Universität Hamburg
Department Informatik
Knowledge Technology, WTM

Cortical Computing for Invariant Object Recognition

Seminar Paper

Models inspired by cortical computing to achieve invariant object recognition

Leena Chennuru Vankdara

Matr.Nr. 6641141
4chennur@informatik.uni-hamburg.de

17.12.2014

Abstract

This paper tries to answer the following questions by providing evidence to either prove a hypothesis or to disprove a null hypothesis formulated to answer the following questions.

1) Can invariant computing be entirely achieved by using computational vision algorithms?
2) Which principles of cortical computing are relevant to solving the problem of invariant object recognition?
3) How can principles of cortical computing be applied to solve the problem of invariant object recognition?
4) Are the principles of cortical computing optimal for invariant object recognition?

Sub questions:

Why is object recognition hard?

Can increasing sparseness in the inputs improve the efficiency of object recognition?

Which computing architecture is better suited to learn invariance in object recognition?

Does a biological vision system learn invariance through temporal correlation?

What is the effect of the parameters like the number of neurons, number of selected features and the parameters used in the hardcoded Gabor filters?

Is it efficient to use hard coded mathematical models that estimate the behaviour of cortical elements or is it more efficient to learn the properties by training on a good statistical sample of images?

What is a good statistical sample that can efficiently represent the variability of the object in a real world scenario?

How can this model be tested on an efficient benchmark?

How can the efficiency of benchmarks for the problem of invariant object recognition be tested.

Contents

1 Why is Object Recognition Hard? Error! Bookmark not defined.

2 Principles of Cortical Computing Error! Bookmark not defined.

3 Models inspired by cortical processing in the ventral stream Error! Bookmark not defined.

3.1 Introduction ..... **Error! Bookmark not defined.**
3.2 HMAX..... **Error! Bookmark not defined.**
3.3 Serre ..... **Error! Bookmark not defined.**
3.4 Mutch and Lowe ……...4
3.5 Wurtz (HTM like Model)…….....4
3.5.a Wurtz (HTM like Model) A Discussion ……..5

4 Analysis Error! Bookmark not defined.

5 Effect of Sparseness of input images on efficiency of object recognition
5 Effect of learning vs usage of hard coded filters
5 Effect of temporal continuity vs spatial continuity in invariant learning
5 Effect of parameters on efficiency of learning

5 Discussion
6 Conclusion

Bibliography Error! Bookmark not defined.

Introduction

Invariant object recognition and coherent object representation have been major hurdles in developing efficient artificial vision systems. Invariant object recognition forms the core of the problem of developing egocentrical object based representations of the external world. Given the incredible ability of the biological vision systems which exhibit invariance to considerable amount of deformations it is natural to take inspiration from biological vision systems to build artificial vision systems which can perform invariant object recognition. In this paper, we evaluate different hierarchical models which attempt to mimic the cortical circuit design and the cortical architecture found in the neocortex of the brain to achieve object recognition. In the first section, we discuss the topic of object recognition and the challenges faced by Artificial Vision systems which try to achieve this goal. In the second section, a brief summary of design principles of cortical computing is presented along with a model (insert citation) linking the designs of cortical computing to behavioural properties of various forms of biological intelligence. In the third section, a brief history of object recognition is presented. In section 4, we evaluate 4 different biologically inspired models for object recognition. In section 5, we provide a comparative evaluation of these models with the cortical architecture model of the neocortex. In section 6, we discuss future directions of research in cortical computing and object recognition.

Section 1

Invariant Object Recognition

Earliest research in artificial vision systems were developed for optical character recognition and pattern recognition tasks. The major hurdle in building artificial vision systems has been to achieve recognitions invariant to translation, rotation, scaling and minor distortions and to build efficient 3D object representations. In the past few decades object recognition algorithms took inspiration from robust and efficient biological vision systems. Most computer vision algorithms extract features from a static image and compare it with a prototype vector by constructing a measure of similarity or using an unsupervised classifier. Some single stage feature extraction algorithms compute a histogram of features to achieve invariance to translation and rotation and lose spatial information.

Geometric blurring, where the template and the input feature are passed through a blurring function is very effective to achieve invariance to small geometric distortions.

Most of the state of the art object recognition algorithms use supervised learning.

