SIMULATING HIERARCHICAL STRUCTURE OF HUMAN VISUAL CORTEX FOR IMAGE CLASSIFICATION

SEPEHR JALALI

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE

2013

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in this thesis. This thesis has also not been submitted for any degree in any university previously.

Sepehr Jalali

31 May 2013

Acknowledgement

I would like to express my deepest gratitudes to my supervisors: Dr Lim Joo Hwee, Prof. Ong Sim Heng and Dr Tham Jo Yew who have led me into this wonderful field. Without their guidance, inspirations, support and encouragement, this research project would not have been possible. I also express my appreciation to Dr Cheston Tan for great guidance, discussions and collaborations.

Gratitudes are also due to Prof. Daniel Raccoceanu, Dr Paul Seekings and Dr Elizabeth Taylor for their support. I would also like to express my gratitude to Prof. Cheong Loong Fah, Dr. Yeo Chuo Hao, Prof. Chong Tow Chong, Dr Shi Lu Ping and Dr Kiruthika Ramanathan, Prof. Tomaso Poggio, Prof. Thomas Serre, Jim Mutch, Dr Christian Theriault and Jun Zhang for discussions and collaborations. I would also like to convey thanks to the A*STAR Graduate Academy (A*GA) for providing the scholarship, tuition fees and conference trip expenses; A*STAR's Institute for Infocomm Research (I^2R) for computational resources and support; and Image and Pervasive Access Lab (IPAL) for providing the financial support, and special thanks also to all my friends who have always been there.

Last but not least, I express my love and gratitude to my beloved family for their support, understanding and endless love, throughout the duration of my studies. I dedicate this thesis to my beloved family for their endless and unwavering love throughout my life.

Contents

Summary

Image recognition is one of the most challenging problems in computer science due to different illumination, viewpoints, occlusions, scale and shift transforms in the images. Hence no computer vision approach has been capable of dealing with all these issues to provide a complete solution. On the other hand, the human visual system is considered a superior model for various visual recognition tasks such as image segmentation and classification as well as face and motion recognition. Exceptional fast performance of human visual system on image recognition tasks under different resolutions (scales), translations, rotations and lighting conditions has motivated researchers to study the mechanisms performed in the human and other mammals' visual system and to simulate them. Recent achievements in biologically inspired models have motivated us to further analyze these hierarchical structure models and investigate possible extensions to them.

In this thesis, we study several hierarchical models for image classification that are biologically inspired and simulate some known characteristics of visual cortex.

We base our investigation on the HMAX model, which is a well-known biologically inspired model (Riesenhuber and Poggio, 1999), and extend this model in several aspects such as adding clustering of features, evaluating different pooling methods, using mean pooling (HMean) and max pooling in the model as well as coding occurrences and co-occurrences of features

with the goal of improving the image classification accuracy on benchmark datasets such as Caltech101 and a subset of Caltech256 (classes with a higher number of training images) and an underwater image dataset. We introduce several self organizing maps and clustering methods in order to build mid-level dictionary of features. We also investigate the use of different pooling methods and show that concatenation of biologically inspired mean pooling with max pooling as well as enhanced models for encoding occurrences and co-occurrences of features on a biological feasibility basis improves the image classification results.

We further propose a new high-level biologically inspired color model, CQ-HMAX, which can achieve better performances than the state-of-the-art using the bottom-up approaches when combined with other low-level biologically inspired color models and HMean on several datasets such as Caltech101, Soccer, Flowers and Scenes. We introduce a new dataset of benthic marine organisms and compare different proposed methods.

We also propose an HMAX like structure for simulating auditory cortex and create sonar images and combine them with visual images for underwater image classification in poor visibility conditions. We also show the use of HMAX and CQ-HMAX models on other tasks such as detection of mitosis in histopatholgy images and propose several future directions on this field of study.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivations

Image classification includes a broad range of approaches to the identification of images or parts of them. In classification of images, each image is assumed to have a series of features that distinguish that particular image from other images. Different approaches are proposed to extract features such as geometric parts, spectral regions, histogram of pixels in color or grayscale, using templates of the target of interest or other features from images. These approaches generally fall into two categories, namely, supervised and unsupervised (or a combination of them).

These approaches can be bottom-up, top-down, or interactive based on the contextual information from the images. Object rotations, occlusions, different viewpoints, scales and lighting in the images are among the factors that make image classification a complex process. As a result, the complete method that can incorporate all these issues based on the computational

approaches of computer vision has not been successful.

On the other hand, human visual capabilities in dealing with these issues have inspired many scientists to study the visual cortex of humans and other mammals to gain a better understanding of it and to simulate how these processes take place in the brain based on the current findings. In addition there is active ongoing research in both directions (biologically inspired methods and computer vision approaches) towards a holistic framework that can deal with all these issues.

1.2 Human Visual Cortex

Research on the human visual cortex suggests a hierarchical structure in which each level of the hierarchy is assumed to be responsible for specific roles and sends its output to the higher levels, as can be seen in Figure 1.1.

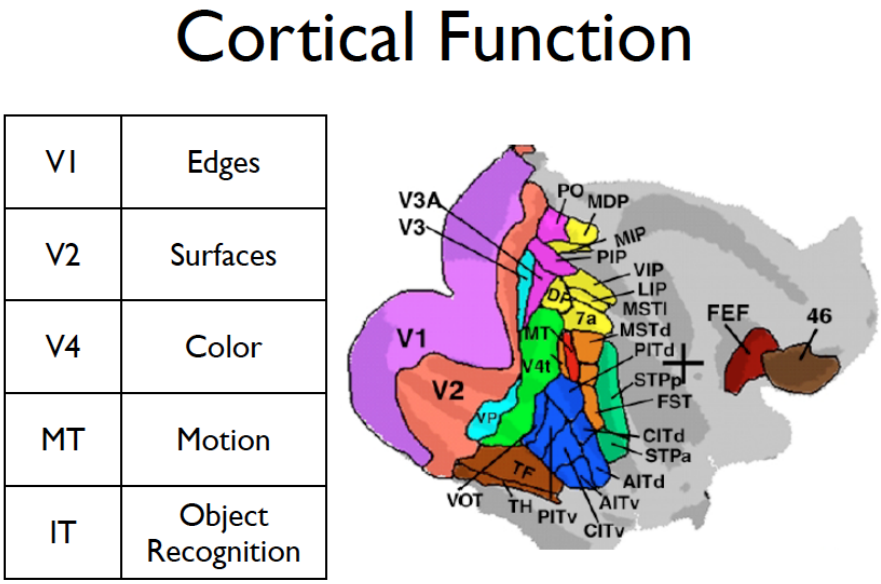


Figure 1.1: Different roles proposed for different layers of human visual system hierarchy in Goldstein (2009).

Visual cortex is a part of the cerebral cortex located in the occipital lobe, which includes striate cortex or $V1$ and extrastriate visual cortical areas such as $V2$, $V3$, $V4$ and $V5$/MT, and is responsible for processing visual information. The information acquired by $V1$ is transmitted in two primary pathways called the dorsal and ventral streams. The dorsal stream begins with $V1$, goes through $V2$ and $V5$/MT and to the posterior parietal cortex. This pathway is also referred to as "Where pathway" or "How pathway". The ventral stream, begins with $V1$, followed by $V2$ and $V4$ and to the inferior temporal cortex (IT). This pathway is also called the "What pathway" which is associated with the recognition and object representation and storage of long term memory (Mishkin et al., 1983). These layers have interactions with each other via feedback, feedforward and inter-level connections.

Object recognition in cortex is thought to be mediated by the ventral visual pathway running from primary visual cortex, $V1$, over extrastriate visual areas $V2$ and $V4$ to inferotemporal cortex, IT Riesenhuber and Poggio (1999).

Over the last decades, several physiological studies in non-human primates have established a core of basic facts about cortical mechanisms of recognition that seem to be widely accepted and that confirm and refine older data from neuropsychology. A brief summary of this consensus knowledge begins with the ground-breaking work of Hubel and Wiesel first in the cats (Hubel and Wiesel, 1962, 1965) and then in the macaque (Hubel and Wiesel, 1968). Starting from simple cells in primary visual cortex, $V1$,

with small receptive fields that respond preferably to oriented bars, neurons along the ventral stream show an increase in receptive field size as well as in the complexity of their preferred stimuli Riesenhuber and Poggio (1999). At the top of the ventral stream, in anterior inferotemporal cortex (AIT), cells are tuned to complex stimuli such as faces. A hallmark of these IT cells is the robustness of their firing to stimulus transformations such as scale and position changes. In addition, as other studies have shown, most neurons show specificity for a certain object view or lighting condition (Sigala et al., 2005; Olshausen et al., 1993).

Since Hubel and Wiesel (1959) introduced simple and complex cells in the early processing in visual system (Figure 1.2), a series of models were proposed to simulate this hierarchical structure. HMAX Riesenhuber and Poggio (1999) and HTM (George, 2008) are among these models. Some other biologically inspired models are tackling the problem with a more probabilistic approach like Deep Belief Networks (DBN) (Hinton et al., 2006) using Restricted Boltzmann Machines (RBM) which will be further discussed in Chapter 2.

There are also computational evidences that hierarchical structures such as spatial pyramid matching and deep belief networks are more powerful than traditional linear approaches. Computationally speaking, functions that can be compactly represented by a depth k architecture might require an exponential number of computational elements to be represented by a depth $k-1$ architecture. Since the number of computational elements one can afford depends on the number of training examples available to tune

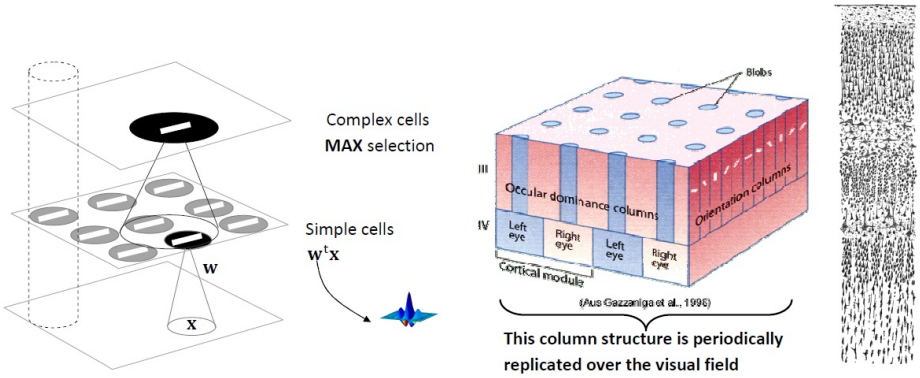


Figure 1.2: Hubel and Wiesel's model of simple and complex cells in visual cortex (right) and HMAX simulation (left).

or select them, the consequences are not just computational but also statistical: poor generalization may be expected when using an insufficiently deep architecture for representing some functions (Bengio, 2009).

The depth of an architecture is the maximum length of a path from any input of the graph to any output of the graph. Although depth depends on the choice of the set of allowed computations for each element, theoretical results suggest that it is not the absolute number of levels that matters, but the number of levels relative to how many are required to represent the target function efficiently (Bengio, 2009). Kernel machines, with a fixed kernel can be considered as two level structures. Boosting usually adds one level to its base learners. Artificial neural networks normally have two hidden layers and can be considered two layer structures. Decision trees are also considered two layer structures. According to the observations we have from the human's visual system, there are several layers in the brain that work in a hierarchical structure to interpret the images and perform cognition and recognition in the brain (Serre et al., 2007a).

1.3 HMAX Biologically Inspired Model

HMAX, proposed by Riesenhuber and Poggio (1999), is a model that simulates the simple-complex cell hierarchy in the visual cortex. The model reflects the general organization of visual cortex in a series of layers from $V1$ to IT to PFC. In the standard HMAX model, there are four layers of hierarchy (namely, $S1$, $C1$, $S2$ and $C2$) that create the features for the classifier and there is a supervised classifier on top as can be seen in Figure 1.3. A pyramid of Gaussian filters are convolved on the images in $S1$ layer, and a local max is calculated on small neighborhoods in $C1$ layer. A handmade dictionary of features that contains more complex features is convolved on the $C1$ layer, and the $S2$ layer is thus created. A global max is taken on $S2$ layer to create the $C2$ layer, and the outputs are then fed to a classifier such as a support vector machine (SVM).

Subsequent extensions to this model have improved it for image classification tasks to compete with the state-of-the-art computational models. We will explain the HMAX model in more detail and provide an extensive review on the extensions to the base model in Chapter 2. Serre and Riesenhuber modified the standard HMAX structure and released a new version of this structure (Serre and Riesenhuber, 2004). Gabor filters were used instead of second order Gaussian derivatives in $S1$ layer, and the number of filter sizes was increased. They also changed the values of scale range and pool range parameters in standard HMAX in $C1$ layer to provide less scale tolerance and therefore narrower spatial frequency bandwidth (Serre and Riesenhuber, 2004). Two other layers were added to the standard model to

simulate bypassing of information. This model includes $S2b$, $S3$, $C2b$, $C3$, and $S4$. They also suggested a random sampling of features from $C1$ layer in order to replace the handmade dictionary of features in HMAX model.

Mutch et al. (Mutch and Lowe, 2008; Mutch et al., 2010a) proposed a series of computational modifications to the structure proposed by Serre et al.'s model. In this model, a fixed size of Gabor filters is implemented on different scales of the images which provides the same invariance to scale for Gabor filters (Mutch and Lowe, 2008, 2006). They also investigated the use of Sparse features. Theriault et al. (2011) suggested using multi-scale sparse features and replaced Guassian response in $S2$ layer with a normalized dot product.

1.4 Scope, Contributions and Organization of Thesis

In this thesis, we propose several modifications, enhancements and applications for HMAX model as follows:

(i) Non-random sampling methods for creation of the dictionary of features such as clustering and saliency points;

(ii) Different pooling methods and encoding occurrences and co-occurrences of features in the intermediate layers;

(iii) A new high-level biologically inspired color model (CQ-HMAX); and

(iv) Applications of HMAX model in other image classification tasks.

All the modification made to the main model are biologically inspired or consistent with the existing evidence from the visual cortex mechanisms, which we will illuminate in detail in the following Chapters.

In Chapter 2, we have an overview, comparison and a discussion on several pertinent models available in the literature. We introduce biologically inspired models such as HTM (George, 2008), LeNet (LeCun and Bengio, 1995), Dynamic Routing Model (Olshausen et al., 1993), Hierarchical Statistical Learning (Fidler et al., 2008), Top-Down Hierarchy of Features (Bart et al., 2004) , NeoCognitron (Fukushima, 1980) and computational approach of bag of features (Li and Perona, 2005), DBN (Hinton et al., 2006) and HMAX model (Riesenhuber and Poggio, 1999).

In Chapter 3 we investigate HMAX model in more detail and review the main modifications made to it. We discuss this model and provide several modifications and improvements built on top of the previous enhancements to the model which are both biologically inspired and result in better classification performances on different datasets over the existing HMAX model performance.

The general structure of HMAX model is shown in Figure 1.3 and the main contribution areas to be covered in this thesis are highlighted by red circles.

In Chapter 4 we present modifications to the creation of the dictionary of features using several self organizing maps, clustering methods and saliency points selection and discuss the significant improvement that is achieved by using spatial and frequency information of the features in the

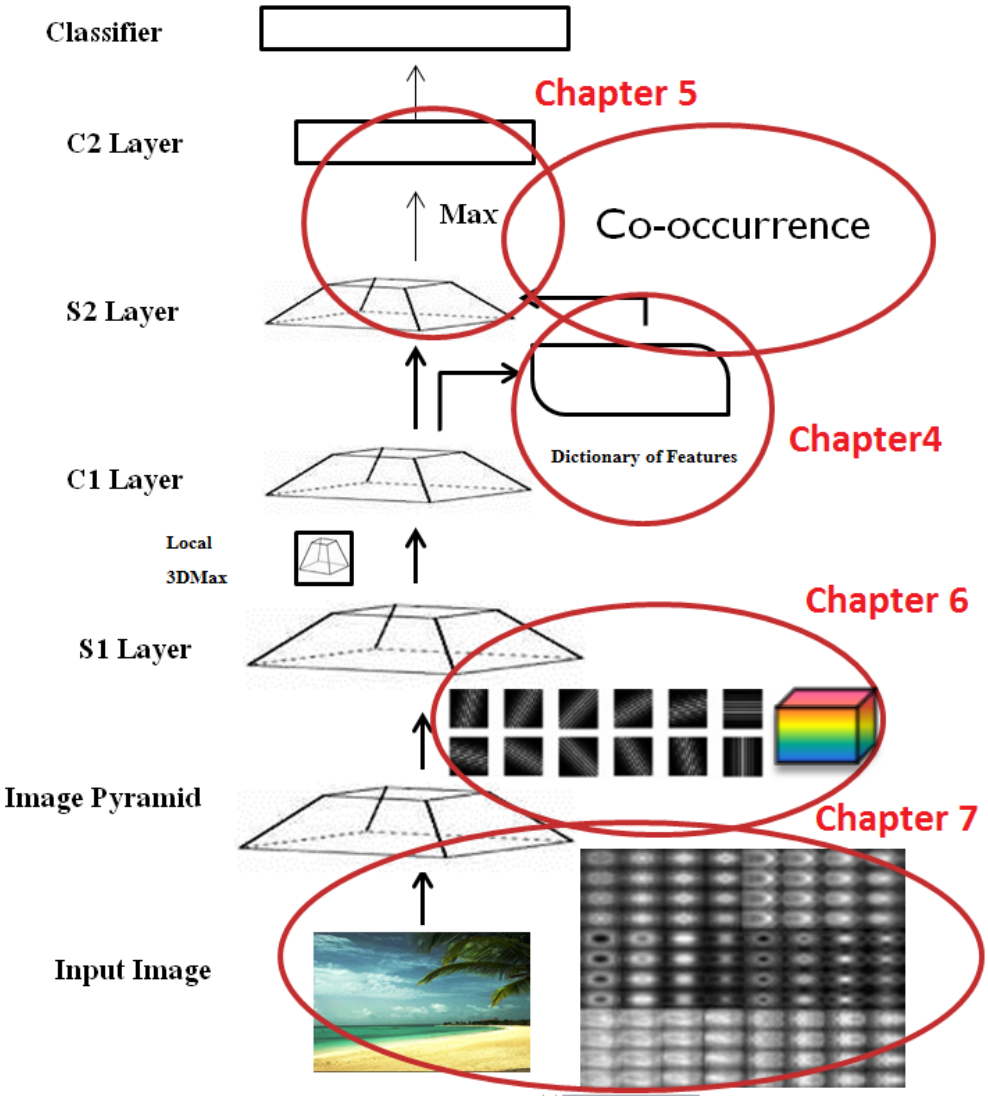


Figure 1.3: A summary of main contributions on the HMAX model.

creation of the dictionary of features.

In Chapter 5 we incorporate the mean pooling method into HMAX (named HMean), and provide different methods for encoding occurrences and co-occurrences of complex features in the HMAX model. The concatenation of HMean and HMAX models results in significant improvements over classification results in several datasets. Encoding co-occurrences of features without any top-down or heuristic interactions further improves the classification results when a higher number of training images is available.

In Chapter 6 we introduce a new biologically inspired high-level color approach, CQ-HMAX which is similar to HMAX in structure and show that using this model, we can achieve higher classification accuracy on several datasets and concatenation of this model with the low-level biologically inspired color model of Zhang et al. (2012) further improves the classification performance to performances as good or better than the state-of-the-art bottom-up approaches on several benchmark color datasets.

Chapter 7 provides some applications of the HMAX model in other datasets such as benthic marine organisms and mitosis detection. We show that higher classification results can be achieved using HMAX feature when compared with some other well-known techniques that deploy popular feature extraction/cassification such as SIFT (Lowe, 1999). We also propose a new structure using HMAX model in simulating acoustic information acquired from underwater sonar systems to resemble the marine mammal auditory and visual systems and show that a combination of visual and

sonar images results in a better classification accuracy in poor underwater visibility conditions.

We provide a discussion in Chapter 8 followed by further suggestions for the future directions for this interesting field of research.

Chapter 2

A Review of Related Models in Image Classification

This chapter introduces the most well-known hierarchical and biologically inspired models that are used for image classification and are related to our model and discuss these models. Chapter 3 will provide a detailed description of the HMAX model and its various extensions.

Here we briefly introduce the following biologically inspired models:

- Dynamic Routing Model;
- Top-Down Hierarchy of Features; and
- Interactive Activation and Competition Model.

Dynamic Routing Model and Top-Down Hierarchy of Features are two hierarchical models that have demonstrated significant improvements over non-hierarchical models. We also introduce Deep Belief Networks (DBN) which have a hierarchical statistical structure that resembles some of the

characteristics of the human visual cortex and introduce Bag of Features (BoF) method which has been among successful computer vision approaches:

- DBN; and
- Bag of Features.

We introduce DBN as a successful hierarchical structure and draw inspirations from the BoF method for encoding the occurrences of features in HMAX model.

We introduce Hierarchical Temporal Memory, LeNet, NeoCognitron, Hierarchical Statistical Learning and HMAX models which have a similar simple-complex cells structure based on the hierarchical structure proposed by Hubel and Wiesel (1959).

- HTM;
- LeNet;
- NeoCognitron;
- Hierarchical Statistical Learning; and
- HMAX and Extensions.

We have a discussion on the above mentioned models and explore HMAX model (Riesenhuber and Poggio, 1999) and it's extensions in Chapter 3 in more detail:

- Serre *et al.*;

- Mutch *et al.*;
- Masquelier *et al.*; and
- Theriault *et al.*.

We compare these models and provide biological inspirations and justifications for the further extensions we have made to the HMAX model including the use of clustering of features, encoding occurrences and co-occurrences of features and the use of color information in our new CQ-HMAX model in the following chapters.

2.1 Overview

Human visual cortex has a hierarchical structure as introduced in Section 1.2. However, different roles are proposed for each layer, and there is no perfect understanding of the processes taking place in each layer and the exact connections among the layers are not known.

Several models are suggested for simulating the human visual cortex and the image understanding capabilities of human. The rest of this chapter briefly discusses several well-known models, followed by a more detailed discussion of the HMAX model.

2.2 Related Models

In this section, we will describe three models: Dynamic Routing Model, Top-Down Hierarchy of Features, and Interactive Activation and Competi-

tion Models. We also introduce Deep Belief Networks (DBN) which have a hierarchical statistical structure that resembles some of the characteristics of the human visual cortex, and the Bag of Features (BoF) methods which have been among the most implemented computational computer vision methods.

2.2.1 Dynamic Routing Model

This model relies on a set of control neurons to dynamically modify the synaptic strengths of intracortical connections so that information from a windowed region of primary visual cortex ($V1$) is selectively routed to higher cortical areas (see Figure 2.1). Local spatial relationships (i.e. topography) within the attentional window are preserved as information is routed through the cortex. This enables attended objects to be represented in higher cortical areas within an object-centered reference frame that is position and scale invariant (Olshausen et al., 1993).

2.2.2 Top Down Hierarchy of Features

Bart et al. (2004) proposed a top-down feature extraction method in which they start by N random large features and select the most informative ones as the top level nodes, and inside each selected patch, they select the most informative sub-patches (see Figure 2.2). If the information is increased using these nodes, they add these as children in the tree and repeat these steps until no more information is added. The last selected nodes are atomic features such as edges, corners, etc.

