Object Recognition with Task Relevant Combined Local Features

Wenjun Zhu and Liqing Zhang

Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai 200240, China
{wilsonzhu,lqzhang}@sjtu.edu.cn

Abstract. A number of cortex-like hierarchical models of object recognition have been proposed these years. In this paper, we improve them by introducing supervision during forming combined local features. The traditional cortex-like hierarchical models always contain three layers which imitate the functions of neurons in ventral visual stream of primates. The bottom layer detects orientation information in a local area. Then the middle layer combines these information to form combined features. Finally, the top layer integrates combined features to form global features which are input into a classifier. In these models, three stages to form global features are all unsupervised. The supervision procedure only occurs after global features are generated, which is implemented by the classifier. But we think the supervision should occurs earlier. For a particular object recognition task, the second stage of generating global features is also supervised because only task relevant combinations are useful. In our paper, we analyze why introducing supervision in this stage is necessary. And we explain task relevant combined local features can be extracted by some feature selection algorithms. We also apply this improved system to a series of object classification problems and compare it with traditional models. The simulation results show that our improvement really boosts object recognition performance.

Keywords: object recognition, task relevant, visual cortex, Gabor filter.

1 Introduction

Human brain can recognize what it concerns in clutter within a very short time. This ability surpasses all artificial systems. So understanding how visual cortex recognizes objects and building brain-like recognition system attract many researchers in the fields of physiology, neuroscience and computer science.

Over the last decade, a number of basic properties about object recognition in cortex have been found through many physiological experiments [1]. Visual signals received by retina are processed stage by stage in the ventral visual pathway, which runs from primary visual cortex (V1), over extrastriate visual areas V2 and V4 to inferotemporal cortex (IT). The following properties have been widely accepted in this stearm. Firstly, simple cells in V1 respond perferably to oriented bars [2]. Secondly, neurons along the ventral stream show an increase in receptive field size as well as in the complexity of their preferred stimuli. V4

D.-S. Huang et al. (Eds.): ICIC 2008, LNCS 5226, pp. 285–292, 2008.
© Springer-Verlag Berlin Heidelberg 2008

provides information about the individual contour elements and these elements are integrated into shapes at the level of area IT [2,3,4,5]. Finally , at the top of the ventral stream, in anterior inferotemporal cortex (AIT), cells are tuned to very complex stimuli such as faces, animals, scenes, etc [6].

Based on these widely accepted properties, several transformation invariant object recognition models have been proposed [7,8,9]. HMAX [10,11] proposed by MIT-CBCL is one of the most significant models among them. This hierarchical model based on feedforward connections ignores any dynamics of the back-projection, although feedback connections really exist in our visual cortex. In the simplest form of the model, it contains fours layers, which are S1,C1,S2 and C2 from bottom to top. S1 detects orientation in a particular position. C1 takes the max over a local area to form local invariant features. S2 combines features output by C1. This procedure is implemented by extracting some patches or templates from positive training images and calculating similarity between the samples and the templates. C2 deals with global invariance by getting the maximal value over the outputs of S2. Finally, the global invariant features formed in C2 are input into a classifier (SVM, neuron network, etc.). The whole recognition procedure is composed of two stages. The stage forming C2 features is unsupervised and the stage using classifier to recognize objects is supervised. However, we find for a particular object recognition task not all C2 features extracted by this unsupervised method are useful. Some of them are task irrelevant. These irrelevant features may reduce the system's performance. So we introduce supervision to form C2 features in our improved model. Task relevant combinations are extracted in the layer S2, because we think only partial combined local features are useful for a particular recognition task. For example, a feature which can discriminate two different people in face recognition task doesn't work in the task to detect whether the object is a face, and vice versa. This improvement boosts the recognition performance, which we will show in Section 3. Another dedication of this improvement is that it avoids the problem in combining features, which is the number of possible combinations is too large. The authors of HMAX resolve this problem by randomly selecting a few ones from all combinations. In our model, we select task relevant combinations. Obviously it is better than random selection.

This paper is organized as follows. In Section 2, we briefly introduce HMAX at first. Then present our improvement and describe detailed implementation of the whole system. In Section 3, we apply our system to both binary and multi-class classification problem on public database and compare the results with HMAX. Finally, we summarize our work and propose some open questions about the limitations and possible improvements of the model in Section 4.

2 Hierarchical Model and Feature Selection

HMAX is one of the most significant hierarchical models. It was first proposed by Riesenhuber and Poggio in [10]. After that, Serre and his co-workers evolved the model and applied it to a number of challenging recognition tasks [11].