1) Temporal sequences to train the network.
2) Feedback /Attention
3) Bottom up filtering
4) Perceptual grouping

Neuroscience Background (Biological Vision System)

Some of the earliest successful models of the vision system stem from Hubel and Wiesel's experiments on a cat's striate cortex. The experiments revealed the pattern of organization of the cells of the primary visual cortex and their function. The visual system is a hierarchical organization of cells grouped together into layers responsible for processing different properties of vision. Input signals from the eye pass through the retina and the Lateral geniculate nucleus (LGN) to reach the visual cortex of the brain. The visual cortex comprises of several cortical layers. Two different streams of hierarchical cortical areas, the ventral stream and the dorsal stream, compute complementary properties. The ventral stream, also called the "What Stream" is responsible for object recognition and the dorsal stream, also called the "Where Stream" is responsible for processing the spatial location of the object and information related to motion of the object. Together, they create egocentrical representations of objects in the external world. The ventral stream, which is responsible for object recognition is comprised of the V1, V2, V4, IT, PFC cortical areas.

Principles of Cortical Computing

Feedforward processing in the case of unambiguous object recognition

Evaluation of Object recognition systems

Most architectures for object recognition consist of a hierarchy of stages encompassing a combination of the following stages

1) Feature extraction
2) A nonlinear function (sparseness, thresholding, divisive normalization, local contrast normalization)
3) A pooling layer(MAX Pooling, Average Pooling) (to obtain local invariance)
4) A Classifier layer

Several low level feature extraction algorithms are available in the literature SIFT, HOW, BOW, HOG.

Neocognitron

The architecture of the current state of the art image recognition models can be traced back to Fukushima's Neocognitron

Serre and Poggio

Mutch and Lowe

Lessmann and Wurtz

One codebook of features per level.

Lessman and wurtz propose a model which is similar to the Memory Prediction Framework proposed by Hawkins and Blakeslee which is based on the ideas of laminar computing, temporal sequence learning and attention modelling to achieve invariance by increasing the sparseness and by using a global dictionary of features for each level. This model focuses on learning invariance in object recognition through temporal correlation. Several state of the art object recognition algorithms focus on feedforward computing of static images which is proposed (Grossberg et al) as the form of computing which is found in the neocortex while processing unambiguous images. However, as several real world recognition scenarios involve a high element of ambiguity, a simple feedforward form of computing which is based on the principles of cortical computing may not sufficiently encapsulate the solution to the problem of invariant object recognition. The current state of the art biologically inspired algorithms are wary to this limitation and thus build models of feedforward computing which could be integrated as a module of a system of higher functionality which can achieve invariant object recognition in the real world scenario. The model by Lessman and Wurtz is a step towards integrating the model of feedforward computing into a higher functional system which can achieve invariant object recognition. The model also varies from several state of the art object recognition algorithms by learning through temporal correlation in contrast to learning through static images.

The basic features of this model are presented here.

1) The network takes a laminar form with several stages (usually 3 or 4) with each layer consisting of two sublayers (1. **S** and 2. **T**)
2) The S layer detects the spatial patterns in the input signal (after training is complete) and the T layer determines the temporal group of the input signal with a certain probability.
3) Each layer consists of nodes and the correspondence between the S and T sub layers is one to one.
4) S layer in each stage is connected to T layer in the same stage and a group of nodes (usually 9) make an afferent convergence into one node in the S layer of the subsequent higher stage in the hierarchy.

5) Connections between the layers is bidirectional and both feedforward and the feedback connections share the same weights.
6) Training is not done simultaneously for all layers but can only be performed layer by layer.
7) Lateral inhibition is used to supress weakly active neurons to increase sparseness

Spatial feedforward input

Spatial pattern at the current node is first extracted. On the lowest level this is an image feature on the corresponding image positions.

A temporal group contains all the features that occur in a spatial node (image position in the lowest layer) in a particular time interval.

1) System has to be trained from sequences that represent images undergoing these transformations.
2) On each level, it builds a database of features (one per level) during learning and then it learns temporal groupings. Which means it saves the sequence of images which occur close in time into one temporal group. Then in the next spatial layer, it concatenates the spatially together temporal groups into one spatial group and then temporal groups are pooled again.
3) Activities are calculated by similarity to the prototype of the neuron. Activities to next layer are given by the weighted sum of the activations in the next layer.
4) In the lowest level Gabor jets are used as filters in various orientations and at each spatial position in the image, the feature is extracted.
5) In the subsequent spatial layers, a concatenation of indices of the most active temporal neurons of the node positions converging into that node. (Max Pooling) (not losing spatial information)
6) Lateral inhibition is done by setting all but K most active neurons to zero. This step can be seen as a step which minimizes the confusion as to which spatial pattern is relevant or which temporal pattern is observed in the current position in space and time.
7) Adds feedback input from the temporal images. This feedback is sent from the temporal group that has been active for the last T images. (like in LAMINART) (A trace of activity is stored in an activity stack to know which temporal group has been active for the past few images.)
8) Activation function is the hyperbolic tangent which prevents the system from having infinite activity values which can happen because of the feedback connections. (This step is analogous to the modulation and contrast divisive normalization observed in the biological neurons)
9) The two nonlinear functions used in this network are the tanh function used to limit the activities and the lateral inhibition used to supress all but the k most active neurons.
10) Activities of all the neurons for the past few image samples are stored in a stack in the same sequence and at every time step if same neurons at all the node positions are active in the previous image and the current image, then