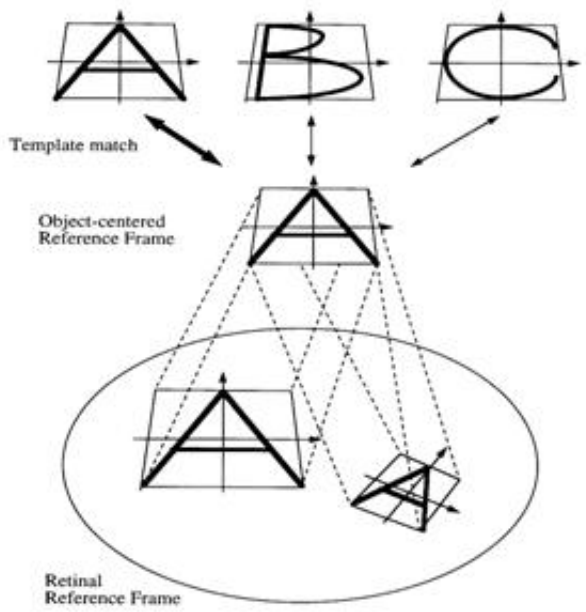


Figure 2.1: Dynamic Routing Model (Olshausen et al., 1993).

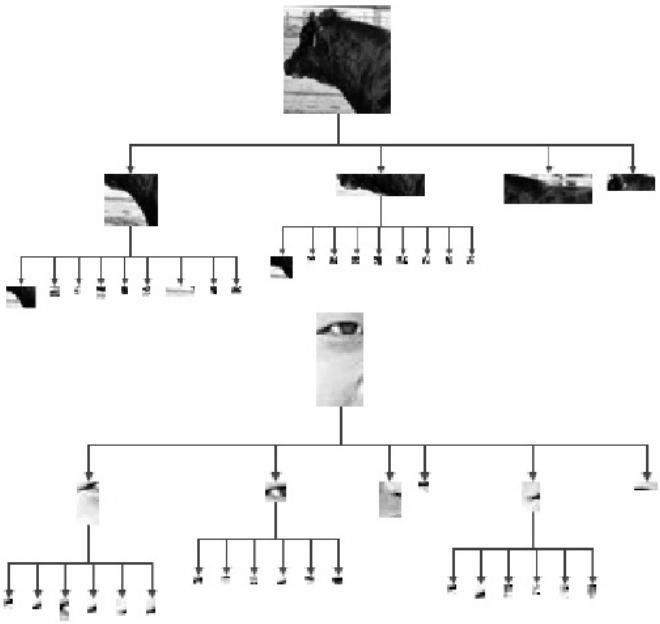


Figure 2.2: Top-Down Hierarchy of Features (Bart et al., 2004) .

This approach is different from the bottom-up segmentation methods that use the continuity of grey-level, texture, and bounding contours. They show that this method leads to improved segmentation results and can deal with significant variations in shape and varying backgrounds. This model is a successful example of hierarchical structure for segmentation (which can be used in classification).

2.2.3 Interactive Activation and Competition Network

The Interactive Activation and Competition Network (IAC) proposed by McClelland and Rumelhart (2002) consists of a number of competitive pools of units (see Figure 2.3). Each unit represents some micro-hypothesis or feature. The units within each competitive pool are mutually exclusive features and are interconnected with negative weights. Among the pools, positive weights indicate features or micro-hypotheses that are consistent. When the network is cycled, units connected by positive weights to active units become more active, while units connected by negative weights to active units are inhibited. The connections are in general bidirectional, making the network interactive (i.e. the activation of one unit both influences and is influenced by the units to which it is connected).

Interactive Activation and Competition model is a model that uses interaction between co-occurring units and enhances their connection weight and decreases the weight of the non co-occurring units. Inspirations from this model can be used for encoding co-occurrence of features in HMAX

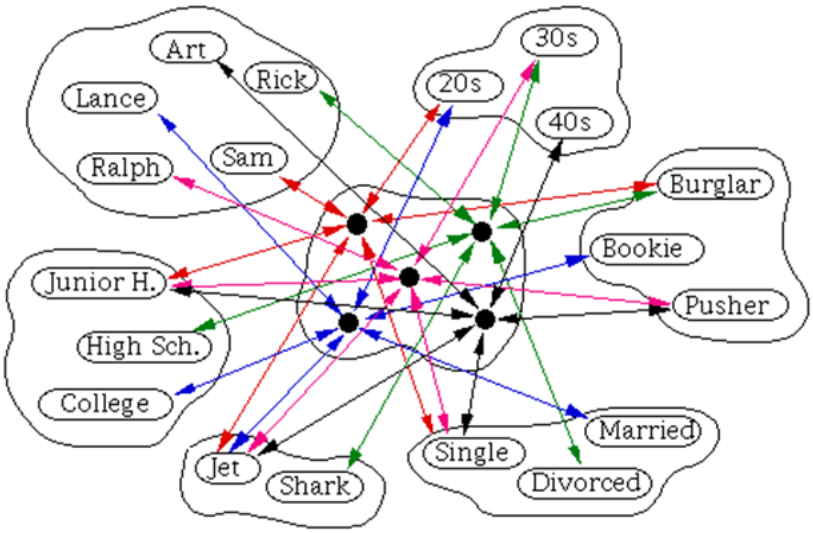


Figure 2.3: Interactive Activation and Competition Model.

model.

2.2.4 Deep Belief Networks

Deep Belief Networks (DBNs) are probabilistic generative models that are composed of multiple layers of stochastic, latent variables (see Figure 2.4). The latent variables typically have binary values and are often called hidden units or feature detectors. The top two layers have undirected, symmetric connections between them and form an associative memory. The lower layers receive top-down, directed connections from the layer above. The states of the units in the lowest layer represent a data vector. DBNs have successfully been used to learn high-level structure in a wide variety of domains, including handwritten digits (Hinton et al., 2006) and human motion capture data (Taylor et al., 2007).

A DBN can be viewed as a composition of simple learning modules, each of which is a type of Restricted Boltzmann Machine (RBM) that contains a

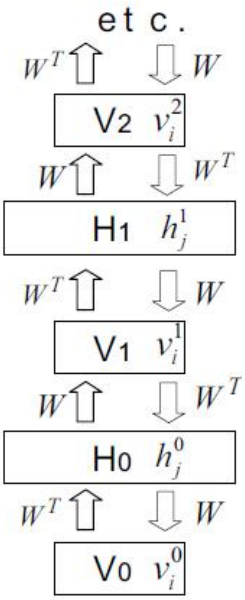


Figure 2.4: Deep Belief Networks (Hinton et al., 2006).

layer of visible units that represent the data and a layer of hidden units that learn to represent features of higher-order correlations in the data. The two layers are connected by a matrix of symmetrically weighted connections W, and there are no connections within a layer. Given a vector of activities v for the visible units, the hidden units are all conditionally independent so it is easy to sample a vector h, from the factorial posterior distribution over hidden vectors, $P(h|v, W)$. It is also easy to sample from $P(v|h, W)$. By starting with an observed data vector on the visible units and alternating several times between sampling from $P(h|v, W)$ and $P(v|h, W)$, it is easy to learn a signal. This signal is simply the difference between the pairwise correlations of the visible and hidden units at the beginning and end of the sampling. DBNs typically use a logistic function of the weighted input received from above or below to determine the probability that a binary latent variable has a value of 1 during top-down generation or bottom-

up inference, but other types of variables can be used and the variational bound still applies, provided the variables are all in the exponential family.

DBNs have been used for generating and recognizing images, video sequences, and motion-capture data (Taylor et al., 2007). If the number of units in the highest layer is small, DBNs perform non-linear dimensionality reduction and they can learn short binary codes that allow very fast retrieval of documents or images (Salakhutdinov and Hinton, 2009; Bengio and LeCun, 2007; LeCun et al., 1998; Hinton et al., 2006).

2.2.5 Bag of Features

A simple approach to classifying images is to treat them as a collection of regions, describing only their appearance and ignoring their spatial structure. Similar models have been successfully used in the text community for analyzing documents and are known as "bag-of-words" models (Harris, 1954), since each document is represented by a distribution over fixed vocabulary(s). Using such a representation, methods such as probabilistic latent semantic analysis (pLSA) and Latent Dirichlet Allocation (LDA) are able to extract coherent topics within document collections in an unsupervised manner. Bag of features is a well known computational approach that uses the histograms of features frequencies for image classification (Li and Perona, 2005). The key idea is to find a series of features in the images and based on the frequency of features perform the classification task (see Figure 2.5). Several approaches have been considered for the problem of finding the best features. Regular grids, interest point detectors such

as SIFT (Lowe, 1999), random sampling and segmentation based patches have been used and compared. In order to perform the classification, these histograms of frequencies are fed to a classifier such as Support Vector Machine (SVM). In other approaches, a fusion of these frequencies and other features in the image are fed to the classifier.

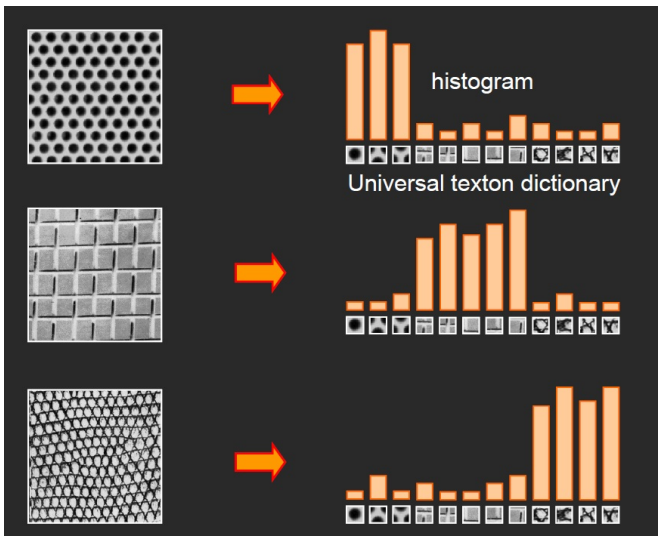


Figure 2.5: Bag of Features (Li and Perona, 2005).

This concept can be used in HMAX model to encode frequency of features and we use this method and introduce the HMean model in the following chapters.

2.3 Simple-Complex Cells Hierarchical Models

A series of biologically inspired models to image classification are proposed based on the simple and complex cells structure introduced by Hubel and Wiesel (1959). They found two types of cells in visual primary cor-

tex called simple and complex cells, and also proposed a cascading model of these two types of cells, as can be seen in Figure 1.2. In this section, we briefly introduce these models and provide a deeper review on HMAX model and its extensions in Chapter 3.

2.3.1 Hierarchical Temporal Memory

Hierarchical Temporal Memory (HTM) is a method proposed by George and Hawkins (2009), inspired from the book "On Intelligence" (Hawkins and Blakeslee, 2005). The HTM network is organized in a 3-level hierarchy. In each level, there is a temporal and a spatial pooler.

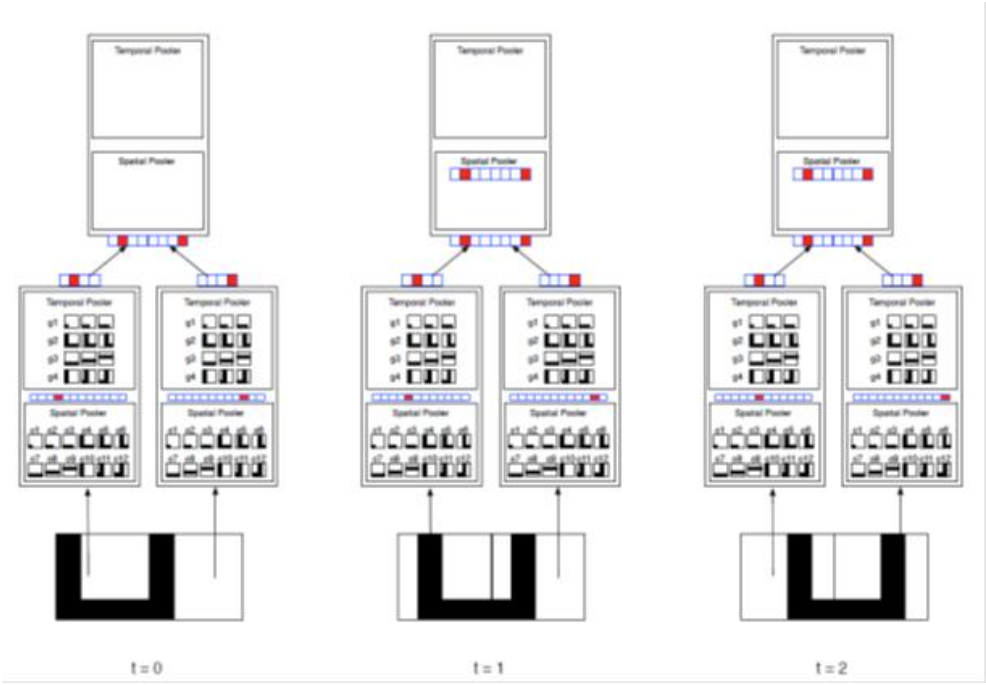


Figure 2.6: Operation of nodes in a hierarchy: this illustrates how nodes operate in a hierarchy. The bottom-level nodes have finished learning and are in inference mode (George and Hawkins, 2009).

The HTM network operates in two distinct stages: training and infer-

ence. As can be seen in Figure 2.6, during the training stage, the network is exposed to movies of images, and the nodes in the network form representations of the world using the learning algorithms. When learning is complete, the network is switched to inference mode. The input to a node, irrespective of its position in the hierarchy, is a temporal sequence of patterns. A node contains two modules:

1. **Spatial Pooling**: Learns a mapping from a potentially infinite number of input patterns to a finite number of quantization centers. The output of the spatial pooling is in terms of its quantization centers. The spatial pooling has two stages of operation: (a) During the learning stage, it quantizes the input patterns and memorizes the quantization centers; and (b) Once these quantization centers are learned, it produces outputs in terms of these quantization centers. This is the inference stage.

2. **Temporal Pooling**: Learns temporal groups of quantization centers, according to the temporal proximity of occurrence of the quantization centers of the spatial pooling. The output of the temporal pooling is in terms of the temporal groups that it has learned. Markov chains are used for the temporal grouping part and Bayesian Networks are employed to do the updates in the feed-forward and feed-back phase. In a modification to this mode, Bayesian networks were replaced by a competitive network and the performance of the structure is reported to be improved on the moving bit-worm dataset (Ramanathan et al., 2009). Competitive networks are replaced with a version of GSOMs (our previous unpublished work) to perform clustering and this show better results in some experiments.

2.3.2 LeNet

LeCun's convolutional neural networks (LeCun and Bengio, 1995) are organized in layers of two types: convolutional layers and sub-sampling layers (Figure 2.7). Each layer has a topographic structure i.e. each neuron is associated with a fixed two dimensional position that corresponds to a location in the input image, along with a receptive field (the region of the input image that influences the response of the neuron). At each location of each layer, there are a number of different neurons, each with its set of weights, associated with neurons in a rectangular patch in the previous layer. The same set of weights, but with a different input rectangular patch, is associated with neurons at different locations.

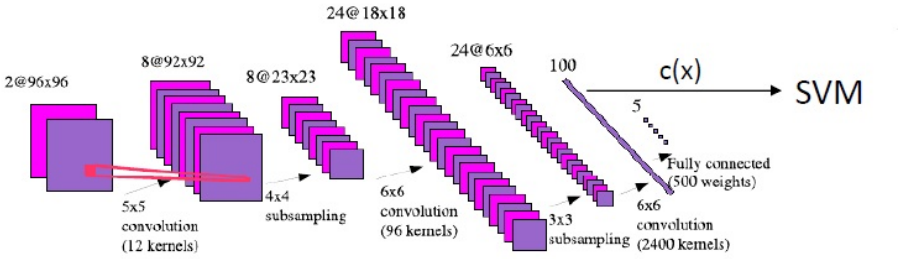


Figure 2.7: LeNet (LeCun and Bengio, 1995).

Even with random weights in the first layers, a convolutional neural network performs well, i.e. better than a trained fully connected neural network but worse than a fully optimized convolutional neural network.

2.3.3 Neocognitron

Neocognitron (Fukushima, 1980) is a hierarchical multi-layered neural network. The Neocognitron is a natural extension of the cascading models

(Figure 2.8). In Neocognitron, which consists of two types of cells called S-cell and C-cell, the local features are extracted by S-cells, and these features' deformation, such as local shifts, are tolerated by C-cells. Local features in the input are integrated gradually and classified in the higher layers. In later extensions to this model (Fukushima, 1988), the idea of 'winner kills loser' in simple layers and sum (instead of max) in complex layers, has been shown to improve the model.

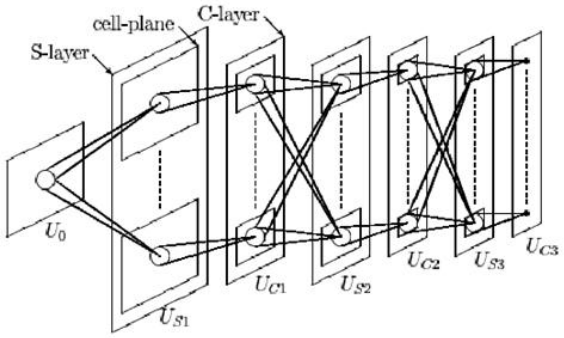
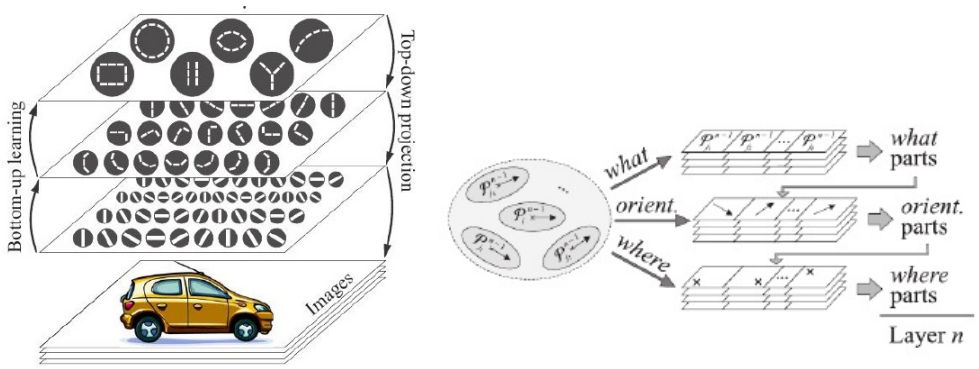


Figure 2.8: Neocognitron (Fukushima, 1980).

2.3.4 Hierarchical Statistical Learning

Fidler et al. (2008) proposed a hierarchical statistical learning approach that is similar to HMAX architecture in lower layers . They use favorable statistics of images to learn parts (Figure 2.9a).

They use Gabor filters in $S1$ layer, and based on the outputs of these filters, they define sub-parts and parts. The position and orientation of each sub-part is described with respect to the center of the mass and orientation of p_i^n where n is the layer number and i is the feature index. Some variance is allowed in the exact position of sub-parts. The simplest parts (layer



(a) Hierarchical statistical learning. (b) Learning parts.

Figure 2.9: Left: Hierarchical Statistical Learning. Right: Learning statistics in images Fidler et al. (2008).

1) are On/Off Gabor filters. A part p_i^n of layer n is built from a list of m sub parts $\{p_{i-1}^n\}_{j=1..m}$. There are bottom-up and top-down learning in this approach. In bottom-up learning of parts, statistics about {*what, orientation, where*} are gathered in the neighborhood of each type of part p_i^{n-1} of layer $n-1$ and only the most frequent configurations of {*what, orientation, where*} are selected to be a part p_i^n of layer n. In the top-down selection, each activated part p_i^{n-1} votes for every part p_i^n that contains p_i^{n-1} in its list of sub-parts, and each part p_i^n that receives votes from all its sub parts is selected (Figure 2.9b). A final subset of parts from step 1 is selected by minimal descriptor length algorithm.

2.3.5 HMAX Model

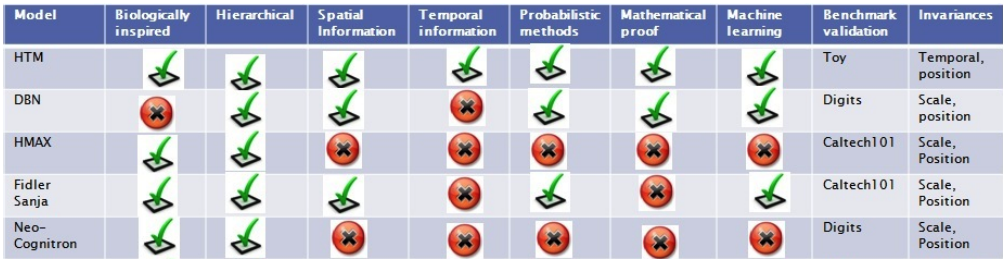
HMAX is a computational model of object recognition in cortex proposed by Riesenhuber and Poggio (1999). The standard model simulates the feed-forward path of the visual cortex and has been first used to classify

animal vs. non-animal images and paper clip images. This model is used to find a good trade-off between invariance and selectivity. $S1$ cells provide selectivity by responding to oriented filters and $C1$ cells provide invariance by pooling over neighboring scales and positions. There are several extensions to this model in the recent years. In Chapter 3, we describe this model in detail and provide a review on the extensions and improvements made to it.

2.4 Comparisons and Discussions

Several researchers have built pattern/object recognition systems with multiple levels. Neocognitron, convolutional neural networks, HMAX and HTM are examples of these models. Boltzmann machines and DBNs provide another set of examples of networks with multiple levels of representations having more computer vision background. Convolutional networks were inspired by the visual system's structure, and in particular by the models proposed in (Hubel and Wiesel, 1959).

The first computational models based on the local connections between neurons and on hierarchically organized transformations of the image are found in Fukushima's Neocognitron (Fukushima, 1980). When neurons with the same parameters are applied on patches of the previous layer at different locations, a form of translational invariance is obtained. Later, LeCun followed-up on this idea and trained such networks using the error gradient, obtaining and maintaining state-of-the-art performances on several computer vision applications (LeCun and Bengio, 1995).



Model	Biologically inspired	Hierarchical	Spatial Information	Temporal information	Probabilistic methods	Mathematical proof	Machine learning	Benchmark validation	Invariances
HTM	✓	✓	✓	✓	✓	✓	✓	Toy	Temporal, position
DBN	✗	✓	✓	✗	✓	✓	✓	Digits	Scale, position
HMAX	✓	✓	✗	✗	✗	✗	✗	Caltech101	Scale, Position
Fidler Sanja	✓	✓	✓	✗	✓	✗	✓	Caltech101	Scale, Position
Neo-Cognitron	✓	✓	✗	✗	✗	✗	✗	Digits	Scale, Position

Figure 2.10: A comparison on the main models introduced above.

Modern understanding of the physiology of the visual system is consistent with the processing style found in convolutional networks (Serre et al., 2007a), at least for the quick recognition of objects, i.e. without the benefit of attention and top-down feedback connections. Vision systems based on convolutional neural networks have been among the best performing systems. This has been shown clearly for handwritten character recognition (LeCun and Bengio, 1995), which has served as a machine learning benchmark for many years. HMAX has some advantages such as the fact that it fits neuroscience data well, and can make a few predictions for biophysics and psychophysics. It can compete with existing Artificial Intelligence (AI) systems on categorization task. On the other hand, it also has some disadvantages. It ignores feedback effects and focuses on the first 150 ms of visual pathway; hence, it is not suitable for some complicated object recognition tasks in which feedback plays an important role. The scale information is lost due to the max operation in C_k layers. HMAX, unlike HTM, does not need long spatio-temporal information processing procedures and can be applied to real images rather than toy object videos for learning. HTM model uses the formalism of Bayesian belief propagation for inference and has the capability of using feedback propagation for

reconstruction of input images and uses temporal information as well. A comparison on different models described above, is summarized in Figure 2.10.