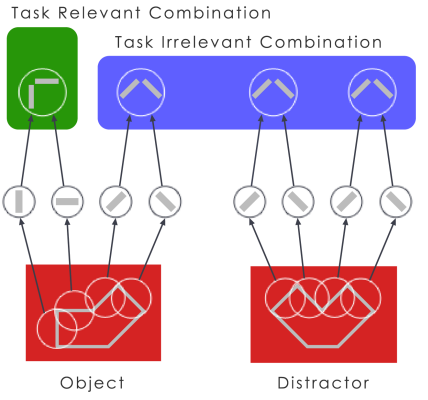


Fig. 1. An example of task relevant and irrelevant combination

HMAX model consists of simple S units and complex C units alternately. In the simplest form, there are four layers (S1, C1, S2, C2). The readers can refer [10,11] for the detailed implementation of these layers.

The difference between our model and HMAX is we employ supervision in the stage of forming combined features. In fact, learning occurs at all stages in ventral visual stream. It is unsupervised in the lower layers and supervised in the higher layers. The properties of simple and complex cells in V1 is learned from natural images by unsupervised proedure [12,13]. In the next stage forming local combined features, the authors of HMAX think it is also unsupervised. However we don't agree this opinion. Maybe this stage seems unsupervised for all kinds of recognition tasks. But for a special recognition task, we believe only a few task relevant combinations are useful and the learning is supervised. In other words, we think there are many cells in cortex corresponding all possible combinations, but only task relevant cells are activated for a particular recognition task and the others can be ignored. We illustrate a example in Fig. 1. The combination in blue rectangle is useless for the object recognition task, and the one in green rectangle is task relevant. In our model, we only keep these task relevant combinations and the others are discarded.

How to extract task relevant features is a feature selection problem. Now, there are many methods to resolve feature selection problem [14,15,16,17,18]. In this version of our model, we use a greedy algorithm to resolve it. The algorithm can be described as:

1. Set $X \leftarrow$ " initial set of n features", $S \leftarrow$ " empty set".
2. $\forall x \in X$, compute $J(\{x\} \cup S)$[1].
3. Find the feature that maximizes $J(\{x\} \cup S)$, set $X \leftarrow F \backslash \{x\}$ and $S \leftarrow \{x\} \cup S$.
4. Repeat until desired number of features are selected.

[1] $J(\bullet)$ is a criterion to evaluate the selected subset of features.

In our experiments, the inter-intra criterion is used, i.e. $J = trace(S_w^{-1} S_b)$, where S_b is between-scatter and S_w is within-scatter.

At the end of this subsection, we summarize our method completely[2]:

1. For every image, we apply a battery of Gabor filters to it. The filters are in 4 orientations θ and 16 scales s. Then we obtain $16 \times 4 = 64$ maps $(S1)_\theta^s$ arranged in 8 bands (e.g., band 1 contains filter outputs of size 7 and 9, in all four orientations, band 2 contains filters outputs of size 11 and 13, etc).
2. Take the max over scales and positions within each band: each band member is sub-sampled by taking the max over a grid with cells of size N^Σ first and the max between the two scale members second, e.g., for band 1, a spatial max is taken over an 8×8 grid first and then across the two scales (size 7 and 9). Note that we don't take a max over different orientations. Hence, each band $(C1)^\Sigma$ contains 4 maps.
3. Extract all possible patches of size $n_i \times n_i \times 4$ from all $(C1)^\Sigma$ of all positive training images. They are $P_{i=1,2,\cdots,K}$.
4. For each C1 image $(C1)^\Sigma$ and each P_i, compute: $Y = \exp(-\gamma \parallel X - P_i \parallel)^2$. X is a patch of $(C1)^\Sigma$, whose size is same as P_i. And we let X run over all positions. Obtain S2 map $(S2)_i^{\Sigma 3}$.
5. Take the max over all positions and bands, we obtain $(C2)_i$. Now, every input image are converted into K C2 features. Every C2 feature correspond a patch P_i in step 3.
6. In the training set, we use our greedy feature selection algorithm to extract k task relevant C2 features ($k << K$). The patches corresponding these features are task relevant combinations.
7. Input these selected C2 features into a classifier (linear SVM in our implementaion).

3 Experimental Results

We evaluate our system by several object detection tasks and a multi-class object classification task in this section. The classifier used in our system is a linear SVM.

The database used in our experiments is Caltech6[4]. This is a public multi-class object database. The Caltech6 database contains six object categories and a background category (used as the negative set). In fact, there are only 5 different kinds of objects in Caltech6 database, because the first two categories are both cars (rear). For cars (rear), we only use the images under the directory *cars_brad* in our experiments.