only the activities are updated and if either new neurons turn active or if previously active ones become inactive then the last entry in the stack is deleted and the activities of the neurons of the current image are recorded into the current stack.

11) The stack is emptied after presenting each object category to prevent learning associations between different object categories.

Learning the prototypes for spatial neurons:

Prototypes are selected from the set of all the features in the codebook and are defined as the set of features for which the similarity between all the features in the codebook and any of these features from the feature set to be above a certain threshold.

To learn the codebook features, at each level, when the input features are presented the codebook is scanned to find the closest neighbour to the corresponding input feature. If the similarity between the input feature and the nearest neighbour is greater than the threshold (same as the one used in generating prototypes) then the codebook is not changed else, the codebook is appended with the new input feature.

Learning the temporal groups

An adjacency matrix representing the probabilities of two neurons to be active in a time interval T is computed. Activity stack holds the data needed to calculate this matrix. The matrix is clustered using spectral clustering using a stopping criterion which determines the size of the clusters.

Temporal groups make afferent connections to the spatial nodes of the immediate higher layer in the hierarchy and they represent the new spatial patterns. The indices are concatenated in a fixed order to maintain the spatial continuity and the same procedure is followed to extract spatial patterns and associate temporal groupings in the consecutive layers in the hierarchy.

Weights between both the spatial and temporal neurons in the same level as well as between the temporal neurons of one layer and the spatial neurons of the next layer are derived based on the similarity measures (temporal similarities of spatial patterns and spatial similarities of temporal patterns).

All parameters involved are computed based on experimentation and determined differently when tested on different images.

To maintain spatial relationships between different temporal groups and to minimize the Probability of erroneous classification, the concatenation of indices is used.

Analysis and Discussion

Even though the recognition rates of these state of the art object recognition algorithms has been considerably significant on the benchmark image datasets, they are far from matching the capabilities of biological vision systems.

Single Processing Unit for all sensory modalities

On temporal correlation of input signals

The idea of temporal correlation of input signals is as follows. One of the suggested mechanisms by which the biological vision system learns invariance to transformations is through temporal connectivity. The idea is that objects that occur close in time have a higher probability to be associated with different views of the same object rather than a different object. Hence in principle, if a network designed to implement this idea is presented with objects changing slowly in time under certain transformations, with enough number of training samples, the system achieves invariance to the corresponding transformations. This idea however is still debated upon. For instance, continuous transformation (Rolls) is suggested as an alternative to temporal correlation to achieve invariant object recognition. The idea of continuous transformations is that as objects change slowly in time under transformations and hence there is bound to be sufficient overlap between input signals to excite the same neuron in the higher level which improved its connection strengths by winning in response to presentation of the previous stimulus. Most state of the art object recognition algorithms do not rely on temporal correlation to achieve invariance but use methods like Max Pooling or Average Pooling to achieve invariance to transformations. In the past years several successful object recognition algorithms like Slow Feature Analysis (Wiskott) on Convolutional Neural Networks have incorporated the idea of temporal correlation to achieve invariance.

The question of whether temporal correlation is used to achieve invariant object recognition has been very interesting and several psychophysical experiments could be found in the literature both to support as well as to counter this idea. For instance, Cox and Carlo designed an experiment based on the idea that if spatiotemporal statistics were thought to be used to learn invariance, it should also be possible to modify the statistics to disrupt invariance to transformations. The hypothesis postulates that if the visual system learns to associate images that occur closely in time across saccades into a single object. Participants are presented with views of different images across saccades for one test group along with a control group in which they are presented with different views of the same object across saccades. After a relatively brief period of exposure participants in the first group exhibited significantly high amount of confusion in associating different views of different object to a single object.

Conclusion