Convolutional neural networks and feed-forward models of the visual cortex like Neocognitron and the HMAX model have been very successful on visual pattern recognition problems. Despite the wide variety of learning algorithms employed in these models, they all share the same structural properties. All of these models have feed-forward hierarchies with alternating layers of feature selection and feature pooling levels. However the proposed models for the visual cortex are simplistic and do not provide a perfect mapping of the brain.

In conclusion, there are two approaches using hierarchical structures. One is the more biologically inspired models that attempt to model the human visual cortex such as HMAX, and the other is the computer vision and probabilistic approaches such as DBN. Since HMAX has a very strong biological base and has been proven to be very successful in image classification, we will employ this concept as a basis for our model. In Chapter 3 we will provide an exposition of our model and investigate the proposed extensions and improvements made to it.

Chapter 3

The HMAX Model and its Extensions

In this chapter we introduce the HMAX biologically inspired model and provide a review on the recent modifications and extensions made to this model. We conclude this chapter by providing justifications and biological inspirations for the modifications and enhancements we have proposed in this thesis, followed by a review on works related to these modifications.

3.1 HMAX Model

The HMAX model (Riesenhuber and Poggio, 1999) simulates the feed-forward path of the visual cortex. This model is used to find a good trade-off between invariance and selectivity. $S1$ cells provide selectivity by responding to oriented filters and $C1$ cells provide invariance by pooling over neighboring scales and positions, as can be seen in Figure 3.1. There were a number of extensions made to this model in recent years. We first

introduce the basic model, and then discuss its extensions.

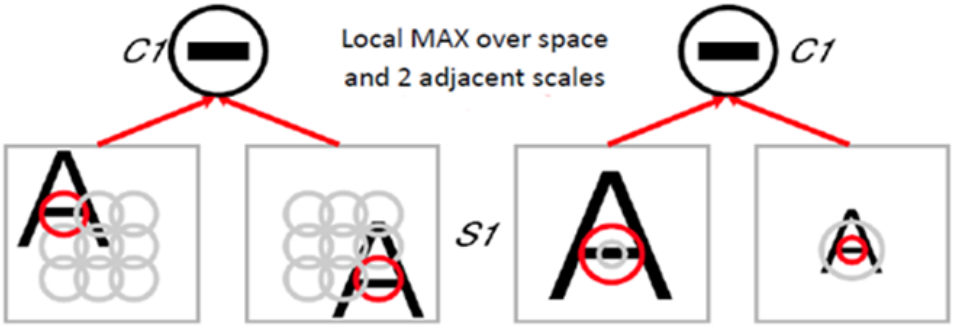


Figure 3.1: Invariance to scale and position in $C1$ layer (Serre and Riesenhuber, 2004).

Event-Related Potential (ERP) data has shown that the process of object recognition appears to take remarkably little time on the order of the latency of the ventral visual stream. This adds to earlier psychophysical studies using a rapid serial visual presentation paradigm (RSVP) which have found that subjects were still able to process images when they were presented as rapidly as 8 images per second. In summary, the accumulated evidence points to six mostly accepted properties of the ventral stream architecture:

- A hierarchical build-up of invariance, first to position and scale and then to viewpoint and more complex transformations requiring the interpolation between several different object views;
- In parallel, an increasing size of the receptive fields;
- An increasing complexity of the optimal stimuli for the neurons;

- A basic feed-forward processing of information (for immediate recognition tasks);
- Plasticity and learning probably at all stages and certainly at the level of IT; and
- Learning specific to an individual object is not required for scale and position invariance (over a restricted range).

These basic facts lead to a standard model, likely to represent the simplest class of models reflecting the known anatomical and biological constraints. It represents in its basic architecture the average belief - often implicit - of many visual physiologists.

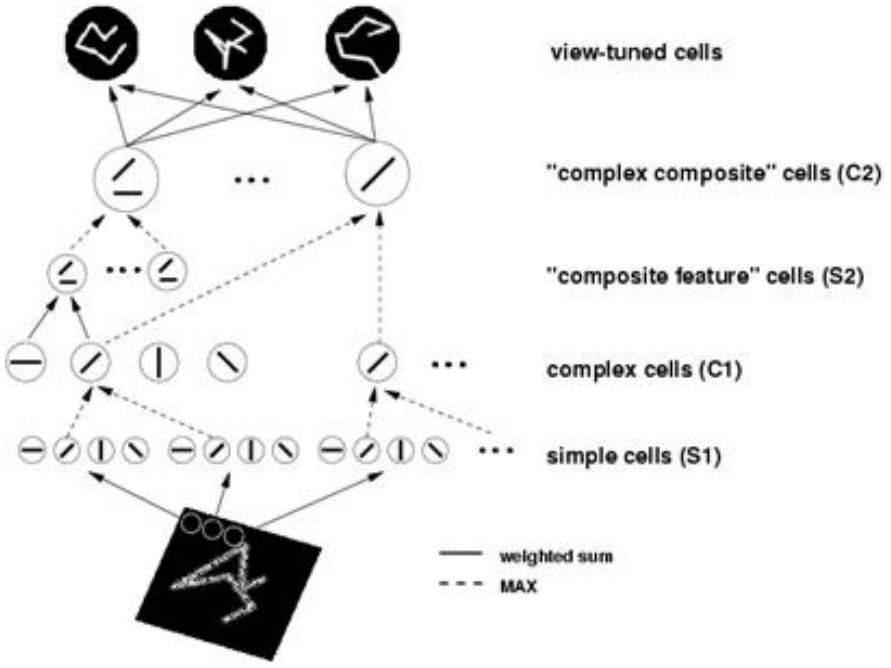


Figure 3.2: The standard HMAX model (Riesenhuber and Poggio, 1999) .

The model reflects the general organization of visual cortex in a series of layers from $V1$ to IT to PFC (as can be seen in Figure 3.2). From the

viewpoint of invariance properties, it consists of a sequence of two main modules based on two key ideas. The first module, shown schematically above, leads to model units showing the same scale and position invariance properties as the view-tuned IT neurons of (Logothetis et al., 2001) using the same stimuli. This is not an independent prediction since the model parameters were chosen to fit Logothetis' data (Logothetis et al., 2001). It is, however, not obvious that a hierarchical architecture using plausible neural mechanisms could account for the measured invariance and selectivity. Computationally, this is accomplished by a scheme that can be best explained by taking striate complex cells as an example: invariance to changes in the position of an optimal stimulus (within a range) is obtained in the model by means of a maximum operation (max) performed on the simple cell inputs to the complex cells, where the strongest input determines the cell's output. Simple cells afferent to a complex cell are assumed to have the same preferred orientation with their receptive fields located at different positions.

Taking the maximum over the simple cell afferent inputs provides position invariance while preserving feature specificity. The key idea is that the step of filtering followed by a max operation is equivalent to a powerful signal processing technique: select the peak of the correlation between the signal and a given matched filter, where the correlation is either over position or scale. The model alternates layers of units combining simple filters into more complex ones in order to increase pattern selectivity with layers based on the max operation and also to build invariance to position and

scale while preserving pattern selectivity (Serre and Riesenhuber, 2004).

In the second part of the architecture, (Figure 3.2), learning from multiple examples, i.e. different view-tuned neurons, leads to view-invariant units as well as to neural circuits performing specific tasks. The key idea here is that interpolation and generalization can be obtained by simple networks, similar to Gaussian Radial Basis Function (GRBF) networks (Riesenhuber and Poggio, 1999) that learn from a set of examples, that is, input-output pairs. In this case, inputs are views and the outputs are the parameters of interest such as the label of the object or its pose or expression (in the case of a face). The GRBF network has a hidden unit for each example view, broadly tuned to the features of an example image.

The weights from the hidden units to the output are learned from the set of examples, that is input-output pairs. In principle, two networks sharing the same hidden units but with different weights (from the hidden units to the output unit), could be trained to perform different tasks such as pose estimation or view-invariant recognition. Depending just on the set of training examples, learning networks of this type can learn to categorize across exemplars of a class as well as to identify an object across different illuminations and different viewpoints (Riesenhuber and Poggio, 2000). The demonstration that a view-based GRBF model could achieve view-invariant object recognition in fact motivated psychophysical experiments (Edelman, 1991). In turn, the psychophysics provided strong support for the view-based hypothesis against alternative theories (for a review, see (Tarr and ulthoff, 1998) and, together with the model, triggered the phys-

iological work of (Logothetis et al., 1995)). Thus, the two key ideas in the model are: (1) The max operation provides invariance at several steps of the hierarchy; and (2) The RBF-like learning network learns a specific task based on a set of cells tuned to example views. More details on how tuning properties are adjusted, in particular invariance ranges in HMAX, depend on pooling parameters.

In the standard HMAX model, there are four layers of hierarchy to create the features for the classifier and there is a supervised classifier on top. The overall framework is described as follows:

S1 Layer: Input images are densely sampled by arrays of two-dimensional Gaussian filters, the so-called $S1$ units (second derivative of Gaussian, of four different orientations and 17 different scales); sensitive to bars of different orientations, thus roughly resembling properties of simple cells in striate cortex. At each pixel of the input image, filters of each size and orientation are centered. The filters are sum-normalized to zero and square-normalized to 1, and the result of the convolution of an image patch with a filter is divided by the power (sum of squares) of the image patch.

C1 Layer: In the next step, filter bands are defined, i.e. groups of $S1$ filters of a certain size range. Within each filter band, a pooling range is defined which determines the size of the array of neighboring $S1$ units of all sizes in that filter band that feed into a $C1$ unit (roughly corresponding to complex cells of striate cortex). Only $S1$ filters with the same preferred orientation feed into a given $C1$ unit to preserve feature specificity. The pooling operation that the $C1$ units use is the max operation, i.e. a $C1$

unit's activity is determined by the strongest input it receives.

S2 Layer: Within each filter band, a square of four adjacent, non-overlapping $C1$ units is then grouped to provide input to a $S2$ unit. There are 256 different types of $S2$ units in each filter band, corresponding to different possible arrangements of four $C1$ units of each of four types (i.e. preferred bar orientation). The $S2$ unit response function is a Gaussian function with mean 0 and standard deviation 1, i.e. a $S2$ unit has a maximal firing rate of 1 which is attained if each of its four afferent fires at a rate of 1 as well. $S2$ units provide the feature dictionary of HMAX, in this case all combinations of 2×2 arrangements of bars (more precisely, $C1$ cells) at four possible orientations.

C2 Layer: To finally achieve size invariance over all filter sizes in the four filter bands and position invariance over the whole visual field, the $S2$ units are again pooled by a max operation to yield $C2$ units, the output units of the HMAX core system, designed to correspond to neurons in extrastriate visual area $V4$ or posterior IT (PIT). There are 256 $C2$ units, each of which pools over all $S2$ units of one type at all positions and scales. Consequently, a $C2$ unit will fire at the same rate as the most active $S2$ unit that is selective for the same combination of four bars, but regardless of its scale or position.

VTU Layer: $C2$ units then again provide input to the view-tuned units (VTUs), named after their property of responding well to a certain two-dimensional view of a three-dimensional object, thereby closely resembling the view-tuned cells found in monkey inferotemporal cortex by Logothetis

et al (Logothetis et al., 1995). The $C2$ to VTU connections are so far the only stage of the HMAX model where learning occurs. A VTU is tuned to a stimulus by selecting the activities of the $C2$ units in response to that stimulus as the center of a 256-dimensional Gaussian response function, yielding a maximal response of 1 for a VTU in case the $C2$ activation pattern exactly matches the $C2$ activation pattern evoked by the training stimulus. To achieve greater robustness in case of cluttered stimulus displays, only those $C2$ units may be selected as afferent for a VTU that responds most strongly to the training stimulus. An additional parameter specifying response properties of a VTU is its σ value, or the standard deviation of its Gaussian response function. A smaller σ value yields more specific tuning since the resultant Gaussian has a narrower half-maximum width.

3.2 Extensions to the Standard HMAX Model

Several modifications and improvement have been proposed to improve the standard HMAX model. In the rest of this section we introduce the existing known modifications to HMAX model, as follows:

Serre et al.

Serre and Reisenhuber modified the standard HMAX structure, and released a new version of this structure (Serre and Riesenhuber, 2004). Gabor filters were used instead of second order Gaussian derivatives in $S1$ layer and the number of filter sizes was increased. They also changed the

values of scale range and pool range parameters in standard HMAX in $C1$ layer to provide less scale tolerance and therefore narrower spatial frequency bandwidth (Serre and Riesenhuber, 2004). Two other layers was added to this standard model. This model includes $S2b$, $S3$, $C2b$, $C3$, and $S4$ as shown in Figure 3.3 (where b indicates bypass layers). This hierarchical structure is repeated once more to create $S3$ and $C3$ layers, and in another implementation, $S2b$ and $C2b$ layers are also created to bypass $S2$ and $C2$ layers in creating the dictionary (Serre et al., 2007a).

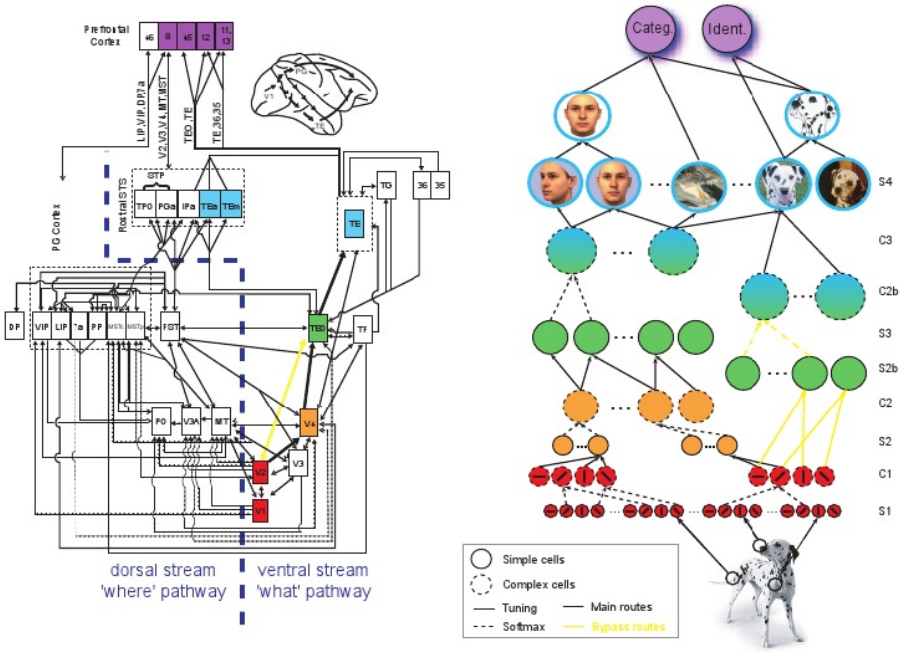


Figure 3.3: Extensions to HMAX in Serre et al. (2007a)

In the basic HMAX model, in $S1$ layer, Gaussian filters are convolved on the whole image to create inputs to the $C1$ layer. In the later models, Gabor filters were used to replace Gaussian filters because Gabor filters were shown to have better performance and to be more similar to $V1$ cells (Serre and Riesenhuber, 2004). Gabor filters have more free parameters

and can be tuned more accurately than Gaussian filters. On one hand, using Gabor filters provides better performance when the model is used to differentiate between edges, bars and gratings. As illustrated in Figure 3.4 (Serre and Riesenhuber, 2004), using Gaussian filters, we obtain consistent tuning curves for sweeping edges, bars and gratings since Gaussian filters have shorter and wider bars. On the other hand, since Gaussian derivatives have only one free parameter, it is impossible for them to provide matching of spatial frequency distribution and bandwidth (Serre and Riesenhuber, 2004).

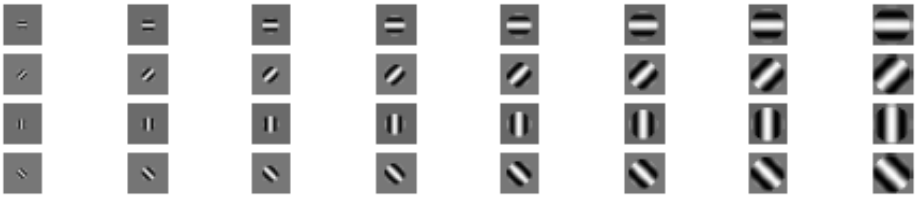


Figure 3.4: (left) Gabor and (right) Gaussian derivatives (Serre and Riesenhuber, 2004).

Different numbers of Gabor filters have been implemented on the images in different implementations. In (Serre et al., 2007a), 17 sizes of filters are used with 4 orientations and 2 phases which sums up to 136 types of units, as shown in Figure 3.5.

Beyond $C2$ the units are increasingly complex and invariance. $S3/C3$ units are combination of $V4$ like units with different selectivity levels. They are like a dictionary of 1000 features, which according to (Fujita et al., 1992), is equivalent to the number of columns in IT. $S4$ units are view-tuned units similar to the standard HMAX but with the difference that

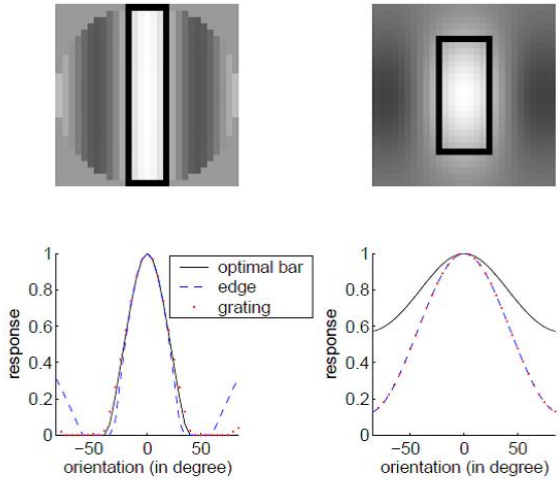


Figure 3.5: Receptive filed organization of the $S1$ units (only units at one phase are shown (left: Gabor, right: Guassian) (Serre and Riesenhuber, 2004).

they are not supervised. Their tuning and invariance properties agree with IT data (Riesenhuber and Poggio, 2000; Logothetis et al., 1995). Using this approach, performance of 35% with 15 training images and 42% with 30 training images has been achieved on Caltech101 test dataset (Li et al., 2004).

Mutch et al.

Mutch et al. (Mutch and Lowe, 2008; Mutch et al., 2010a) proposed a series of computational modifications to the structure proposed by Serre et al.'s model (See Figure 3.6).

In this model, a fixed size of Gabor filters is implemented on different scales of the images which provides the same invariance to scale for Gabor filters (Mutch and Lowe, 2008, 2006). In this model, an image is fed into the structure and 10 different scales of the image are created as input to $S1$ layer. Gabor filters in 4 directions in their standard model, and 12

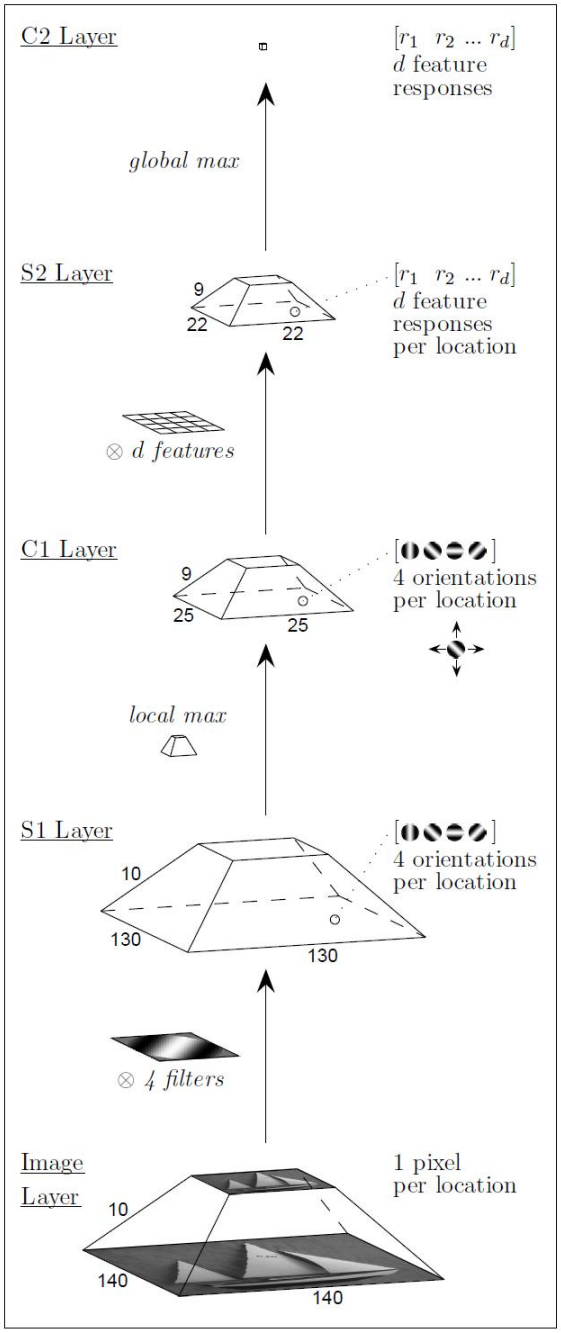


Figure 3.6: Modified HMAX model in (Mutch and Lowe, 2008).

directions in their extended model, are created based on Equation 3.1 and convolved on the images:

$$G(x, y) = \exp\left(-\frac{(X^2 + \gamma^2 Y^2)}{2\sigma^2}\right) \cos\left(\frac{2\pi}{\lambda} X\right) \tag{3.1}$$

These outputs are sent to $C1$ layer, which performs a local max operation on both size and position of the filter responses. The response of a patch of pixels X to a particular $S1$ filter G is given by:

$$R(x, y) = \left| \frac{\sum X_i G_i}{\sqrt{\sum X_i^2}} \right| \tag{3.2}$$

The output of this layer will be between 500-2000 different patches of size 4×4, 8×8, 12×12 and 16×16 depending on the size of the input image.

A dictionary of features is randomly sampled from these patches. One or two samples are randomly sampled from each training image, and a feature's dictionary of size 4096 of prototypes is created. This dictionary is then made sparse by selecting the highest response from each orientation and setting the rest to 0, as portrayed in Figure 3.7.