[2] In this version, the implementation of S1 and C1 are same as HMAX. So you can refer [11] for more information.

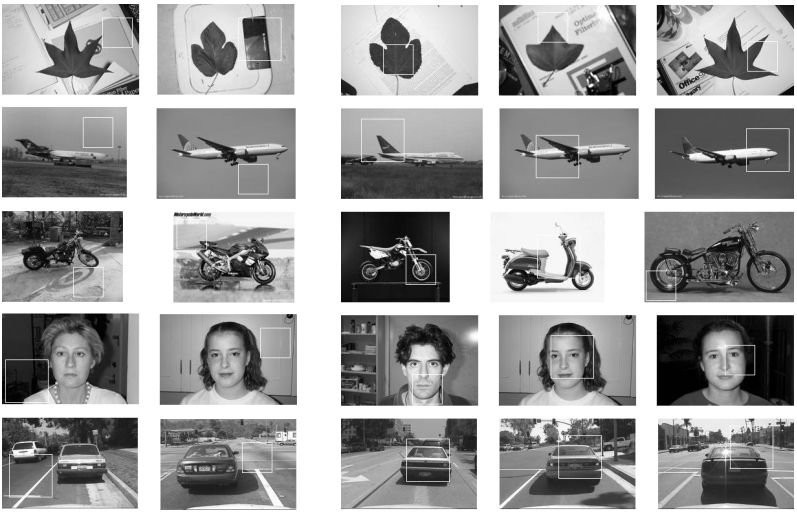
[3] For a train image, we should calculate all S2 features. However we need not compute all S2 features for a test image, because only task relevant patches are useful for them.

[4] http://www.vision.caltech.edu/Image_Datasets/Caltech6/

Fig. 2 illustrates some examples in Caltech6. These images contain the target object embedded in a large amount of clutter. The challenge is to learn from unsegmented images and discover the target object class automatically. As a preprocessing procedure, we normalize all images to 140 pixels in height (width is rescaled accordingly so that the image aspect ratio was preserved) and converted to gray values.

3.1 Object Detection

In our object detection experiments, we selected 40 images in a category as the positive train samples and 50 background images as the negative ones. We also extracted 100 positive and 100 negative test samples from the remaining images. The performance of HMAX was averaged over 10 runs. The code of HMAX in our experiments was downloaded from MIT-CBCL (http://cbcl.mit.edu/software-datasets/). Only 6×6 and 10×10 patches of band 2 were extracted in our experiments for saving computation time (Using patches of more sizes and bands will improve the performance.). In HMAX, we randomly extract 100 patches in each size and totally 200 C2 features. In our system, we calculated correlation coefficient between every C2 features and train label. Then run our greedy feature selection algorithm on the most correlated C2 features (1000 features) to extract task relevant patches (50 patches). Some features extracted by HMAX are task



(a) task irrelevant patches (b) task relevant patches

Fig. 2. (a) and (b) are some examples of task irrelevant and relevant patches. These patches are extracted from C1 outputs. We use white rectangles to denote the areas corresponding these patches in the original images. Task irrelevant patches in (a) come from background or other objects and task relevant patches in (b) come from the target objects.

Table 1. Performance of HMAX and our model on five object detection tasks

Categories	HMAX (200)	Our Model (50)
Leaves	94.9	97.5
Airplanes	91.6	94
Motorbikes	95.8	98
Faces	98.2	99
Cars	97.9	99.5

irrelevant and useless for classification. But in our system almost all features are task relevant. So our system need fewer features than HMAX for the same classification problem. You can see the performance of our system with only 50 features even better than HMAX with 200 features in Table 1.

Fig. 2(a) illustrates some examples of task irrelevant patches extracted by HMAX. We can see these patches come from background or other objects rather than the target object. So the features formed from these patches are useless for recognize the target object. Fig. 2(b) illustrates some patches extracted by our system. We can see these areas are the most important parts of the target objects and features formed from these areas are very discriminative between the objects and distractors. Although size and position of target objects vary in different images, our system can detect where they are and extract features from these places.

Table 1 summarize the experimental results. We use 200 features in HMAX and 50 features in our model. The classification rates of our model are better than HMAX in these challenging applications even with lesser features. And the results show that our system is very efficient in object detection.

