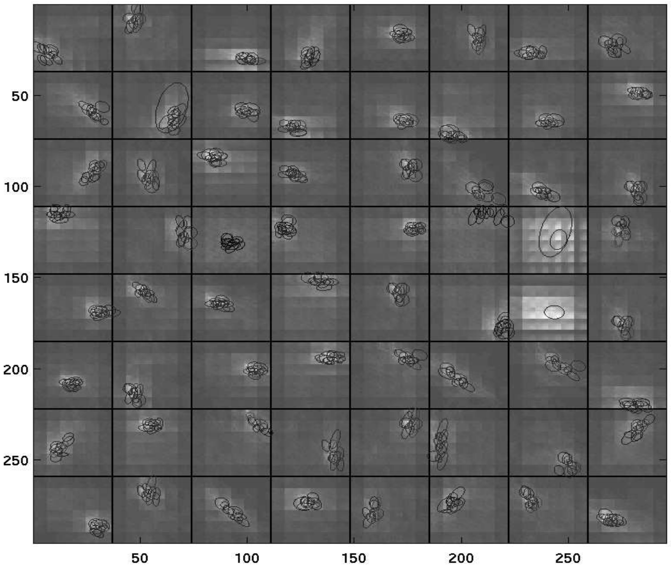

Figure3 --V2 filters

Experiments

Two types of experiments were carried out to test the performance of the learned filters. In the first experiment, six classes were chosen at random from Caltech 101 with a training sample size of 1, 3, 6, 15, 30, and 40 respectively, as described in [2]. Each category was tested against the background class in Caltech 101, which consists of random images downloaded from Google. In the second experiment, each class was tested against samples randomly chosen from the other five classes and the test was repeated for ten times to get the average recognition rate.

Table 1 below shows the two category classification results of replacing S1 filters by 8x8 ISA filters, ISA filters with scale invariance (by grouping filters of two different sizes into the same band), fitted ISA filters with scale invariance, and 8x8 TICA filters respectively. The last column shows the results using the original gabor filters of [2] for comparison. It can be seen that the fitted ISA filters with scale invariance give the closest results with those obtained using the original gabor filters, although other filters give results that are comparable as well.

Class ***(number of samples for training)***	***ISA (8x8)***	***ISA (8x8 + 12x12)***	***Fitted ISA filters (8x8+ 12x12)***	***TICA (8x8)***	***Gabor filters***
Crocodile_head (1)	0.5	0.5	0.5	0.5	0.5
Cellphone (3)	0.53	0.54	0.56	0.52	0.51
Elephant (6)	0.51	0.5	0.51	0.53	0.54
Bonsai (15)	0.59	0.58	0.64	0.57	0.65

Class *(number of samples for training)*	*ISA (8x8)*	*ISA (8x8 + 12x12)*	*Fitted ISA filters (8x8+ 12x12)*	*TICA (8x8)*	*Gabor filters*
Kangeroo (30)	0.84	0.81	0.83	0.78	0.74
Airplanes (40)	0.93	0.91	0.95	0.94	0.97

Table 1—two-category classification with V1 replaced

Table 2 below shows the results of the second experiment for ISA filters with scale invariance (learned), fitted ISA filters with scale invariance, and the original gabor filters. From the table, except for the elephant and the bonsai classes, which show unexpectedly poor results for all three filters, recognition rates obtained using the fitted ISA filters are comparable with that of the original gabor filters.

Class (no. training samples)	*Crocodile (11)*	*Cellphone (11)*	*Elephant (12)*	*Bonsai (13)*	*Kangeroo (16)*	*Airplanes (18)*
Learned	0.7537	0.7081	0.5778	0.5849	0.6908	0.8959
Fitted	0.7122	0.8767	0.6000	0.5368	0.7154	0.9521
Gabor	0.6600	0.8600	0.7100	0.6600	0.6800	0.98

Table 2—one versus five categories classification with V1 replaced

Table 3 shows results of the second experiment when both layers are replaced by learned filters—V1 is replaced by ISA filters and V2 is replaced by sparse coding filters. Since sparse coding is not a linear algorithm, S2 response cannot be computed by simply convolving the filter with C1 responses. C1 responses have to be divided into smaller patches in order to run gradient descent to get their S2 responses. When dividing C1 responses into smaller patches, there could be overlapping areas between these patches. Table 3 shows the recognition rate of one versus five categories classification as the overlap area varies. As shown, the results do not seem to be improved as one would expect when increasing the proportion of overlap area.

Class (no. training samples)	*Crocodile (11)*	*Cellphone (11)*	*Elephant (12)*	*Bonsai (13)*	*Kangeroo (16)*	*Airplanes (18)*
½ overlap	0.5630	0.8279	0.5811	0.6632	0.6223	0.5610
1/3 overlap	0.6293	0.8105	0.5889	0.6292	0.6823	0.5925
1/6 overlap	0.5512	0.7291	0.5900	0.7028	0.6192	0.6541

Table 3—one versus five categories classification with V2 replaced

Conclusion

Comparable recognition rates are obtained by replacing the gabor filters in [2] by filters learned from training on natural images. In particular, fitted ISA filters with scale invariance demostrate the best results. While TICA filters give similar rates as ISA filters, it takes longer to compute because of the large number of bands and therefore are not used in other experiments.

To actually improve the recognition rates and the efficiency of the algorithm, it is expected that machine learning techniques have to be applied to the second layer as well. Preliminary results from applying sparse coding on C1 reponses are not satisfactory. Possible changes that can be made include applying spatial pooling on the sparse coding responses to increase shift invariance and to reduce the size of the feature vector fed into the support vector machine. Another possiblility is to apply ISA or the independent subspace version of sparse coding to C1 reponses to give S2 responses with groupings for max pooling.

Acknowledgement

This project is done in collaboration with Alexis Battle, Honglak Lee, and Ben Packer, under the supervision of Professor Andrew Ng. I would like to thank them all for their advice, support, and help throughout the quarter.

Reference

1. Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature,* 381, 607-609.
2. Serre, T., Wolf, L., and Poggio, T., Object Recognition with Features Inspired by Visual Cortex. *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), IEEE Computer Society Press,* San Diego, June 2005.
3. Bell, A. J. and Sejnowski, T. J. (1997) The 'independent components' of natural scenes are edge filters. *Vision Research,* 37(23) 3327-3338.
4. Hyvarinen, A., and Hoyer, P.O. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation,* 12(7), 1705-1720.
5. Hyvarinen, A., Hoyer, P. O., and Inki, M. (2001) Topographic Independent Component Analysis. *Neural Computation, 13.*
6. Hyvarinen, A., and Hoyer, P. O. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research,* 42(12) 1593-1605.