The response of a patch of $C1$ units X to a particular $S2$ feature/prototype P, of size $n \times n$, is given by a Gaussian radial basis function:

$$R(x, P) = \exp\left(-\frac{\| X - P \|^2}{2\sigma^2 \alpha}\right) \tag{3.3}$$

In order to train the Support Vector Machine (SVM), they find the distance of each sample from each training image, with each entry on the

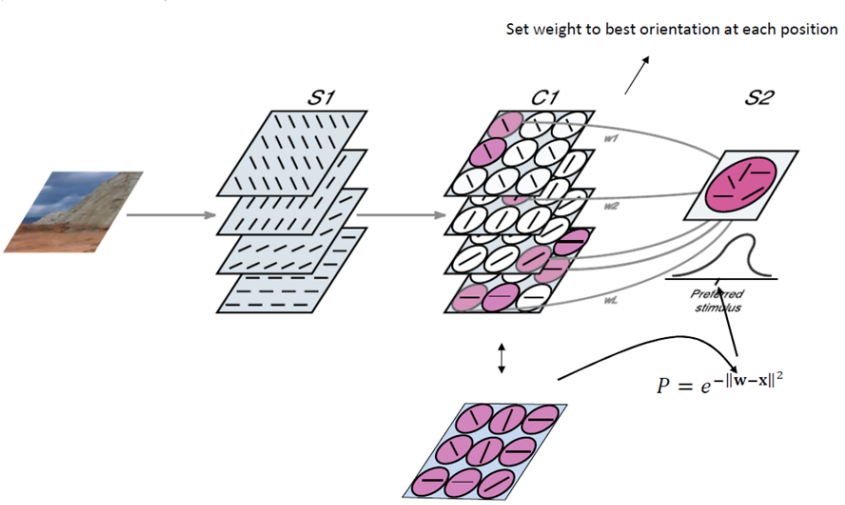


Figure 3.7: Dense and sparse features (Theriault et al., 2011).

dictionary, and find the max in $C2$ layer. These features are sent to the SVM for training.

For testing images the same hierarchical procedure is repeated and the performance of the system is calculated. They proposed a few modifications to improve the performance of the system such as running a SVM normals method (Mladenić et al., 2004) to select the features with higher weights. SVM is run a few times, and each time features with lower weights are dropped. Using this approach, performance of 51% on 15 training images and 56% on 30 training images has been achieved on Caltech101 dataset (Li et al., 2004). This model is performing similar to the standard model, meanwhile using a fewer number of layers. A GPU based implementation of hierarchical architectures is provided in (Mutch et al., 2010a) which runs about 100 times faster in creating the layers of hierarchical structures such as HMAX. We used this source code as the kernel of our project, modeled our modifications and performed the experimental simulations.

Masquelier et al.

Masquelier and Thorpe (2007) proposed unsupervised learning of features in $S2$ level of HMAX structure (see Figure 3.8). The structure they proposed starts with 4 directions of Gabor filters on 5 scales of images. The orientation with the best response is selected in $C1$ layer, and after performing spatial lateral inhibition, responses of $C1$ layer are summed and $S2$ features are created. The max is selected in $C2$ layer and another sum is performed to create higher level features to be fed to the classifier. They start with 20 prototypes initialized with random weight matrices W_i. They present one training image and compute activation of each prototype layer and the strongest output of each prototype triggers learning which is based on Hebbian unsupervised learning rule. This approach has been tested on a few classes of Caltech101 dataset(Li et al., 2004) and the features created contain object specific characteristics and look like the gist of the image.

Theriault et al.

Theriault et al. (2011) suggested a few modifications to the HMAX model and proposed S-HMAX model by selecting their sparse features from neighboring scales rather than one random scale chosen in (Mutch and Lowe, 2008) as illustrated in Figure 3.9.

They reported classification accuracy of 59% on Caltech101 dataset using multiple scale features and reported that the use of normalized dot product (instead of Gaussian radial basis function) at $S2$ layer (Equation 5.3) increases performance by another 2%. However, in our experiments,

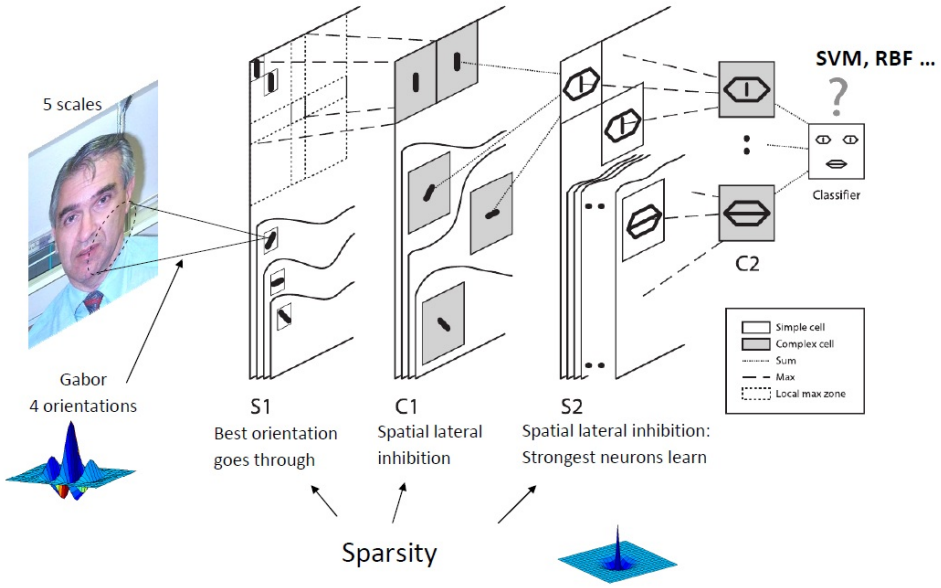


Figure 3.8: Unsupervised learning of S2 prototypes (Masquelier and Thorpe, 2007).

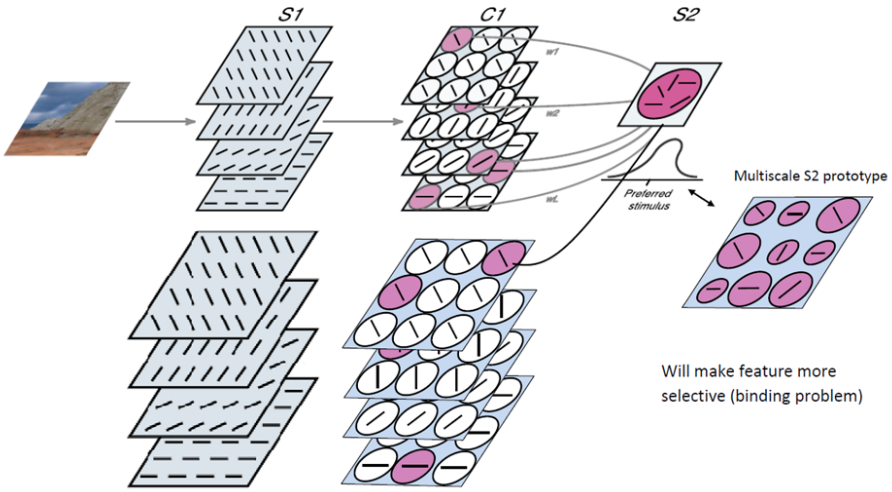


Figure 3.9: Multiple-scale sparse features (Theriault et al., 2011).

this replacement had resulted in 6% lower performance.

Use of Color in HMAX Model

Recently, more sophisticated modeling of single-opponent and double-opponent cells in $V1$ has shown that adding more biological realism to color descriptors can significantly improve object and scene categorization performance. Nonetheless, this improvement is attained with a relatively low-level color machinery. Zhang, Barhomi and Serre (Zhang et al., 2012) proposed a new biologically inspired color descriptor that encodes color information in a low-level manner. In their model, they create 8 channels of opponent colors: R^+G^-, R^-G^+, R^+C^-, R^-C^+, Y^+B^-, Y^-B^+, Wh, Bl and used these channels to calculate the Gabor filters on different orientations and used them to create Single-Opponent and Double-Opponent channels. In order to evaluate the combination of the SODO-HMAX model of (Zhang et al., 2012) with our proposed color model 6, which is a high-level color model, we concatenated their SODO-HMAX features with our CQ-HMAX features. In SODO-HMAX, Single-Opponent features encode color regions and Double-Opponent features encode color edges.

3.3 Discussions and Proposed Modifications

We present several modifications to the HMAX model in the following chapters and propose a new biologically inspired high-level color model. We provide the motivations for our modifications in three main aspects and conclude this chapter by providing possible applications for HMAX

model.

3.3.1 Visual Dictionary of Features in HMAX Model

As described in Section 3.1, in order to create the dictionary of features in HMAX main model (Riesenhuber and Poggio, 1999), a handmade dictionary of features in $S2$ level is used. This set was later replaced with a random selection of patches of different sizes selected from $C1$ layer (Serre and Riesenhuber, 2004). In another extension, these features were made sparse (Mutch and Lowe, 2008), and multiple-scale sparse (Theriault et al., 2011). Inspired by the better classification accuracy achieved by replacing the Bayesian network with a Growing Self Organizing Map (GSOM) for clustering in the inference phase in HTM model, we proposed several clustering methods and investigated the use of saliency regions for sampling the features for creation of the dictionary of features in Chapter 4. In this chapter, we also provided a method to use the most frequent clusters in different regions of the images of each class as candidates features.

3.3.2 Encoding Occurrences and Co-Occurrences of Features in HMAX Model

In Chapter 5 we introduce several methods for pooling at $C2$ layer and propose the HMean model in which a mean operator replaces the max operator proposed. We also investigate the classification accuracy achieved by concatenating these pooling methods. We further provide another extension to the HMAX model by adding a higher level dictionary that is created

using co-occurrences of the features in the lower level dictionary of features. We propose several methods for encoding occurrences and co-occurrences of features and provide the biological inspirations for these modifications.

3.3.3 Color Processing in HMAX Model

In Chapter 6, a new biologically inspired model (CQ-HMAX) is introduced in which a structure similar to HMAX is implemented for encoding color information of the images. We provide biological inspirations for our model and show that the classification results achieved by this model are among the best in the bottom-up approaches on several benchmark color datasets and show that the concatenation of our model with that in (Zhang et al., 2012) performs better than the state-of-the-art performances.

3.3.4 Applications of HMAX Model

We have used HMAX and CQ-HMAX models for classification of images extracted from histopathological images in order to detect mitosis. Use of HMean-HMAX and CQ-HMAX resulted in better classification accuracy than other computational methods such as SIFT on benthic marine mammals dataset. We also propose a new HMAX-like structure for encoding images captured from SONAR and use it for underwater object recognition where the quality of images is not high and the use of sonar information along with images results in better accuracy.

Chapter 4

Enhancements to the Visual Dictionary in HMAX Model

4.1 Introduction

Since Riesenhuber and Poggio (1999) proposed the HMAX model, a series of models were proposed to provide modifications to it in order to make it more suitable for real image classification tasks (Serre and Riesenhuber, 2004; Serre et al., 2007a; Masquelier and Thorpe, 2007; Fidler et al., 2008; Mutch and Lowe, 2008; Theriault et al., 2011). A review on these models and the differences among them were provided in Chapter 3.

In the original HMAX model (Riesenhuber and Poggio, 1999), a hand-

[0]The models and experiments in this chapter are partially presented in International Conference on Neural Information processing (ICONIP2010) and published in Proc. of Springer Neural Information Processing, Models and Applications (Jalali et al., 2010) and partially published in the Proceedings of IEEE International Joint Conference on Neural Networks 2012 (IJCNN2012) (Jalali et al., 2012).

made dictionary of features is used in $S2$ layer. In (Serre et al., 2007a) a dictionary of features of size 4075 is randomly sampled from training images from $C1$ layer which computes a local max operation on different scales and orientations of Gaussian filter responses. Mutch et al. (Mutch and Lowe, 2008) performed random sampling on Gabor filters of different orientations with the same size, on different scales of images and made features sparse. In (Masquelier and Thorpe, 2007), a Hebbian learning rule was provided to update and learn features.

The use of non-random sampling to create the dictionary of features for the model, was a prospective investigation, that motivated us to compare the performance of non-random sampling methods with random sampling.

In this chapter, we introduce several Self Organizing Maps (SOM) and K-means clustering methods for the creation of the dictionary of features as well as the use of saliency points in the creation of this dictionary. We use the same model provided in Mutch and Lowe (2006, 2008) that was described in Section 3.2 and use the Graphical Processing Unit (GPU) based codes provided in Mutch et al. (2010b) as a part of the Cortical Network Simulator (CNS) package that form the basis for our experiments and comparison.

We investigate different Self Organizing Maps (SOM) and clustering methods and propose clustering as a means of reducing the size of the dictionary of features in HMAX Model in Section 4.2. We introduce our implementation of the model, provide the experimental results of our modifications to the dictionary of features created by the model, and discuss

the modifications followed by conclusions in Section 4.3.

4.2 Proposed Methods for Creation of the Visual Dictionary

In order to investigate the role of features selected in the dictionary of features in this structure, we performed a series of experiments on Caltech101 dataset (Li et al., 2004), which includes 101 classes of objects plus a background category. Each class contains between 31 to 800 color images that have a size of about 300×200 pixels. A total of 30 randomly chosen images from each class are used for training and the rest of the images are used in the test phase.

Samples are selected from $C1$ pyramid of each image in different positions and scales using a random generator function with a Gaussian distribution in (Mutch and Lowe, 2008) based on the number of images per class and by taking a different number of samples from each image. We performed different non-random sampling methods and compared their performances using an extensive set of experiments.

Since learning in $S1$ and $S2$ layers is not purely genetic, and in an experiment by Kohonen (1982), it is shown that cats that have been kept in an environment with only horizontal lines, did not develop sensitivity to vertical lines, and since top-down supervision is absent or very weak in $V1$ layer, there shall be a self organizing structure in this layer to learn the statistics of the input data. SOM has been proposed in Barlow et al.

(1975) to resemble this biological inspirations. On the other hand, self organizing maps with a small number of nodes behave similar to K-means and larger self-organizing maps rearrange data points while preserving their topological structure in a lower-dimension manifold. Hence the main difference between SOM and K-means clustering is in the neighborhood update which is absent in K-means clustering.

A self organizing map produces a low-dimensional discretized representation of the input space of the training samples Kohonen (1982). In a self organizing map, after the random initialization of the values in the map, the closest value to each entry from the input space to each neuron is found, and the neighborhood of that particular winning neuron is updated according to their distance to the winning map feature.

A self organizing map for an n-dimensional input space and m output neurons includes the following steps:

1. Randomly initialize the weight vector w_i for neuron i, i = 1,...,m.

2. Sampling: choose an input vector x from the training set.

3. Determine winner neuron k:

$\| w_k - x \| = min_i \| w_i - x \|$ (Euclidean distance)

4. Update all weight vectors of all neurons i in the neighborhood of the winning neuron $i(x)$:

$w_j(n+1) = x_j(n) + \eta(n) h_{j,i(x)}(n)(x - w_j(n))$, j = 1,...,M

5. If convergence criterion met, STOP. Otherwise, go to Step 2.

$h_{j,i(x)} = \exp(-\frac{d_{j,i}^2}{2\sigma^2})$ where $d_{j,i}$ is the Euclidean distance from neuron j to the winning neuron i

$\sigma(n) = \sigma_0 \exp(-\frac{n}{\tau_1})$

$\eta(n) = \eta_0 \exp(-\frac{n}{\tau_2})$

$\eta_0 = 0.1$, $\eta(n) = 0.1 \exp(-\frac{n}{T})$, n = 0,1,2,...

where T is the total number of iterations.

We perform K-means clustering after sampling more samples from $C1$ pyramid of each image in different approaches. Different number of samples and different number of clusters are tested in a series of experiments. Whenever an empty cluster is created in the batch update phase, we create a new cluster consisting of the one point furthest from its centroid. Squared Euclidean distance is chosen as the distance measure, so that each centroid is the mean of the points in the corresponding cluster. We use the K-means function in Statistical toolbox of Matlab® for clustering. We compared our results with those from Mutch and Lowe (2008) which have reported 54% classification accuracy before the use of weighted SVM.

Let X be a set of features sampled from images at the $C1$ layer in a $k - dimensional$ feature space, i.e. $X = [x_1, x_2, ..., x_M]^T \in \Re^{M \times K}$.

K-means clustering is used to solve

$$\min_{D} \sum_{m=1}^{M} \min_{n=1...N} \parallel x_m - d_n \parallel^2 \tag{4.1}$$

where $D = [d_1, d_2, ..., d_N]^T$ are the cluster centers to be determined and $||.||$ denotes the $l2 - norm$.

4.2.1 SOM and Clustering over Images from All Classes

In the first approach, we sampled between 5 to 20 random patches from all of the images to achieve a more dense sampling and added these samples

to the dictionary of features, resulting in a very big dictionary of features of size 15000 to 60000. We then performed SOM and clustering over the whole dictionary and created a dictionary of size 1000 to 9000. This method is illustrated in Figure 4.1. The results of this experiment are shown in Table 4.1.

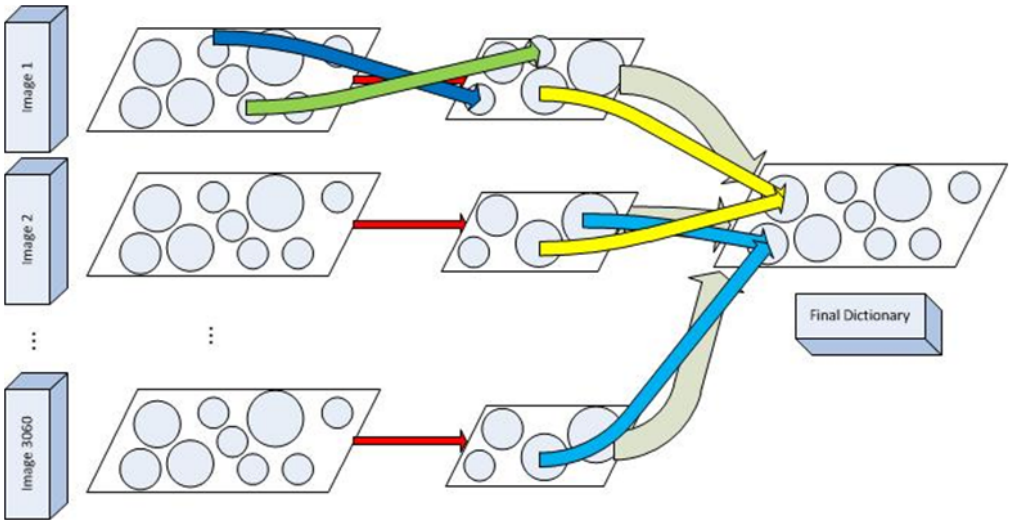


Figure 4.1: Sampling over all images and performing clustering over all samples to create the dictionary of features.

The value of M in Equation 4.1 is all the possible patches on all images of all classes. If we have NP possible patches on each image and we have NI number of images in each class and NC classes, then $M = NP \times NI \times NC$.

4.2.2 SOM and Clustering over Images Individually

In another approach, we sampled all of the possible positions of $C1$ features for each image; this resulted in a number of samples between 500 to 2000 for each image (depending on the image size). In this approach, all of the patches of different sizes are extracted non-randomly from $C1$ layer with

a step size of 1 without overlapping. Clustering is then performed on each image, and between 3 to 10 clusters per image are added to the dictionary which results in a dictionary size of 9180 to 30600. The results of this experiment are shown in Table 4.1. Furthermore we performed sampling on more features from each image in another set of experiments, and performed a second clustering on the whole dictionary to reduce the number of features to 4075. We sampled 10 clusters per image, and generated a dictionary of size 30600. We performed another clustering to reduce the size to 4075 and the performance is almost the same with sampling less from each image, and creating a dictionary of size 4075. The method for this model is illustrated in Figure 4.2. The same experiments were carried out using SOM.

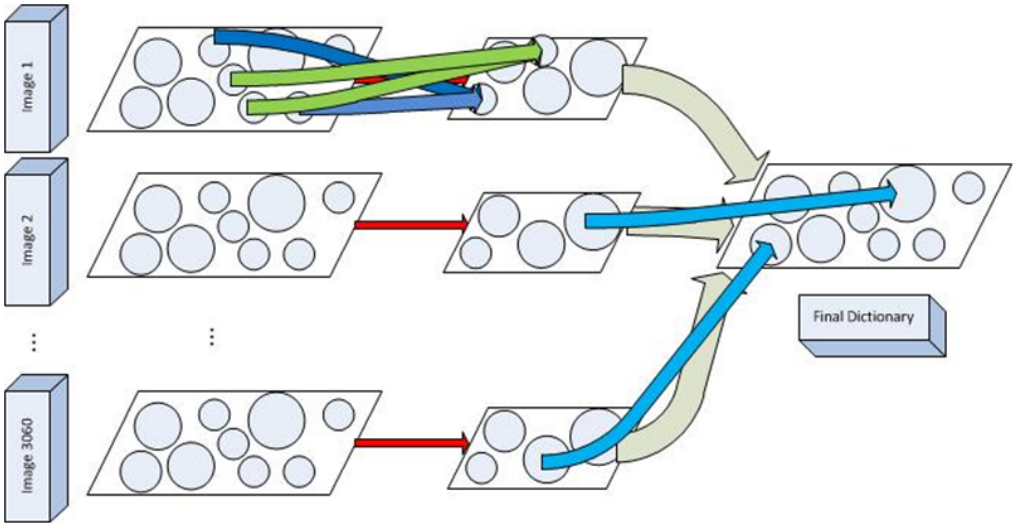


Figure 4.2: Sampling over one single image and performing clustering at image level to create a dictionary of features.

The value of M in Equation 4.1 is all the possible patches on all images of each class. If we have NP possible patches on each image, NI number of images in each class, and NC classes, then $M = NP$ for each image.

The clustering is done individually for each image ($NI \times NC$ times). Then the clusters of each image are added to a final dictionary.

4.2.3 SOM and Clustering over Images in Each Class

In the third approach, we performed sampling on images of each class separately. Different numbers of samples were chosen for each image, both randomly and non-randomly. In the case of non-random sampling, a regular scanning over all possible sizes and positions was conducted using a step size depending on the ratio of number of possible positions for sampling over number of desired samples from each image. We sampled different numbers of samples from each image and created 3 different categories. Small number of samples standing for 100-300 samples in each image, 300-600 samples per image for medium number of sampling, and 600-2000 for large number of sampling. For all the methods, the ratio of the number of clusters over each class to the number of samples per image, was kept to be approximately 0.1 since this showed the best performance among different ratios of 0.2, 0.3 and 0.5. The same experiments were carried out using SOM. The method for these experiments is shown in Figure 4.3.

The value of M in Equation 4.1 is all the possible patches on all images of each class. If we have NP possible patches on each image, NI number of images in each class, and NC classes, then $M = NP \times NI$ for each class. Then the clusters of each class are added to a final dictionary.

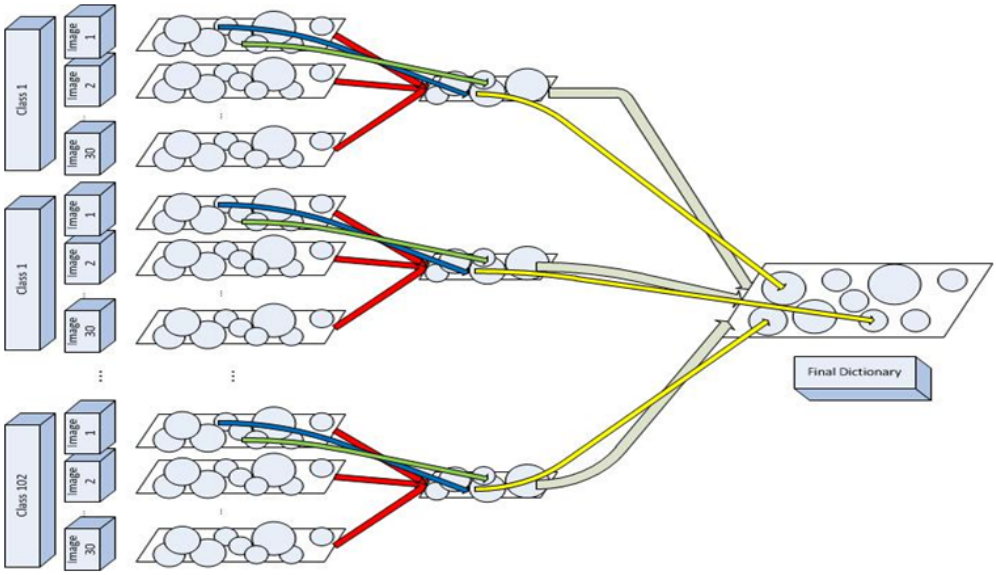


Figure 4.3: Clustering on samples from the center quarter of the images from each category to create a dictionary of features.