3.2 Multi-class Object Classification

In the multi-class object classification, we still used 5 object categories in Caltech6. We used all images of these five categories in our experiment. In each category, we randomly selected 30 samples for training and the rest for testing. As in object detection, we selected the most correlated C2 features (500 features) for each category and extracted task relevant ones (50 features) among them by our feather selection algorithm. We extracted totally 250 patches, 50 patches for each category. During selecting object relevant patches, we regarded the images of this category as positive samples and the others as negative samples.

Table 2. Performance of multi-class object recognition (see text)

Categories	Train	Test	One VS Rest	All
Leaves	30	156	99.6	
Airplanes	30	1044	91.7	
Motorbikes	30	796	96.8	95.0
Faces	30	420	98.8	
Cars	30	496	98.4	

Table 2 is the results of our multi-class object classification experiment. The second and third columns are number of train and test samples respectively. The fourth column is the classification rates between one category and the rest (only 50 features relevant to this category are used). The last column is the performance of our multi-class object recognition task by total 250 features. The results show that our system performs well even with a small number of training samples.

4 Summary and Discussion

In this paper, we have proposed a new object recognition model by introducing supervision to the local feature combination phase of traditional HMAX. HMAX model proposed by MIT-CBCL is a famous hierarchical object recognition model in cortex. But randomly selecting patches in S2 layer is one limitation of it. We find introducing supervision and extracting task relevant patches can resolve the problem effectively. By this improvement, higher recognition performance with less features are achieved (see Section 3).

One limitation of our system is long computation time in training, because computing C2 features for a patch is time consuming and we compute C2 features for all patches on training data. Fortunately, the testing procedure is mush faster than training, for only a few task relevant C2 features need to be calculated (50 features in our experiments).

In HMAX and our model, only shift and scale transformations are considered. In real world, we will face all kinds of other transformations, such as rotation, illumination, occlusion, etc. How to extract invariant features under these complex transformations is a challenging work. In the next step, we will modify our model and try to resolve these more challenging problems.

Acknowledgments. The work was supported by the National High-Tech Research Program of China (Grant No. 2006AA01Z125) and the national natural scinece foundation of China (Grant No. 60775007).

References

1. Roelfsema, P.R.: Cortical Algorithms for Perceptual Grouping. Annual Review of Neuroscience 29, 203–227 (2006)
2. Hubel, D., Wiesel, T.: Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. The Journal of Physiology 160, 106–154 (1962)
3. Kobatake, E., Tanaka, K.: Neuronal Selectivities to Complex Object Features in the Venral Visual Pathway of the Macaque Cerebral Cortex. Jounal of Neurophysiology 71, 856–867 (1994)
4. Logothetis, N.K., Sheinberg, D.L.: Visual Object Recognition. Annual Review of Neuroscience 19, 577–621 (1996)
5. Tanaka, K.: Inferotemporal Cortex and Object Vision. Annual Review of Neuroscience 19, 109–139 (1996)

6. Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., Fried, I.: Invariant Visual Representation by Single Neurons in the Human Brain. Nature 435(7045), 1102–1107 (2005)
7. Fukushima, K.: Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics 36(4), 193–202 (1980)
8. Perrett, D.I., Oram, M.: Neurophysiology of Shape Processing. Image and Vision Computing 11, 317–333 (1993)
9. Wallis, G., Rolls, E.T.: A Model of Invariant Object Recognition in the Visual System. Progress in Neurobiology 51, 167–194 (1996)
10. Riesenhuber, M., Poggio, T.: Hierarchical Models of Object Recognition in Cortex. Nature Neuroscience 2(11), 1019–1025 (1999)
11. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust Object Recognition with Cortex-like Mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3), 411–426 (2007)
12. Hyvarinen, A., Hoyer, P.O.: Topographic Independent Component Analysis as a Model of V1, Organization and Receptive fields. Neurocomputing 38, 1307–1315 (2001)
13. Olshausen, B.A., Field, D.J.: Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images. Nature 381, 607–609 (1996)
14. Inza, I., Larranaga, P., Etxeberria, R., Sierra, B.: Feature Subset Selection by Bayesian Network-based Optimization. Artificial Intelligence 123(1-2), 157–184 (2000)
15. Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. Artificial Intelligence 97(1-2), 273–324 (1997)
16. Kwak, N., Choi, C.H.: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(12), 1667–1671 (2002)
17. Oh, I.S., Lee, J.S., Moon, B.R.: Hybrid Genetic Algorithms for Feature Selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(11), 1424–1437 (2004)
18. Verikas, A., Bacauskiene, M.: Feature Selection with Neural Networks. Pattern Recognition Letters 23(11), 1323–1335 (2002)