4.2.4 Sampling over Center of Images

Since most of the images in Caltech101 are focused on the center, we investigated another approach where we performed the sampling around the center of images with the objective of capturing more meaningful information related to objects of interest and less information about the image background. We tried two different methods for patch selection. In the first method, we created the dictionary of features from the center of images from all of the images of each class, and in the training and testing phases, use the full size images for calculating the $S2$ layer. The method for this approach is shown in Figure 4.4.

In the second method, we sampled from the center quarter of each image and created a dictionary of features from these patches. The method for

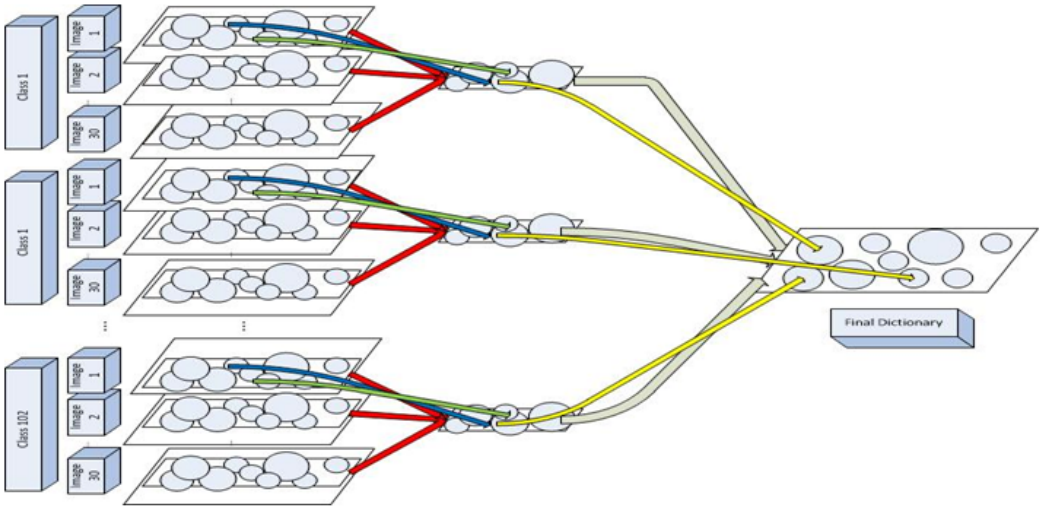


Figure 4.4: Creating the dictionary of features from the center of images rather than the whole image to create a dictionary of features.

this approach is shown in Figure 4.5. This approach, however, did not outperform random sampling but we could achieve better performances than random sampling using a dictionary of features that is smaller in size.

The value of M in Equation 4.1 is equal to the number of all of the possible patches on the center of all images of each class. Consider the example in which we have NP possible patches on each image (NP in this method is smaller since we are only sampling from the center of the images), NI number of images in each class, and NC classes, in the first approach, we have $M = NP \times NI$ for each class; hence, the clustering is performed $NP \times NC$ times and the clusters of each class are added to a final dictionary. In the second approach, we have $M = NP$ for each image. The clustering is done individually for each image for $NI \times NC$ times.

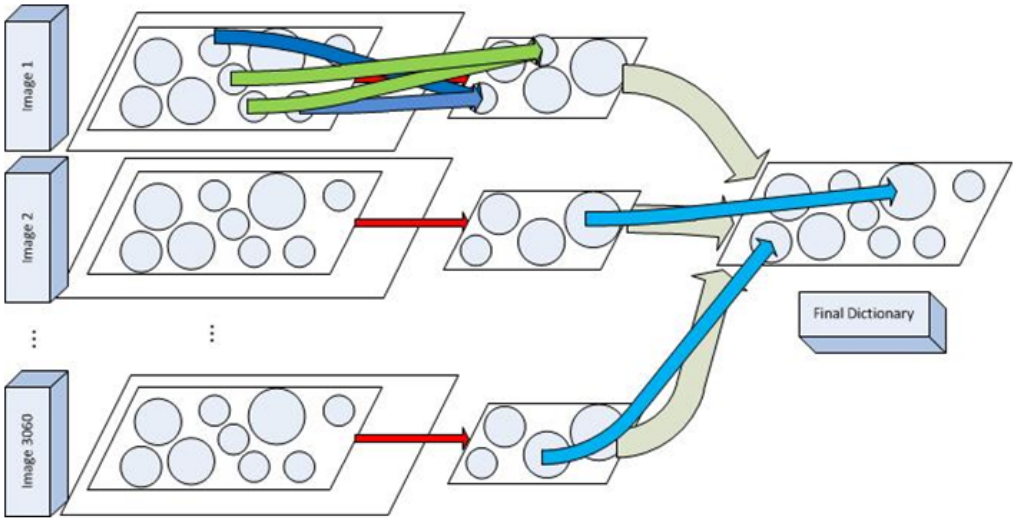


Figure 4.5: Clustering on samples from the center quarter of all of the images to create a dictionary of features.

4.2.5 Sampling over Saliency Points

In an extension to these experiments, we also performed sampling on saliency regions in the images to investigate the performance of sampling on these points. Walther (2006) has proposed a combination of Saliency map with HMAX model. In this model, which is based on the saliency map proposed in (Itti et al., 1998), a bottom-up salient region selection is provided, and the points with higher saliency response are selected as prospective points for sampling in the HMAX model, as depicted in Figure 4.6.

A saliency map is calculated for each image using three different features, namely, Red-Green and Blue-Yellow color conspicuity map, intensity conspicuity map, and orientation conspicuity map. After their combination, a saliency map is created and those points with higher response are selected with their neighborhood for sampling. They sample from each

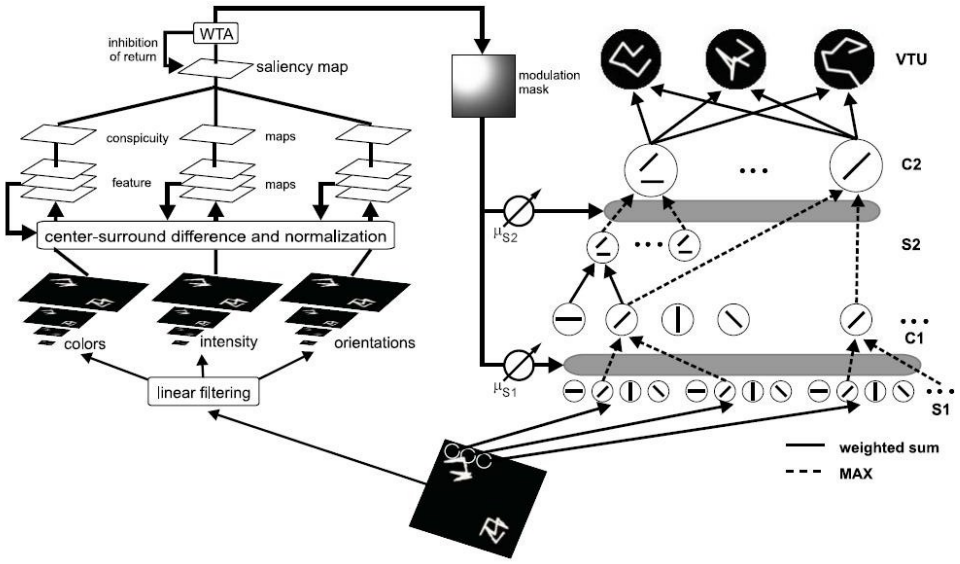


Figure 4.6: Combined model of bottom up attention and object recognition (Walther, 2006).

training class separately; these samples are used for training the SVM. A very similar approach to performing clustering within classes was discussed in Section 4.2.3. We found the top 10 most salient points in each image based on the methods described in (Walther, 2006), and selected a window size of 16×16 around them, and randomly sampled patches of size 4×4, 8×8, 12×12 and 16×16 on these windows. It was observed that the performance of the system did not improve over that of the random sampling.

4.2.6 Spatially Localized Dictionary of Features

We explore the use of frequency and spatial information in clustering features. Each image is divided in 3×3 grids, and clustering is done on the features located within each grid for each class respectively. By using grids, spatial information is encoded such that features that are clustered

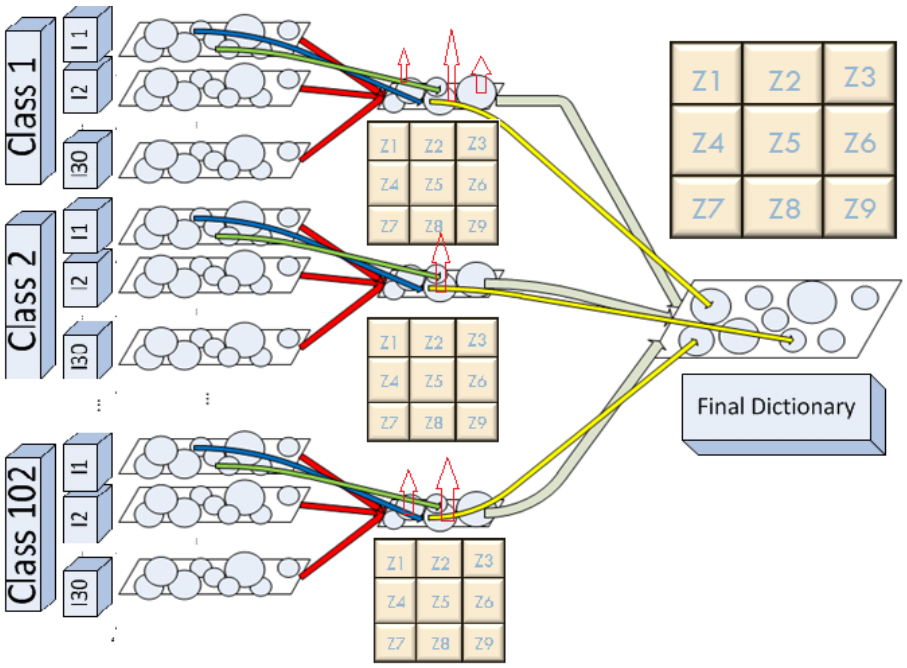


Figure 4.7: Use of zones and frequency of features in clustering inter classes using most frequent features in each zone for each class of images.

together are found in a region. Once these N clusters are created, the next step is to reduce the quantity so that the computational complexity of the classification task is reduced. In our previous model, a lower number of clusters was selected and that had resulted in blurring and degraded classification performance. In this method, a higher number of clusters is selected and a term frequency approach (borrowed from statistical text analysis) is adopted to pick out the features of higher frequency as representatives for every specific region in each class respectively, as shown in Figure 4.7. A small subset of created clusters is chosen for each region R in every category, and the clusters with the most patches are selected and added to the final dictionary of features, D.

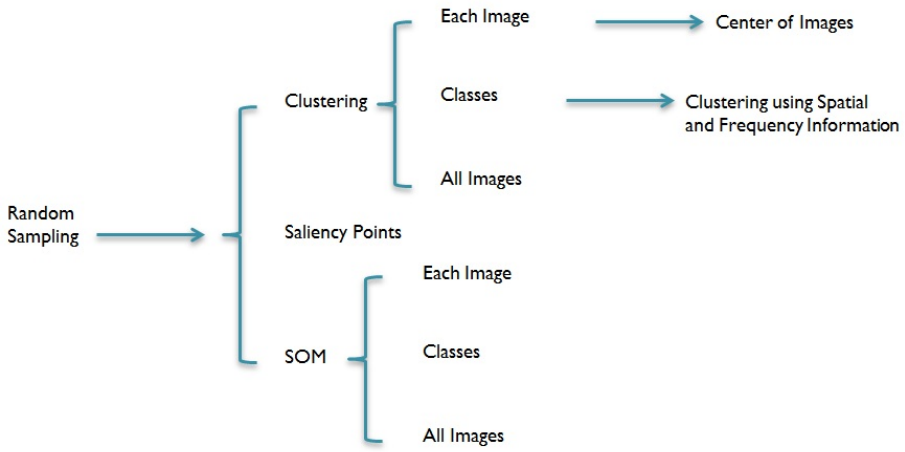


Figure 4.8: Different methods for creation of the dictionary of features.

In this experiment, we have eliminated the $S2$ inhibition in (Mutch and Lowe, 2008) and calculated the max in $C2$ layer in a ± 1 scale neighborhood and $\pm 10\%$ position neighborhood for patches on the borders, a ± 2 scale neighborhood and $\pm 20\%$ position neighborhood is considered. We have also eliminated patches that have invalid values (i.e. those that are partially sampled from the border areas as well as some parts of the patch that do not have a value in the image). Figure 4.8 illustrates different feature learning methods carried out.

Blurring Images

Inspired from the work in (Lowe, 1999) on Scale Invariant Feature Transform (SIFT) based methods, we blurred the images before processing in another experiment to investigate if this step helps with HMAX model. However, the use of Guassian low-pass filter did not result in any significant improvement in classification results.

Eliminating the SVM Layer

In another experiment, we tried skipping the higher levels of the hierarchy to evaluate the performance of $S2$ dictionary in a K nearest neighbors (K-NN) method. We created a dictionary of features as described above (sampling over each class of images) and labeled each sample according to its class of images. In the test phase, we sampled 10-90 random samples from each image, and found the K-NN matches (K varies between 10 to 100) with the existing dictionary of features, and assigned the image to the class based on a majority voting of the minimum distance with dictionary prototypes. In the classification phase, K is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label that is the most frequent among the K training samples nearest to that query point. Different approaches were taken here such as sampling randomly, non-randomly, sampling more features from each image, and performing a clustering afterwards, but none of them resulted in a performance better than 10 percent, which is very low in comparison with the 52% performance we had achieved with the support vector classifier.

4.3 Discussions

As can be seen in Table 4.1, the best performance is achieved when clustering is done using frequency of clusters in their respective spatial distribution. Several other clustering methods did not result in any significant improvement despite higher computational costs of the model.

Method	**Performance** (%)
Random sampling	52.35 % ±1.2
Clustering on all images	48.62 % ±2.1
SOM on all images	37.71 % ±3.3
Clustering on each image	52.69 % ±0.8
SOM on each image	42.13 % ±2.7
Clustering on each class	51.22 % ±1.1
SOM on each class	41.87 % ±2.3
Clustering on center of images	50.04 % ±3.7
Random sampling on center of images	42.04 % ±2.1
Sampling over saliency points	52.18 % ±0.9
Clustering using frequency and spatial info	**60.12** % ±2.2

Table 4.1: Comparison between random and non-random sampling methods for creation of the dictionary of features in Caltech101 dataset classification task using 30 training images per category.

During the previous experiments, we arrive at the conclusion that with sampling only from the center of the images, a better performance is achieved, in comparison with random sampling, when a smaller number of features is sampled in order to create the dictionary of features.

The averaging and blurring effects of clustering can be the reason of equal performance with random sampling when spatial and frequency information are ignored. The reason that the results achieved using SOM are significantly lower than clustering may be because SOM performs an update in the neighborhood of the winning neuron. Hence a method with less blurring effect in creation of the dictionary of features seems to provide a better classification accuracy. The clusters created from images are both from background and objects, and clustering features of the center of the images, where most features are from objects rather than background, resulted in better performance in comparison with random sampling on the same dictionary size. Mutch and Lowe (2008) had shown that using using a local max in $C2$ layer increases the classification performance in Caltech101 dataset by 5% which confirms the significance of spatial information in this specific dataset.

In another experiment, we used only images of one class, and created a dictionary from those images, and used this dictionary to classify the whole dataset, and we achieved similar classification performances. This experiment was repeated over other randomly chosen classes from Caltech101 dataset such as airplanes, accordion and bonsai and the results were simi-

lar. One reason for this may be the huge size of sampling region. When the size of sampling space is huge, random sampling is shown to be among the best sampling methods in statistics. However, as can be seen in 4.1, when the clustering is limited to the windows in the images and the clusters with the highest frequencies are selected as representatives, the blurring effect is reduced and the performance is increased significantly.

In (Rutishauser et al., 2004), interest points achieved from saliency maps are used for classifying images after creation of dictionary of features, and an improvement has been reported in the performance. One prospective extension to the experiments in this chapter is to use wavelet transform to find points with higher information as interest points which are also used in compressive sampling methods.

Chapter 5

Encoding Occurrences and Co-occurrences of Features in HMAX Model

5.1 Introduction

There is evidence that "Max" spatial pooling is present at multiple levels in the visual system. Importantly, however, these studies cannot be (and have not been) interpreted as evidence for only Max pooling taking place. Each of these studies also showed evidence for "Average" pooling occurring, to various extents which can be interpreted as the occurrence or

[0]A subset of the models and experiments presented in this chapter are published in Proceedings of IEEE International Joint Conference on Neural Networks 2012 (IJCNN2012) (Jalali et al., 2012) and another subset is accepted in the annual meeting of the Cognitive Science Society (CogSci 2013) (Jalali et al., 2013c) and are in preparation for submission to the Journal of Neural Networks.

frequency of the features. There is strong evidence that the primate visual system is also tuned to the co-occurrence statistics. This refers to either the joint or conditional probabilities of two (or more) features occurring together within images belonging to a certain object category or across categories. The detailed discussion on these evidences and studies will be presented in Section 5.2.

We develop and implement a series of experiments to investigate the role of different pooling methods in HMAX model which are biologically inspired and fit well to the visual cortex mechanisms. In this chapter, we investigate the use of mean pooling (thereafter HMean) and use this pooling method along with max pooling conventionally performed in HMAX model to show that the information encoded using these two different pooling models results in better classification performance on Caltech101, Scenes, Soccer and Flowers datasets. We also investigate encoding co-occurrence of features and show that adding a higher layer to the HMAX structure where co-occurrence of features is encoded as a new dictionary of features improves the classification accuracy in a subset of Caltech256 dataset where a higher number of training images is available.

5.2 Background on Biological Inspirations

In this section, we introduce the biological inspirations on Mean pooling and discuss the biological inspirations for co-occurrence and provide justifications.

5.2.1 Biological Inspirations for Mean Pooling

There is evidence for Max spatial pooling occurring at multiple levels in the visual system in the primary visual cortex (also known as $V1$) of cats (Finn and Ferster, 2007; Lampl et al., 2004), as well as in the higher visual areas of monkeys, such as areas $V4$ (Gawne and Martin, 2002) and IT (Sato, 1989). Despite that monkey parietal cortex is associated with attention, rather than invariant object recognition, Max pooling has also been found in this area (Oleksiak et al., 2011).

Generally speaking, these studies investigated the relationship between the responses to single stimuli versus pairs of stimuli. They compared the response to a pair of stimuli (each placed at a different position) to the responses when each stimulus was shown separately (but at the same positions as the paired-stimuli case) for each neuron being recorded from. Overall, there was evidence that the response to the pair of stimuli is about the same as the larger of the two responses to each stimulus. This is consistent with the idea that Max pooling is performed over spatial location. Importantly, however, these studies cannot (and have not) been interpreted as evidence for only Max pooling is taking place. Each of these studies also showed evidence for "Average" pooling occurring, to various extents.

Sato (1989), showed that the response to two bars was "usually similar or less than the stronger response in the single stimulus condition". Max pooling only accounts for the cases in which the response to two bars was similar to the stronger response. More quantitatively, the study used a summation index (SmI) to calculate spatial summation. An SmI of 1.0 cor-

responds to Sum pooling, 0.0 to Max pooling, and -0.5 to Average pooling. Over the population of monkey IT neurons that they recorded from, the mean SmI was either 0.01 or -0.18, depending on whether the two stimuli were in different or the same halves of the visual field, respectively. This means that overall, the pooling was either Max-like, or between Max and Average. These results were from the experiment in which the monkeys were simply fixating. Interestingly, when the monkeys were made to actually perform a visual discrimination task, the mean SmI values became more negative. In other words, spatial pooling became more Average-like.

The study in Gawne and Martin (2002) of $V4$ neurons in monkeys also found that Max pooling is not the only type of pooling present. While Max pooling was a good model for "a substantial fraction" of the neurons, "for many neurons, Gawne and Martin (2002) could not determine any clear relationship". Like the study in (Sato, 1989), there was little evidence for Sum pooling. However, while Max pooling generally predicted responses better than Average pooling, the correlation between residual MSE for the two types of pooling was high ($r = 0.83$). Also, Average pooling was better for a number of neurons. In other words, overall there is evidence that Average pooling may also occur.

The study in (Lampl et al., 2004) of $V1$ neurons in cats did not examine Average pooling per se, but their results mirror those of the previous two studies, in that there was strong evidence that Max pooling is a better model than Sum pooling. Using the same spatial summation index as Sato (1989), the mean index value was again close to 0 (corresponding to

Max pooling). However, the variation in index value was large ($\pm 0.55SD$), suggesting that Max pooling alone does not account for all neurons. A follow-on study (Finn and Ferster, 2007), also of cat $V1$ neurons, found similar results, concluding that there is a continuum of pooling behaviors.

Two other studies have also found evidence for Average pooling. Reynolds et al. (1999) studied $V4$ neurons in monkeys, and found that neural responses to pairs of bar stimuli were generally a weighted average of the responses to individual bars. However, similar to the other studies mentioned, this weighting factor varied from cell to cell, and does not preclude Max-like behavior (i.e. weighting that is close to 1.0, strongly favoring one response over the other). On the other hand, Zoccolan et al. (2005) explicitly compared monkey IT neural responses to predictions from Max and Average pooling, and found that Average pooling was the better model.

Overall, there is evidence for both Max and Average pooling, while Sum pooling is clearly not performed in biological visual systems. Due to experimental noise and other reasons, Max and Average pooling are hard to distinguish with absolute certainty. This is especially the case because most studies use pairs of stimuli that already elicit significant responses individually, therefore the difference between the Max and Average is mathematically limited to only a fraction of the full range of possible responses.

Another possible reason for the difficulty in distinguishing Max and Average pooling, could be that biological systems do in fact utilize a continuous range of pooling functions. As suggested by the study of Sato (1989), the pooling functions could also be dependent on attentional state,

as well as the task being performed.

On a final note, the difference between Average and Sum is in the nature of the denominator. If it is a constant, for example, if the size of the neighborhood (i.e. number of neurons) being pooled over is the same across all features and locations, then Average and Sum are effectively the same, except for a constant factor. Computationally speaking, a classifier would produce the exact same result in both cases.

5.2.2 Biological Inspirations for Co-occurrence

Beyond just being tuned to the statistics of feature occurrences, there is strong evidence that the primate visual system is also tuned to the co-occurrence statistics. Since a "feature" is not a well-defined concept, how can the co-occurrence of two features be distinguished from the occurrence of a single feature that happens to be comprised of two elementary features? To make this distinction unambiguous, experiments are designed such that the elementary features are visually distinct, due to explicit segmentation, due to spatial separation, or from the task context. We term such features, which are the result of sensitivity to co-occurrence, as "co-occurrence features".

In some sense, mid-level features themselves can be considered as co-occurrence features, with their elementary features being simple orientation-sensitive filters (corresponding to orientation-sensitive neurons in primary visual cortex). Since lines, curves and contours are ubiquitous in images, the presence of a short line segment of a certain orientation strongly pre-

dicts that the orientation of a neighboring line segment will be similar. This is particularly so, if the relative position of that neighboring line segment is such that the two line segments have the possibility of being collinear.

Our focus here is on high-level features whose elementary features are more complex than simple oriented filters. These high-level features approach the level of semantic object parts or possibly even objects themselves. In the rest of this section, we will review the experimental evidence that the primate visual system develops sensitivity to such high-level co-occurrence features.

In the field known as visual statistical learning (VSL), it has clearly been shown that adult humans develop sensitivity to co-occurrence statistics in images (Fiser and Aslin, 2001; Aslin and Newport, 2012). In a groundbreaking study by Fiser and Aslin (2002) it was shown that, amazingly, 9-month-old infants similarly develop sensitivity to visual co-occurrence statistics.

There is also an abundance of evidence from monkeys that their visual systems develop sensitivity to co-occurrence statistics. In an early work by Miyashita (1988); Sakai and Miyashita (1991), they trained monkeys to recognize pairs of stimuli, in paradigm known as paired-associate learning. Neurons were found that were sensitive to such trained stimulus pairs, but not other stimulus pairs. The pairings were arbitrary, making the likelihood that such neurons had already possessed such sensitivity vanishingly small. More recently, Hirabayashi and Miyashita (2005) found that populations of IT neurons are sensitive to feature configuration within objects.

Direct evidence for sensitivity to co-occurrence above and beyond sensitivity to occurrence was found by Baker et al. (2002). Monkeys were trained to discriminate objects that were each composed of two distinct parts linked by a line, forming "baton" objects. Compared to untrained objects, selectivity for trained objects was enhanced. This was for both the individual parts, as well as the combined "baton" objects. Crucially, selectivity for the two parts together (i.e. the whole object) was greater than the combined (summed) selectivity for each individual part.

Under what conditions does sensitivity to co-occurrence develop? In human adults, this is an implicit process that develops without awareness of the co-occurrence statistics, using a "cover task" or even through mere exposure (Turk-Browne et al., 2005, 2009; Aslin and Newport, 2012). This is also true for human infants (Fiser and Aslin, 2002; Aslin and Newport, 2012). In monkeys, most work has been done using active task learning. This is so that the neural selectivity for trained objects can be compared to the control set of untrained objects. Since neural selectivity is enhanced for features that are diagnostic for active task learning (Sigala and Logothetis, 2002), passive viewing may not be sufficient to produce selectivity that is large enough to be statistically significant when measured from electrode recordings.

How has sensitivity to co-occurrence been measured experimentally? The methods have generally been constrained by the nature of the subjects. Adult human subjects have generally been tested behaviorally, i.e. through their explicit responses (usually simple yes/no tests). More re-

cently, fMRI has been shown to be able to detect co-occurrence sensitivity (Turk-Browne et al., 2009). In human infants, due to their inability to understand or respond explicitly to verbal instruction, experiments have been constrained to using tests for novelty detection that are ubiquitous for infants. In monkeys, due to the ability to conduct invasive experiments that are not possible with humans, scientists have conducted electrophysiological experiments (i.e. using electrodes to record the responses of individual neurons). Such experiments allow for a detailed, "close-up" analysis of the effects of co-occurrence at the level of individual neurons e.g. Baker et al. (2002); Sakai and Miyashita (1991). However, there are limitations, such as the presence of noise, limited recording time, and the ability to record from only a few hundred neurons at most.

Beyond just "being sensitive" to co-occurrence statistics, what are the characteristics of such sensitivity? It is specific to spatial configuration, such as the relative position of the elementary features (Hirabayashi and Miyashita, 2005). In addition, this sensitivity is reflected not in strength of neural responses per se, but rather in the selectivity for co-occurring features relative to non-co-occurring features (Baker et al., 2002).

One special case of sensitivity to co-occurrence of features is that of faces. The elementary features are semantic face parts such as the eyes, nose and mouth. It is very well-established that humans and monkeys are sensitive to the combination and relative configuration of face parts. Specifically, any change to the normal configuration of the face leads to reduced neural responses and poorer recognition accuracy. One manifestation of

this is the Face Inversion Effect (FIE), whereby inverted faces are much more poorly recognized than upright faces (Yin, 1969). Faces with the parts in scrambled configurations are also poorly recognized. Furthermore, the sensitivity to co-occurrence seems to be unavoidable. In what is known as the Composite Face Effect, people are sensitive to the bottom halves of faces, even when they are explicitly instructed to ignore them during a discrimination task (Young et al., 1987).

Generally, such sensitivity requires normal visual experience during infancy in order to develop (Le Grand et al., 2004). It also develops quickly, reaching adults levels (at least qualitatively) by age 4 (de Heering et al., 2007); this is consistent with the notion that passive exposure is sufficient for co-occurrence sensitivity to develop (see above). Evidence for sensitivity to co-occurrence for face parts has also been found at the level of single neurons. Freiwald et al. (2009) found that in one of the brain regions that respond selectivity to faces, neurons on average responded to combinations of two to three face parts, rather than individual parts. Co-occurrences have been studied in a series of experiments such as Edelman et al. (2002).

Use of co-occurrences of features for creating more complex features in Fidler et al. (2008) shows an improvement in classification accuracy, and bag of features approaches show improvements in classification results using frequency of patches in the images in (Li and Perona, 2005). Co-occurrence information can be used to find part-part and part-whole relations of features of different receptive field sizes. If a feature is occurring too often in a class (and not likewise in other classes), it is more likely to be a discrim-

inant feature in that class and if two features are co-occurring in a class often in a neighborhood, they may be part of a more complex feature and can have a part-part relation and they may be more related to the object rather than the background (unless the background is also repetitive e.g. sky in airplane images). In addition, if there exist features of different sizes and they are co-occurring in the same position on different scales they are likely to have a part-whole relationship. We introduce our model to encode these characteristics in a biologically inspired model in short term and long term memory aspects.

In the rest of this chapter, we propose several pooling methods in Section 5.3 and introduce several approaches for encoding co-occurrence of features in Section 5.4. We show the experimental results and compare these models in Section 5.5 and a discussion and conclusion is provided in Section 5.6.

5.3 HMean

We use the HMAX model presented in Mutch and Lowe (2008) in the first three layers ($S1$, $C1$ and $S2$) as explained in detail in Chapter 3. Here we have a brief review on this model and show our modifications to it. In this implementation, an image is fed into the structure and 10 different scales of the image are created as inputs to $S1$ layer, as can be seen in Figure 5.3.

Gabor filters in 12 orientations are created as $S1$ layer filters:

$$G(x,y) = \exp\left(-\frac{(X^2+\gamma^2Y^2)}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}X\right). \tag{5.1}$$

where $X = x\cos\theta - y\sin\theta$ and $Y = x\sin\theta + y\cos\theta$. The values of x and y vary between -5 and 5, and θ varies between 0 and π. The parameters γ (aspect ratio), σ (effective width), and λ (wavelength) are all taken from Serre et al. (2005) and are set to 0.3, 4.5, and 5.6 respectively.

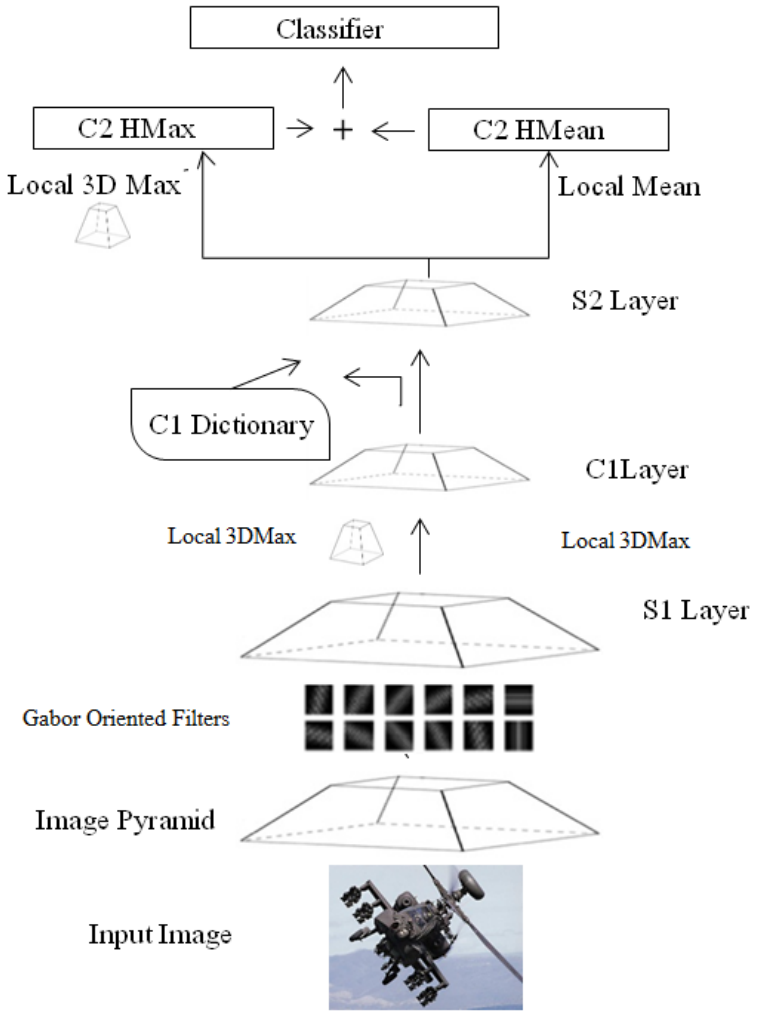


Figure 5.1: The use of Average pooling (HMean) and Max pooling (HMAX).

A fixed size of Gabor filters is implemented on different scales of the images where the smaller edge of the biggest image is set to 140 pixels while

maintaining the aspect ratio (the image pyramid of 10 scales created each layer by a factor of $2^{1/4}$ smaller than the last using bicubic interpolation). The response of a patch of pixels X to a particular $S1$ filter G is given by:

$$R(x,y) = \left| \frac{\sum X_i G_i}{\sqrt{\sum X_i^2}} \right| \tag{5.2}$$

These outputs are sent to the $C1$ layer, which performs a local $3D$ max operation on both scale (± 1) and position (3×3 neighborhood) of the filter responses. The output of this layer is a pyramid consisted of between 500-2000 different patches of size 4×4, 8×8, 12×12 and 16×16 in 8 scales depending on the size of the input image. In this level one or two samples are randomly sampled from each training image (from random scales and positions) and a dictionary of features of size 4096 is created. This dictionary is then made sparse by selecting the highest response from each orientation and setting the rest to 0.

The response of a patch of $C1$ units X to a particular $S2$ feature/prototype P (a dictionary feature), of size $n \times n$, is given by a Gaussian radial basis function:

$$R(X,P) = \exp\left(-\frac{\| X - P \|^2}{2\sigma^2 \alpha} \right) \tag{5.3}$$

The values of R are stored as $S2$ layer. The distance of each sample from each training image with each entry on the dictionary is calculated and a local max is taken in $C2$ layer in $\pm 1 scale$ and $\pm 10\%$ spatial neighborhood (despite a global max in Serre et al. (Serre et al., 2005)). These $C2$ features are sent to the SVM for training. For testing images the same hierarchical

procedure is repeated. In (Mutch and Lowe, 2008) sparse prototypes are calculated and the max response from all directions for each window is taken and SVM normals method (Mladenić et al., 2004) is used to select the features with higher weights. In this approach, SVM is run a few times, and each time features with lower weights are dropped. In this HMAX implementation, once $S2$ features are calculated, the $C2$ layer is calculated as:

$$C2(n) = \max(V_k^n) \text{ for } \forall k \in M$$
$$\text{for } n = 1, ..., N \tag{5.4}$$

As can be seen in Figure 5.2 in conventional HMAX approaches, the max on the columns is taken as the value for $C2$ either in a local neighborhood of each feature or globally. Since taking the max in a local neighborhood (in ± 1 scale and $\pm 10\%$spatial neighborhood) is shown to improve the performance by about 5% in Caltech101 dataset in (Mutch and Lowe, 2008), in our experiments we also use a local neighborhood for calculating the responses. We also eliminate the local inhibition in $S2$ level proposed in (Mutch and Lowe, 2008) as it increased the performance by another 0.5%. Once a feature belongs to the first or last scale in the pyramid, we extend the neighborhood to two neighboring scales. The same method is used for features which fall in the borders of each scale, and $+20\%$ or -20% of their neighborhood is used for comparisons. In the rest of this section, we investigate different pooling methods in order to create the $C2$ layer.

We calculate the $C2$ layer in 4 different combinations:

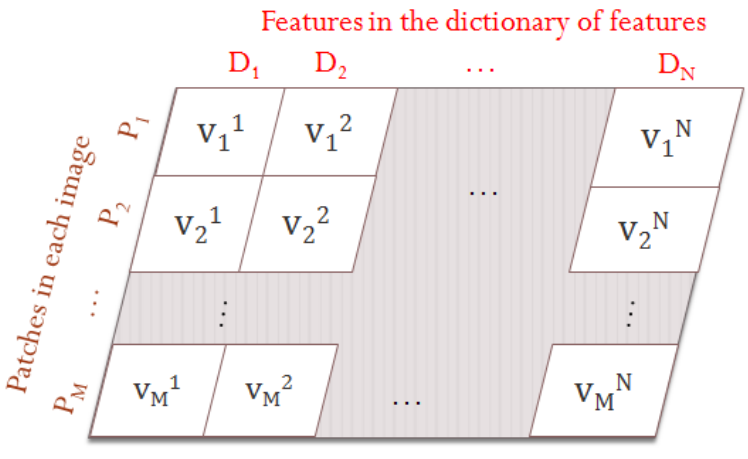


Figure 5.2: The use of frequency of features vs. the use of the best matching unit (BMU) response. In HMAX implementations, the max on the columns is taken as the response for creating $C2$ output vector. In contrast, histogram approaches using SIFT methods, use the statistics of occurrences of features, i.e. the normalized sum of the max values on the rows.

Hard Max:

$$V_m^n = \begin{cases} 1 & \text{if } V_m^n = \max_{n=1}^{N} V_m^n \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

Threshold:

$$V_m^n = \begin{cases} 1 & \text{if } V_m^n >= Threshold \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

Soft Max:

$$V_m^n = \begin{cases} 1 & \text{if } V_m^n = \max_{n=1}^{N} V_m^n \\ V_m^n & \text{if } Threshold < V_m^n < 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

and **Actual Value** in which all of the values are used without any change.

Once the values of V_m^n are updated in different modes, the $C2$ response is calculated as follows:

$$C2(n) = \sum_{k=1}^{M} V_k^n \text{ for } n = 1, ..., N \tag{5.8}$$

We name these $C2$ vectors Frequency $C2$ or $FC2$ for short. The $C2$ vector is normalized to a value between 0 and 1 in all modes. For simplicity we name $C2$ vectors created by Average mode: $FC2AV$, Hard Max: $FC2HM$, Soft Max: $FC2SM$ and Threshold: $FC2T$. Since the $FC2AV$ is more biologically inspired and it also shows the best classification results (presented in Section 5.5.1), we name it HMean. This model is shown in Figure 5.3. The HMean is equivalent to the occurrence or frequency in "bag of features" methods, as we calculate the sum of the values on the rows in Figure 5.2 and normalize it. The terms "HMean", "occurrence" and "frequency" are used interchangeably in the rest of this thesis.

In summary, in HMAX model, the maximum response of the $S2$ layer is chosen as the $C2$ layer to be fed to the classifier. However in HMean, the average response (mean pooling) is taken as the response to be fed to the classifier as the $C2$ layer.

In Section 5.5 we show the classification results acquired by these different methods on Caltech101 dataset and show different concatenation of these features with conventional max features from HMAX and show that the use of these features in concatenation with $C2$ features improves the classification performance on several datasets such as Caltech101, Flowers, Soccer and Scenes.

5.4 Encoding Co-occurrence of Features

In order to calculate the co-occurrence of features, we use the $C2$ features calculated using Equation 5.8 on actual values calculated, called HMean. In the next step, we find the most occurring features (MOF) in each class as follows:

$for \quad i = 1 \quad to \quad$ NMOF

$MOF(i) = \max_{n=1}^{N} C2(n)$

$IMOF(i) = \text{argmax}_{n=1}^{N} C2(n)$

$C2(\text{argmax}(C2(n))) = 0;$

end

Using the loop shown above, we find the value and index of the most frequent features in each class. The next step is to encode the co-occurrence of these features as can be seen in Figure 5.4. For every class, we calculate the co-occurrence of the most frequent features and store it as a $S3$ dictionary feature. Hence a new dictionary of features is added to the model which is composed of NMOF $\times$ NMOF entries for each class. In this dictionary of features, the value of each dictionary feature is calculated as:

$$C3(i,j) = C2(i)C2(j) \exp\left(-\frac{\parallel S_i - S_j \parallel^2}{2\sigma^2\alpha}\right) \tag{5.9}$$

where S_n represents the spatial position of the $C2$ feature and $\sigma = 0.5$ (among different values chosen for σ in the experiments). We also tried eliminating the spatial distance part of the equation which resulted in lower performances.

This dictionary encodes the value of co-occurrence of every pair of fea-

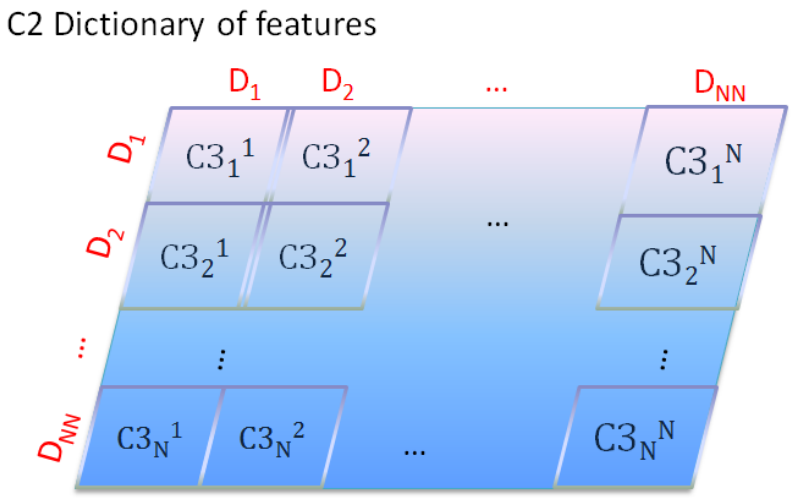


Figure 5.3: Creation of C3 dictionary for encoding co-occurrence of features.

tures selected for each class. Hence we will have NN dictionaries where NN stands for the number of categories in the classification task. These dictionaries are concatenated to create the $C2$ dictionary of features. In the training and test phases, the respective feature to each dictionary feature is found (the most similar feature in every image) and the similarity of the values in dictionary of features are calculated for every image. This results in a NMOF $\times$ NMOF$\times$ NN feature as the $C3$ feature and it is concatenated to $C2$ feature vector and sent to the classifier for classification. The extended model for encoding the co-occurrence of features is shown in Figure 5.4.

Another interpretation of the co-occurrence information in this structure is to provide a probabilistic base for it. In order to implement probabilities in this approach we can encode frequency of features as prior probabilities of having the feature in a specific class, and co-occurrences as joint probabilities of two features. Using steps described before, we calculate

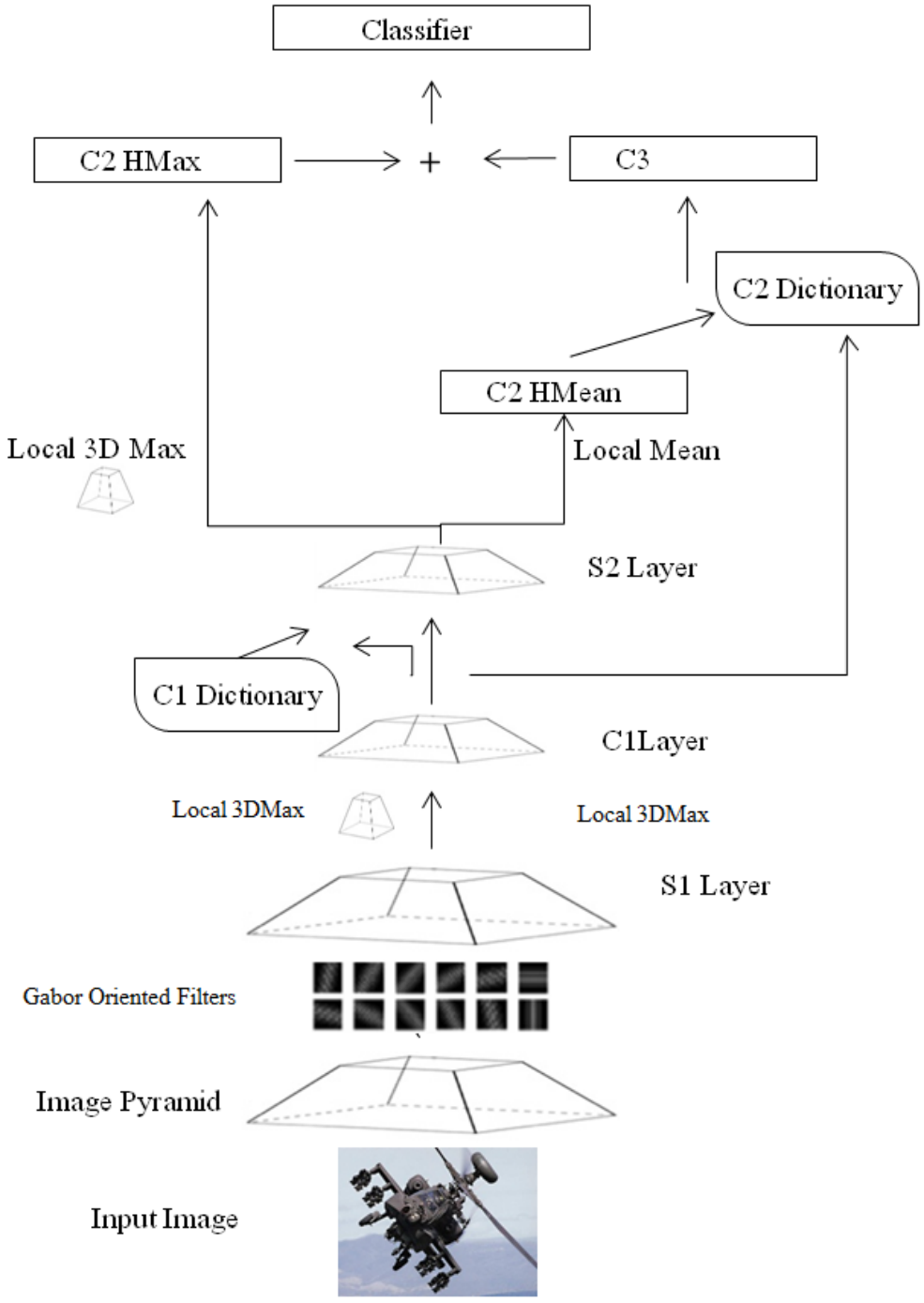


Figure 5.4: The main model encoding co-occurrence of features.

$p(f_i|c_k)$ for every feature for every class which is the prior probability of the feature in a class. We also calculate $p(f_i, f_j|c_k)$ which is the joint probability of two features co-occurring for every class c_k. Once all prior and joint probabilities described above, are calculated, we will find $p(c_k|f_i, f_j, ...)$ for every class, and the class with higher probability is more likely to be the class that image belongs to. Since the number of features is high, this approach needs extensive computations. Hence we only considered the co-occurrences of every two features in the statement above. This proposed method is an approximation of the Bayesian probabilities when two features are independent. However, if two features are independent, encoding their co-occurrence does not add any information. Hence the variable of the distances of two features provides a degree of dependence between them, and in Section 5.5 it is shown that removing the spatial parameter which encodes the distances between two features, results in poorer classification results.

$$P(Y = y|X_1 = x_1, X_2 = x_2) = \frac{P(Y = y, X_1 = x_1, X_2 = x_2)}{P(Y, X_1 = x_1, X_2 = x_2)} = \frac{\sum_{X_3,..,X_n} P(Y = y, X_1 = x_1, X_2 = x_2, X_3, ..., X_n)}{\sum_{Y,X_3,..,X_n} P(Y, X_1 = x_1, X_2 = x_2, X_3, ..., X_n))} \quad (5.10)$$

Here we describe two other models for encoding co-occurrence of features using a long-term and a short-term memory for storing the co-occurrence weights.

A Neural Network Based Model with a Long-Term Memory

Once the dictionary of features is created using the techniques described above, we code the occurrence and co-occurrence information of features in different classes. This information is stored in a neural network structure in a long-term memory for every class. In the training and test phase, we create a long term memory encoding occurrence and co-occurrence information of the features for every image and feed it to a neural network. The occurrence and co-occurrence information of features of every class is encoded separately. This model is illustrated in Figure 5.5.

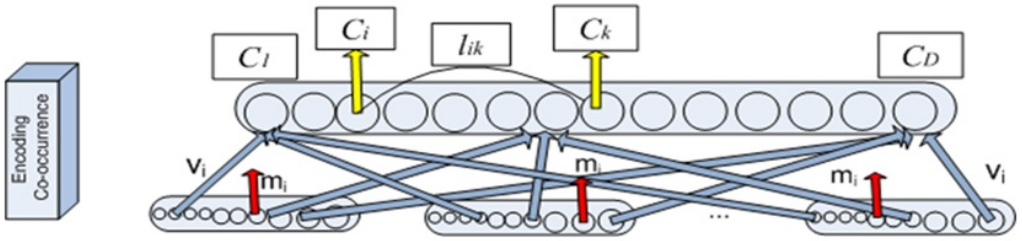


Figure 5.5: The neural network model with long-term memory for encoding co-occurrence of features.

We calculate the occurrence of each feature in the dictionary of features, using Equation 5.3) v ($v_p = \sum_X R(X, P)$. We find the best matching unit to each feature in the dictionary, and store their similarity using Equation 5.3 as m.

$$c_i = \sum_j m_i^j v_i^j + \sum_k c_i c_j l_i^k \tag{5.11}$$

where j is the dimension of patch i (4×4, 8×8, 12×12, 16×16) and k is the dimension of dictionary of features. The network is then trained using

all of the information for every patch response in the previous layer using:

$$v_i(n+1) = v_i(n) + \Delta v_i \tag{5.12}$$

$$\Delta v_i = \alpha v_i m_i \tag{5.13}$$

and

$$l_i^k(n+1) = l_i^k(n) + \Delta l_i^k \tag{5.14}$$

$$\Delta l_i^k = \alpha_i^k c_i c_k \tag{5.15}$$

where l_i^k is the lateral weight of two features c_i and c_k. The value for variable α in Equation 5.13 can be chosen as a constant parameter and in our case, we set it to 0.5 and in Equation 5.15 as a closeness measure of the two features as shown in Equation 5.16. The value of v_i is initialized to 0 and updated using:

$$\alpha_i^k = \exp\left(-\frac{\parallel s_i - s_k \parallel^2}{2\sigma^2}\right) \tag{5.16}$$

where s_i is the spatial position of the feature in the image and $\sigma = 1$. α is a normalizing factor which is $\left(\frac{m}{4}\right)^2$ where $m = 4, 8, 12, 16$ is the feature size to boost the weight of bigger patches, as their similarity is normally less.

This network will be trained for all training images for every single class, and the information calculated here, will be saved for every class separately and will be considered as a long-term memory storing co- occurrence information of images of every class separately. Once this information

is calculated, we feed all of the training images to this network again, and get a response for every node in the C layer.

We can use this model for classification in an image level instead of class level. In this mode, the dimension of vector to be fed to classifier is 4000 for every image, since we do not store any class information anymore and variables v_i and l_i^k which store the class information (long term memory) are not required to be updated anymore and variable l_i^k in Equation 5.11 can be substituted by α in Equation 5.16.

A Neural Network Based Model with a Short-Term Memory

In order to encode co-occurrence information in images for each class, we proposed a neural network based model. In this approach, once the occurrence of features for each class are created, we convolve all patches on the dictionary of features, and for every patch in the image X, we find its best matching unit in the dictionary of features (over all patches p, in the dictionary), and add this distance to value v, standing for occurrence of that specific feature in the dictionary of features($v_p = \sum_X R(X, P)$ calculated using Equation 5.3).

Once this information is calculated for all patches in the image, we have a distribution of occurrences of patches in that specific image. Based on these values, we calculate the co-occurrence values W for every two features as:

$$W_{k,l} = \sum_{\forall k,l} v_k^n * v_l^m * e^{-\left(\frac{\|s_k^n - s_l^m\|^2}{2\sigma^2 \alpha}\right)} \tag{5.17}$$

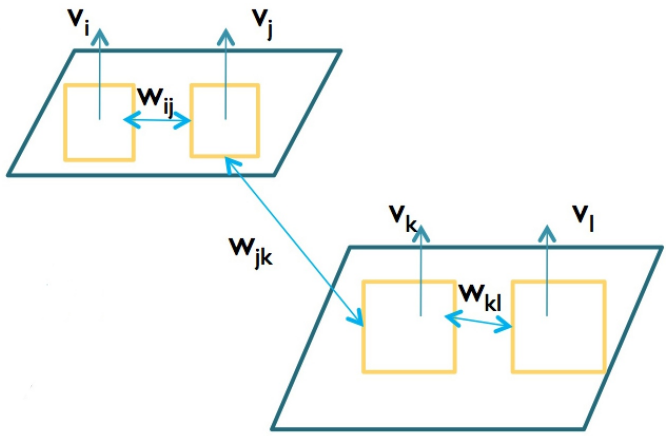


Figure 5.6: The neural network model with short-term memory for encoding co-occurrence of features.

where s_k^n is the spatial location of feature f_k^n where k is the index of feature f, and n is the respective receptive field size which f belongs to. In the next step, we calculate this information for every training image to train the classifier based on:

$$Sc(C_i^k) = \sum_{R_k} sim(f_j^{C^i}, f_j^T) * sim(f_l^{C^i}, f_l^T) *$$
$$sim(v_j^{C^i}, v_j^T) * sim(v_l^{C^i}, v_l^T) * sim(w_{jl}^{C^i}, w_{jl}^T) \quad (5.18)$$

where $sim(x, y) = e^{\frac{-||x-y||^2}{2\sigma^2}}$ and f_j^T is feature j in image T (T stands for both training and test images) and v_j^T is the frequency information of feature f_j^T and w_{jl}^T is the co-occurrence value of features f_j^T and f_l^T. This model is illustrated in Figure 5.6.

Figure 5.7: Sample images of (a) Caltech101 (b) Outdoor Scenes (c) Soccer and (d) Flowers datasets.

5.5 Experimental Results

In order to evaluate the proposed models in Section 5.3 and 5.4, we ran several experiments on Caltech101, Caltech256, Soccer, Scenes and Flowers datasets.

5.5.1 HMean

In this section, we show the performance of concatenation of HMean with HMAX on Caltech101, Scenes, Soccer and Flowers datasets. Sample images of these datasets are shown in Figure 5.7. The HMAX model used in these experiments is from Mutch and Lowe (2008).

Caltech101 Dataset

Caltech101 dataset contains 101 categories of objects plus one background category and is introduced in more detail in Chapter 4. We used 30 training images per category for training the model and used the rest as test images. (between 1 to 800 per class). The results of classification on Caltech101 dataset in different modes are shown in Table 5.1

As can be seen from Table 5.1, the best performance in different pooling methods is achieved on Caltech101 dataset when the actual values are summed and nomalized, which is equaivalent to a mean pooling operator. We name this method, HMean. The best performance is achieved when the Max pooling and Mean pooling are concatenated.

Soccer Dataset

The Soccer team data set contains images from 7 soccer teams taken from the web, containing 40 images (approximately 300×300 pixels) per class, divided into random 25 training and 15 testing images per class. Although players of other teams were allowed to appear in the images, no players being a member of the other classes in the dataset were allowed (Van De Weijer and Schmid, 2006). As shown in Table 5.1, the combined use of HMean and HMAX models provides significant improvements over using HMAX model alone. Since the images from different classes share similar shapes, using color results in better performance than shape; this is investigated in more detail in Chapter 6.

Flowers Dataset

The 17-category Flowers dataset (Nilsback and Zisserman, 2006) consists of 17 categories of common flowers in the UK with 80 images (approximately 600×600 pixels) per class. The images have large scale, pose and light variations and there are also classes with large variations of images within the class and close similarity to other classes. The classification results of our model on this dataset are shown in Table 5.1. We use 50 random images from each category for training and 30 for testing as in (Nilsback and Zisserman, 2006).

8 Scenes Dataset

This dataset contains 8 outdoor scene categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways. There are 2600 color images of size 256×256 pixels (Oliva and Torralba, 2001). We used 100 random images per category for training and the rest for testing (an average of 236 per category). As shown in Table 5.1, the use of HMAX and HMean significantly improves classification performance when concatenated. Further investigation on use of color in classification results of this dataset is provided in Chapter 6.

Scenes, Flowers and Soccer datasets share similar shapes in different colors among categories that make use of color information more important in them as will be discussed in Chapter 6.

Method	**Caltech101**	**Scenes**	**Soccer**	**Flowers**
HMAX	54.7% ±1.2	71.48% ±2.1	24.76% ±4.2	42.54% ±3.7
$FC2AV$	42.7% ±1.1	73.71% ±1.8	26.67% ±2.3	36.67% ±1.2
$FC2HM + C2$	57.2% ±0.9	71.27% ±1.4	49.19% ±1.2	48.24 % ±1.9
$FC2T + C2$	55.9% ±1.5	69.18% ±2.4	45.72% ±3.1	45.15% ±2.8
$FC2SM + C2$	56.3 % ±0.9	70.14% ±1.2	47.58% ±1.8	46.13% ±0.5
$FC2AV + C2$	**58.9**% ±0.6	**81.24**% ±0.9	**52.17**% ±1.2	**51.12**% ±1.1
$FC2AV.C2$	44.6% ±1.4	65.45% ±1.8	28.32% ±1.7	38.74% ±1.0

Table 5.1: Classification performance on four datasets by use of frequency of features in different modes. $'+'$ and $'.'$ stand for concatenation and inner product of two vectors respectively. FC2AV is for Actual Value FC2, FC2HM+C2 is for concatenation of HMAX C2 features with hard max FC2, FC2T+C2 is for threshold, FC2SM+C2 is for soft max and FC2AV+C2 is for actual values of C2 vectors described in Section 5.3.

5.5.2 Co-occurrence

We evaluated our co-occurrence model proposed in Section 5.4 on the Caltech101 dataset Li et al. (2004). The model was trained on 30 images per category (standard for this dataset; see Mutch and Lowe (2008)), and tested on all the other images. We also used the Caltech256 dataset, because it allows for more images per category for training than Caltech101. In particular, we considered only the 14 (out of 256) categories which had 200 or more images. We trained the model on 150 images (so that there would be at least 50 images for testing), and tested on the rest. We also

examined classification accuracy as a function of number of training images for Caltech256. This was motivated by the concern that co-occurrence features could require more data for reliable co-occurrence statistics to be extracted, before the advantage of co-occurrence could be properly manifested. Using the co-occurrence methods proposed in Section 5.4 for neural networks with short and long term memories, a very low accuracy is achieved since all the information of occurrence is transformed into a lower dimension and the accuracy achieved by these two methods is 20% and 10% respectively.

We also evaluated the performance of our model on a new dataset consisting of images of underwater targets. The main challenge with underwater images is the existence of particles that limit the visibility in unclear waters and results in scattering, reflection and absorption of light, and the differential absorption of light of different wavelengths by water itself. This dataset consists of 1664 images (roughly 740×420 pixels in size) from 13 categories. Example images from this dataset are shown in Figure 5.8. We used 30 images per category for training, and the rest for testing.

Results are shown in Table 5.2. For all images, only intensity (luminance) information was used. All results were derived using 8 random train/test splits. For all three datasets, the combination of HMAX and co-occurrence features gave better results (classification accuracy) than either type of feature alone (Caltech101: 59.3% vs. 54.7% vs. 57.7%; Caltech256: 64.4% vs. 60.2% vs. 48.6%; Underwater Images: 98.7% vs. 92.9% vs. 92.2%). Since co-occurrence features were derived from the co-occurrence

Figure 5.8: Examples from TMSI Underwater Images dataset.

of HMean features, we also compared which of these two feature types (co-occurrence vs. HMean) gave better results when combined with HMAX. Again, for all three datasets, combining co-occurrence features with HMAX produced better results than combining HMean with HMAX (Caltech101: 59.3% vs. 58.9%; Caltech256: 64.4% vs. 61.3%; Underwater Images: 98.7% vs. 98.3%). Furthermore, for all datasets, the combination of all three feature types was better than just HMAX and HMean together (Caltech101: 60.1% vs. 58.9%; Caltech256: 64.1% vs. 61.3%; Underwater Images: 99.0% vs. 98.3%).

We also examined the effect of disregarding spatial distance (i.e. the exponential in Eq. 5.9). As seen in Table 5.2, for all datasets, results were better when spatial distance was taken into account (Caltech101: 57.7% vs. 55.1%; Caltech256: 48.6% vs. 44.2%; Underwater Images: 92.2% vs. 83.3%).

In order to evaluate the effect of number of training images for the creation of co-occurrence features, we trained the model with varying numbers of training images per category. As shown in Figure 5.9, the performance boost was observed when adding co-occurrence features was greatest with the use of 150 training images. However, for fewer than 150 training images, the boost from adding co-occurrence features is unreliable. Nonetheless, looking at just HMAX alone, performance seems to asymptote at 150 training images, but for the combination of HMAX and co-occurrence features, performance seems to increase roughly linearly with the number of training images. While empirically, co-occurrence may help performance in

Method	**Caltech101**	**Caltech256 (subset)**	**Underwater Images**
HMAX	54.7% ±1.4	60.2% ±1.7	92.9% ±2.1
Co-occurrence (no distance)	55.1% ±2.2	44.2% ±3.1	83.3% ±0.8
Co-occurrence	57.7% ±1.1	48.6% ±1.8	92.2% ±1.4
HMAX + Co-occurrence	59.3% ±1.3	64.1% ±1.2	98.7% ±1.1
HMAX + HMean	58.9% ±1.6	61.3% ±1.1	98.3% ±1.0
HMAX + HMean + Co-occurrence	**60.1**% ±0.6	**64.4**% ±0.8	**99.0**% ±0.2

Table 5.2: Classification performance on the Caltech101, Caltech256 (subset – see text for details), and TMSI Underwater Images datasets.

all datasets, similar analyses (i.e. performance boost as a function of number of training images) for the other 2 datasets may not be meaningful, since the maximum number of training images is only 30 per category.

5.6 Discussions

In this chapter, we introduced several approaches for pooling and encoding occurrence and co-occurrence of features in a biologically inspired hierarchical structure. As shown in Section 5.5, the use of HMAX and HMean encodes more information when concatenated together (Jalali et al., 2012). The main difference between HMAX and HMean is in the pooling in the $C2$ layer. In the combined model (HMAX+HMean), the Average pooling

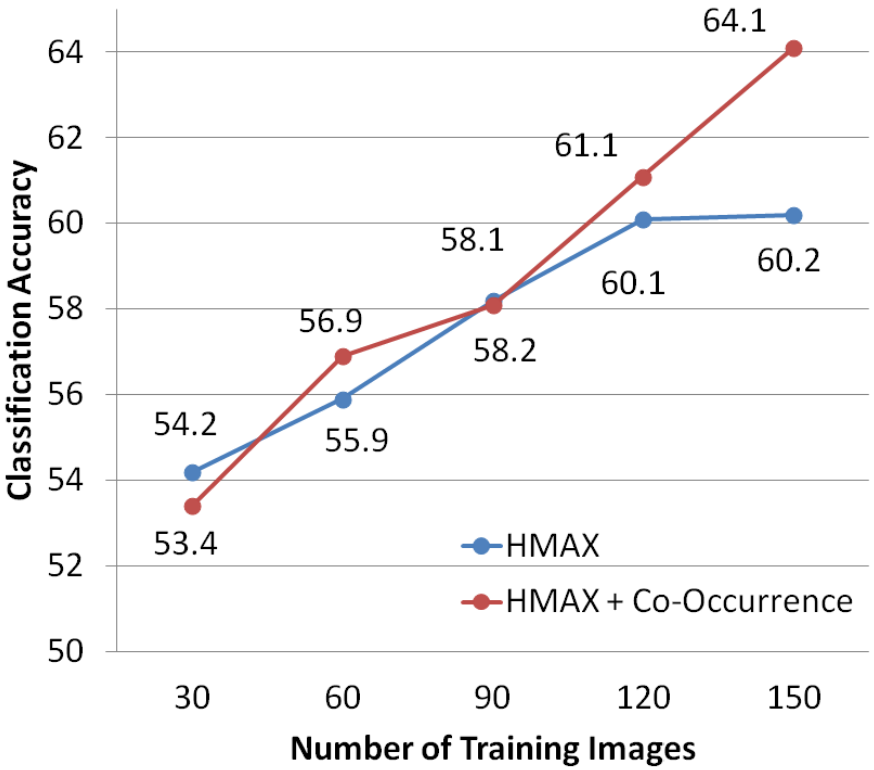


Figure 5.9: Classification accuracy on Caltech256 as a function of number of training images.

is performed along with max pooling and the classification results show a significant improvement. This is also biologically inspired as expounded in Section 5.1 as there is evidence for both pooling methods in the visual cortex. The use of co-occurrence of features also provides an increased performance however, the increase in Caltech101 is not significant due to the low number of training images and the high number of categories. Since the number of categories is high in Caltech101 dataset, selection of discriminative features becomes more important as there may be many redundant features that are occurring in many categories. In Caltech256 subset chosen where more images are available for training the model, we observed an improvement in classification results. These experiments open a path to further investigation of different methods for encoding co-occurrence of features. A top-down cross layered approach is shown to improve the performance in (Fidler et al., 2008). However in our experiments we did not use any top-down connections to provide heuristic information about features that are used for encoding the co-occurrence and encoding co-occurrence is solely performed in an unsupervised approach.

In this chapter, we showed that combining co-occurrence features with regular HMAX features leads to better classification performance than using either feature type alone. Furthermore, adding co-occurrence features to HMAX increases performance more than adding occurrence features. The three types of features encode different information, and therefore the combination of all three feature types gave the best overall performance. For co-occurrence, the spatial distance between the two co-occurring fea-

tures also contributes to better performance. In this chapter, we focused solely on HMAX. However, in future work, our co-occurrence method can be applied to other vision algorithms.

We also experimented with creating co-occurrence features from HMAX features (rather than HMean features). However, this resulted in either a drop in performance or no change.

Figure 5.9 suggests that the performance boost from using co-occurrence may be limited by the number of training images. More detailed investigation is limited by the relatively small number of images per category in these datasets. Further investigation may require utilizing or creating larger datasets.

Another prospect for further improvement is to encode co-occurrence of more than two features. However, besides possibly requiring even more training data than two-feature co-occurrence, there may be diminishing returns for such "higher-order" co-occurrences. This is because relatively fewer classes will have the underlying visual structure that will benefit from encoding such co-occurrences.

In this chapter, the choice of features for encoding co-occurrence was based on their frequency. Choosing discriminative (rather than frequent) features for co-occurrence encoding may be a more direct approach to maximizing classification performance. To choose discriminative features, one approach is to train the SVM several times and remove features with low weights, as in Mutch and Lowe (2008), or to simply use features with mean response values that differ the most between different classes.

Chapter 6

CQ-HMAX: A New Biologically Inspired Color Approach to Image Classification

We develop and implement a new approach to utilizing color information for object and scene recognition that is inspired by the characteristics of color- and object-selective neurons in the high-level inferotemporal (IT) cortex of the primate visual system. In our hierarchical model, we introduce a new dictionary of features representing visual information as quantized color blobs that preserve coarse, relative spatial information. We

[0]Part of the models and experiments presented in this chapter are accepted in the annual meeting of the Cognitive Science Society (CogSci 2013) (Jalali et al., 2013d) and are in preparation for submission to the Journal of Pattern Recognition.

run this model on several datasets such as Caltech101, Soccer, Flowers and Outdoor Scenes. The combination of our color features with (grayscale) shape features leads to double-digit average increases in performance over shape features alone. Using our model, performance is significantly higher than using color naively, i.e. concatenating the channels of various color spaces. This indicates that usage of color information per se is not enough to produce good performance, and that it is specifically our biologically-inspired approach to color that results in significant improvement. Among approaches that use bottom-up information only, the combination of three sets of biologically-inspired features (our high-level color features with existing low-level color features and grayscale shape features) achieves the best performance to date on the Soccer and Flowers datasets. In this chapter, we implemented our approach by extending one specific model (HMAX), but this approach to encoding color information can also be incorporated into other models.

6.1 Introduction

Many models are inspired by the hierarchical organization of the visual cortex proposed by Hubel and Wiesel (1959), such as Fukushima (1980), Riesenhuber and Poggio (1999), Hinton et al. (2006) and Hawkins and George (2006). Most of these models focus on image grayscale information and ignore color information. On the other hand, the primate visual system devotes impressive resources to color information processing (Zhang et al., 2012). While the broad use of color information in the primate visual sys-

tem is well-known, the details are still under active investigation (Conway et al., 2010). This is true not only for color, but also for shape and form (Op de Beeck and Baker, 2010; Lyon and Connolly, 2012). Nonetheless, in this chapter, we attempt to utilize what is currently known about the use of color to enhance object and scene recognition by computer algorithms. In this chapter we utilize the HMAX model (Mutch and Lowe, 2008) but this approach can be extended to be used with other computational models.

In our experiments we use the HMAX model (Riesenhuber and Poggio, 1999) in concatenation with our color model in order to evaluate the use of both shape and color. HMAX is a biologically inspired model which focuses on the shape processing capabilities of the ventral visual pathway, and has been used to perform classification tasks (Mutch and Lowe, 2008; Serre et al., 2007b).

We focus on modelling the high-level usage of color by incorporating insights from cognitive psychology and neuroscience. The broad intuitive inspiration for our model follows from the fact that colors are recognized categorically just as object classes are, even though color discrimination and matching is continuous (Palmer, 1999). Interestingly, people of different races (Boynton and Olson, 1987; Uchikawa and Boynton, 1987), as well as chimpanzees (Matuzawa, 1985), organize colors into the same basic color categories, such as red, blue, yellow, green.

More importantly for object and scene recognition, the categorical recognition of color suggests that, if color information is incorporated into object and scene classification, then fine-grained color information (e.g. precisely

specified hue) may not be necessary. For example, a beach scene might be recognized from the blue (sky and sea) and brown (sand) regions. It may not be important exactly how blue the sky/sea or how brown the sand grains are. In fact, it may be important to disregard such details in order to perform classification that is tolerant to variations in lighting, and so on.

In addition, the coarse relative spatial position of such color regions may be important. A blue region above a brown region might suggest a beach scene. If the relative positions are reversed, then the image is probably not a beach scene (or might be an upside-down one). Not only is the detailed spatial information unnecessary, it may be crucial to discard it and only retain coarse spatial information, since the exact spatial relations will depend on factors such as the shape of the beach and the camera angle.

Overall, our model can be loosely described as performing object and scene classification by reducing a given image to a "coarse arrangement of categorical color blobs", similar to the idea of spatial aggregation of visual keywords (Lim, 1999), but with realization on the HMAX model. This is different from approaches that utilize color information in a low-level fashion e.g. Zhang et al. (2012), although the two types of approaches are not mutually exclusive and can even be complementary (see Section 6.4). Crucially, our biologically-inspired approach clearly outperforms the naive use of color, where an image is decomposed into separate color channels that are processed independently until the final classification stage.

First, we go beyond the intuitive motivation for our approach and re-

view the specific biological evidence that the primate visual system utilizes color information in a manner that is broadly consistent with our model. Specifically, we review studies of color processing in the high-level visual area of the primate brain known as the infero-temporal cortex (IT for short), which is commonly associated with invariant object recognition (Logothetis and Sheinberg, 1996).

In the broadest terms, the IT is known to play an important role in color discrimination (see Komatsu (1998) for a review). A majority of the IT neurons are color-selective (Desimone et al., 1985) and two independent studies estimated this proportion to be roughly 70% (Komatsu et al., 1992; Edwards et al., 2003). Contrary to the theory that color processing occurs after more rapid luminance-only processing, no evidence was found that colored images evoke responses that are delayed relative to achromatic images (Edwards et al., 2003).

There is also more direct causal evidence for the role of the IT in color processing. Color discrimination is severely disrupted by IT lesions (Dean, 1979; Heywood et al., 1988) or cooling (Horel, 1994). Using position emission tomography (PET) imaging, color discrimination activates the IT more than brightness or position discrimination (Takechi et al., 1997).

Color-selective neurons in the IT are found in clusters, suggesting that they may form a roughly segregated and independent processing network (Conway and Tsao, 2006; Conway et al., 2007). As further evidence of this, a color cluster in one part of the IT (the anterior IT) received projections from a color cluster from another part of the IT (the posterior IT),

suggesting that these clusters of color-processing neurons form reciprocally-connected modules within a distributed network in the IT (Banno et al., 2011).

The IT neurons are selective for both hue and saturation (Komatsu, 1993). Different cells have different preferred hues, and as a population, the cells' preferred color spans most of the color spaces (Komatsu et al., 1992; Conway et al., 2007). The colors for which the IT neurons are selective for tend to correspond to the basic color names (Komatsu, 1997, 1998). Komatsu (1998) proposed that the IT has templates corresponding to color categories and may be involved in determining color category by finding the best match over these categories. More recently, the distribution of color-selective neurons found in the IT seems to correspond to the three to four most basic colors (Stoughton and Conway, 2008). The largest peaks align with red, green, and blue, in order of size of peak, with a smaller peak corresponding to yellow. These peaks roughly correspond to colors perceived by humans. Prior to this, neural representation of such unique hues (Hurvich, 1981) had not been found (Valberg, 2001; Mollon and Jordan, 1997). Note that in the low-level primary visual cortex, the axes defined by cone opponency should more accurately be denoted bluish-red/cyan and lavender/lime opponency (Conway and Livingstone, 2006; Stoughton and Conway, 2008; Derrington et al., 1984), rather than the commonly-termed red-green and blue-yellow opponency.

Finally, the region of the IT where color-selective neurons are found is coarsely retinotopic (Yasuda et al., 2010), meaning that spatial information

is maintained in a coarse manner, rather than completely discarded or maintained with high fidelity.

Overall, these studies are broadly consistent with our proposed "coarse arrangement of categorical color blobs" model of high-level color processing in the primate visual system.

In contrast most computer vision algorithms utilize color information in a relatively low-level manner. The simplest color extension of a non-color algorithm would be to apply it independently to the R, G and B channels, and then concatenate the features from all 3 channels just before the final classifier stage. Most algorithms are variants of this basic idea, either using some other color space, or fusing the channels before the classifier stage (usually at the dictionary or keyword learning stage). For example, SIFT features can be computed separately for each channel in HSV color space (Bosch et al., 2008), while Brown and Susstrunk (2011) do this for RGB space, along with an NIR (near infra-red) channel. Besides SIFT features, other algorithms use (non-orientation based) histograms in the HSV (Tang et al., 2012), Gaussian opponent color (Burghouts and Geusebroek, 2009; Geusebroek et al., 2001), normalized RGB or opponent color spaces (Gevers and Stokman, 2004). A comparison of such variants was done by Van de Sande et al. (2010).

What these algorithms have in common is that in terms of the biology of color vision, they correspond to at most the level of color-opponent cells in the primary visual cortex (also known as V1), the lowest level in the hierarchically-organized visual cortex.

Recently, more sophisticated modeling of single-opponent and double-opponent cells in V1 has shown that adding more biological realism to color descriptors can significantly improve object and scene categorization performance (Zhang et al., 2012). Nonetheless, this improvement is attained with a relatively low-level color machinery.

One notable exception is the approach of learning semantic color names that are used by humans (Van de Weijer and Schmid, 2007; Van de Weijer et al., 2009; Shahbaz Khan et al., 2012). Our approach is different, but not mutually exclusive, and these two approaches are discussed in Section 6.4.

The rest of this chapter is organized as follows: In Section 6.2, we describe our model in more detail. In Section 6.3, experimental results of our model on several datasets are provided followed by the results of concatenating features of our model with other similar models. In Section 6.4, we provide a discussion of the model.

6.2 CQ-HMAX

In this section, we describe our new biologically inspired model, CQ-HMAX (Color Quantization Hierarchical Max) which uses color information in a hierarchical organization of simple and complex cells. The combination of our Max-Mean model (Jalali et al., 2012) with CQ-HMAX is investigated and the final model that encodes both color and shape information is then presented. HMAX is a hierarchical model that uses Gabor filters to find simple and complex shapes in the images. Our model has a similar hierarchical structure. However, we use color quantization cores

and not Gabor filters, hence our model encodes color information. When combined with HMAX and HMean, the overall model includes both color and shape information.

Our color model has a hierarchical structure of simple and complex cells as can be seen in Figure 7.2. We first introduce the model briefly followed by a more detailed description of each layer. An image pyramid is created in YIQ color space. The pyramid has 10 scales, with each neighboring scale different by a ratio of $1/(2^{1/4})$. In order to evaluate the use of color information in our model, we determined that the YIQ color space produced the best results in comparison with HSV and RGB color spaces. (The Y channel represents luminance information and I and Q represent chrominance information). A set of representative values from each color channel is selected as color cores and used to find the best matching unit to each individual pixel value in the pyramid. The $S1$ layer is created on 10 scales indicating the index of the best matching YIQ core to each pixel in the image pyramids. At the $C1$ layer, a local max pooling is computed over $\pm 10\%$ spatial neighborhoods of approximately 6×6 on ± 1 neighbor scales to find the most frequent color core in each neighborhood. A dictionary of features is sampled randomly from the $C1$ layer of images. The distance of each dictionary feature to all patches in a neighborhood of that dictionary feature is calculated to create the $S2$ layer and the best response to each dictionary feature in each image is chosen as the $C2$ layer to be fed to the SVM layer for classification. We describe each layer in more detail below.

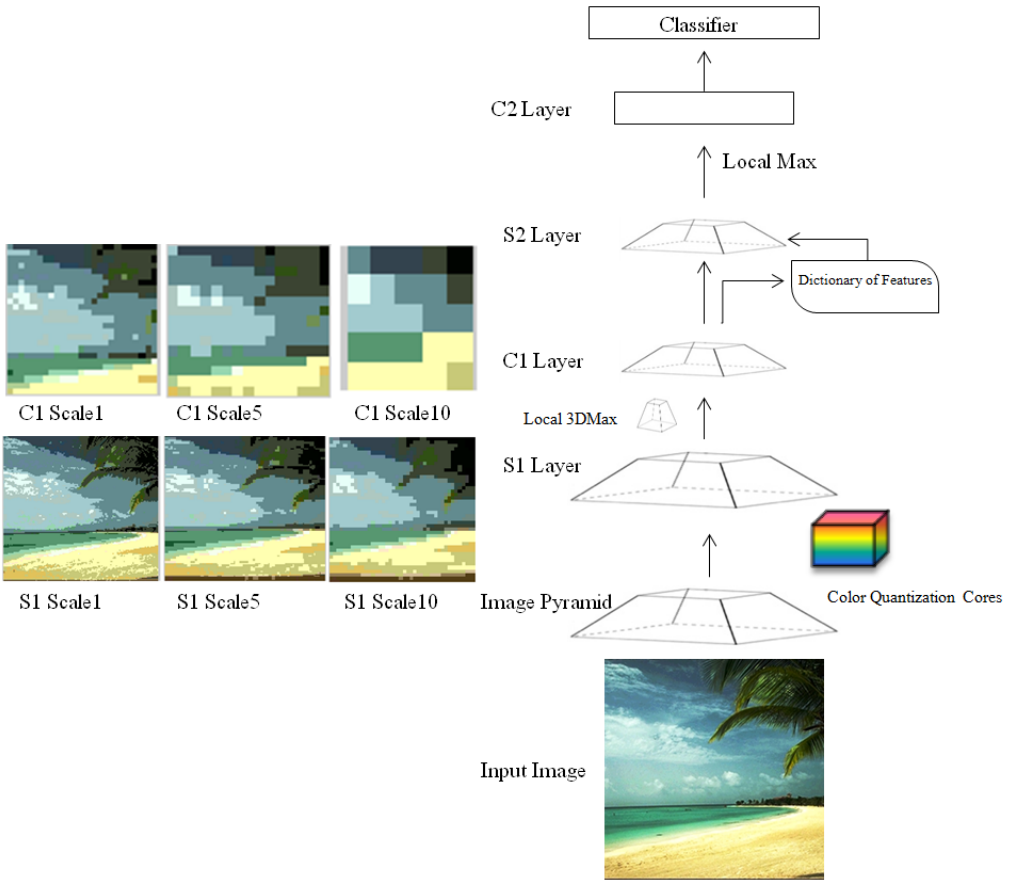


Figure 6.1: The hierarchical structure of CQ-HMAX and an example image of a beach scene in the S1 and C1 layers.

$S1$ Layer and Quantization Cores

The input images are first converted into YIQ color space and a pyramid of 10 scales with a ratio of $2^{1/4}$ is created, with the first scale having the shorter side set to 140 pixels, maintaining the aspect ratio of the original image. This image pyramid is then used as the input to the $S1$ layer. A series of YIQ quantized "color cores" over YIQ channels are created to be used as filters for this layer. We experimented with different numbers of quantization values per color channel, and chose 5 per channel as the optimal number (which results in $5 \times 5 \times 5 = 125$ cores). In order to

choose the optimal cores, 500 images were randomly selected and the color range of these images in YIQ color space was calculated after normalization to the range $[0, 1]$. The values of YIQ channel are mostly in the range $[0, 1]$, $[0.4, 0.7]$ and $[0.4, 0.6]$ respectively. These ranges were selected and divided into 5 bins. The quantized values of Y, I and Q after normalization to $[0, 1]$ were therefore chosen as follows: $Y = (0, 0.25, 0.5, 0.75, 1)$, $I = (0.4, 0.47, 0.55, 0.63, 0.7)$, $Q = (0.4, 0.47, 0.5, 0.53, 0.6)$. Using these values results in better classification performance than using the full range $[0, 1]$ in each YIQ channel. The outputs at the $S1$ layer are the index values (i.e. 1,2,...,125) of the best-matching color core for each element in the image pyramid.

$C1$ Layer

The $C1$ layer provides local invariance to position and scale as it pools nearby $S1$ units, and as a result, subsamples $S1$ to reduce the number of units. The $S1$ pyramid is convolved with a $3D$ max filter to set the $C1$ layer size of the bottom of the pyramid to 25×25 and the highest layer of the pyramid to 5×5 accordingly. The max is calculated over $\pm 10\%$ spatial neighborhood on ± 1 neighbor scales in the middle of the pyramid and -2 on the highest level and $+2$ on the lowest layer of the pyramid (hence it is called a $3D$ max, as it takes the max over $2D$ spatial distribution and over ± 1 scale). This layer provides a model for $V1$ complex cells. Figure 7.2 also shows an example image of $S1$ and $C1$ layers. $S1$ and $C1$ layers have a distribution of quantization cores from coarse to fine. The higher layers

of the $S1$ pyramid are taken from smaller scales of the images in the input pyramid and respectively the higher levels of $C1$ layer are computed by taking a $3D$ max over higher levels of $S1$ layer. As can be seen in Figure 7.2, the higher levels of the pyramid in the $S1$ and $C1$ layers represent less detailed information from the image. All levels in the $C1$ intermediate layer are used for sampling a dictionary of features.

Dictionary of Features and Distance Table

Once the $C1$ layer is created, sampling is performed by centering patches of size 4×4 at random positions and scales using a normalized random number generator function. A distance table is created to store the actual weighted Euclidean distances of the indices from YIQ quantization cores. Since the values of the Y channel are normally distributed between $[0, 1]$, but the values of I and Q channels fall in the approximate range of $[-0.6, +0.6]$ and $[-0.5, +0.5]$ respectively, and as in most of the images the actual values of these two latter channels fall between $[-0.1, +0.2]$ and $[-0.1, +0.1]$ (before normalization to $[0, 1]$) we weighed the distances to have an equal effect in the distance calculation. The distance table weights are calculated as:

$$DistanceTable(i, j) = \sqrt{D(1) + \gamma D(2) + \beta D(3)}$$

$$\text{Where } D(k) = (YIQCore(i, k) - YIQCore(j, k))^2 \tag{6.1}$$

with $\gamma = 3.3$ and $\beta = 5$. In (Jalali et al., 2010) and (Jalali et al., 2012) various clustering methods in the creation of the dictionary of features were

implemented and it is shown that by use of random sampling in HMAX model, relatively good results can be achieved with a lower computational cost in comparison with clustering of features.

$S2$ Layer

Once the dictionary of features and the distance table are created, each entry in the dictionary of features is used as a filter to be convolved on $C1$ patches of size 4×4 on the neighbor scales of the dictionary feature in the pyramid. The responses $V(d, p)$ of each dictionary feature, d to all of the neighbor patches of the same size in ± 1 scale and $\pm 10\%$ in position, p are calculated using a Euclidean distance equation as:

$$V(d, p) = \exp\left(-\frac{\| d - p \|^2}{2\sigma^2\alpha}\right) \tag{6.2}$$

where d is a feature in the dictionary and p is a patch in the image $C1$ pyramid. σ and α are set to 0.5 and 1 respectively as in (Mutch et al., 2010a).

$C2$ Layer

Once the $S2$ layer is generated, the maximum values for each patch in the dictionary are taken as the $C2$ output. This layer outputs a vector of the same size as the dictionary of features. We chose different sizes for the dictionary of features and in most cases a dictionary of size 10000 was chosen which results in slightly better performances than smaller sizes of about 1000 dimensions.

Classification Layer

The $C2$ vectors are classified using a multi-class one-vs-rest linear kernel support vector machine. The algorithm used to train the classifier is weighted regularized least-squares after the data is sphered and the mean and variance of each dimension are normalized to zero and one respectively as in (Mutch and Lowe, 2008).

Use of HMAX and HMean for Encoding Shape Information

In order to implement the use of shape information, we use the HMAX model presented in (Mutch and Lowe, 2008) with the code provided in (Mutch et al., 2010a) and our HMean model presented in (Jalali et al., 2012) as described in detail in Section 5.3. The final model that encodes all shape and color information, is shown in Figure 6.2.

When using HMAX and HMean model, we have used different parameters for different datasets to achieve the best performance using shape information following (Zhang et al., 2012). However, the parameters used for CQ-HMAX are uniform in all datasets. In Section 6.3 an extensive set of experiments are provided and these methods are explored and compared in classification tasks over several datasets.

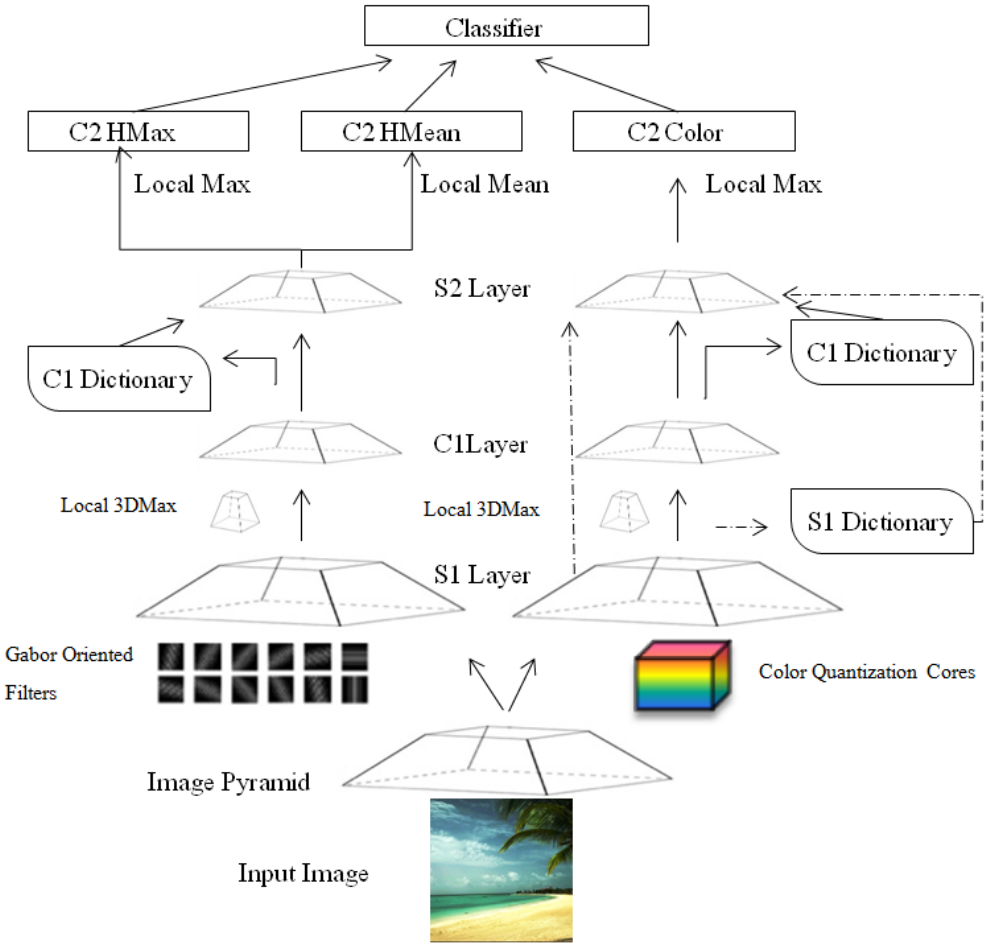


Figure 6.2: The overall model using both shape and color information. Dotted lines represent an extension in which $C1$ layer is eliminated and $S1$ information are directly used to create a dictionary of features and to calculate $S2$ and $C2$ features.

6.3 Experimental Results

First we examine the naïve use of color by computing various color spaces (RGB, HSV, YIQ) on the Caltech101 dataset (Li et al., 2004) and compare the results with grayscale images. The Caltech 101 dataset, includes 101 classes of objects plus a background category. Each class con-

tains between 31 to 800 color images of different sizes. The size of each image is approximately 300×200 pixels on average. We used 30 randomly chosen images for training from each class and the rest of the images were used in the test phase. Some sample images of this dataset are shown in Figure 5.7a. We first divide the images into three channels and feed them to the unmodified HMAX (Mutch and Lowe, 2008) directly and evaluate the classification performance.

Color Component	**Performance**
Y channel (i.e. gray scale)	54.65% ±1.2
I channel	35.20% ±2.1
Q channel	26.86% ±3.2
YIQ channels concatenated	**55.06**% ±1.0
RGB channels concatenated	26.53% ±4.3
HSV channels concatenated	31.32% ±5.7

Table 6.1: Naïve use of various color channels and color spaces.

As shown in Table 6.1, the use of three different channels and concatenating the $C2$ vectors of all channels to the SVM does not provide any significant improvement. Hence, we explore the use of color using the CQ-HMAX model described in detail in Section 6.2. In the rest of this chapter, we evaluate our model on four datasets: Caltech101, 8 Scenes, 17 Flowers and Soccer.

Caltech101 Dataset

The results of using CQ-HMAX on Caltech 101 are shown in Table 6.2. All experiments are performed 8 times on random splits of training and test sets and the average performance is reported. As can be seen, the use of our color model in this dataset does not outperform the HMAX performance. However, when the $C2$ features of the color model are concatenated with $C2$ features of HMAX and HMean models, the classification results outperforms the existing state of the art performances on biologically inspired approach of HMAX models and results in an approximately 7% improvement on HMAX and about $9-10\%$ when used with both HMAX and HMean $C2$ vectors and the final model which uses both shape and color information has the best performance. HMAX is a computationally expensive model as Gabor filter responses over different orientations in $S1$ layer are calculated. However, CQ-HMAX is relatively faster than HMAX as it performs a quantization with 125 cores in the $S1$ layer instead of Gabor filters. Adding HMean also does not add much computational costs as it uses the S2 responses calculated in HMAX model.

8 Scenes Dataset

This dataset is introduced in 5.5.1.

As can be seen in Table 6.2, the use of HMAX and HMean significantly improves classification performance when concatenated with color and it outperforms the use of GIST algorithm by about 3% proposed in

Model	**Caltech101**	**8Scenes**	**17Flowers**	**Soccer**
HMAX (i.e shape)	54.65% ±1.2	71.48% ±2.1	42.54% ±3.7	24.76% ±4.2
CQ-HMAX (i.e. color)	38.11% ±2.1	69.21% ±1.2	77.64% ±0.8	**77.14**% ±0.9
CQ-HMAX + HMAX	61.09% ±1.7	78.97% ±1.3	69.21% ±2.1	66.67% ±3.1
CQ-HMAX + HMAX + HMean	**64.39**% ±0.6	**86.54**% ±0.9	**78.31**% ±0.6	71.42% ±1.2

Table 6.2: Experimental results of the use of CQ-HMAX color model in concatentation with HMAX and HMean on Caltech101, 8 Scenes, 17 Flowers and Soccer datasets.

(Oliva and Torralba, 2001). Sample images of this dataset are shown in Figure 5.7b. The classification results achieved using a combination of CQ-HMAX, HMAX and HMean are as good as the state-of-the-art performance of 87.1% in (Zhang et al., 2012).

Flowers Dataset

The 17-category Flowers dataset (Nilsback and Zisserman, 2006) is introduced in Chapter 5.5.1.

The classification accuracy in different classes ranges from 43.33% (for category 1) to 100% (for category 2). As can be seen in Figure 6.3, the distribution of the color cores in category 1 (Figure 6.3a) and the average

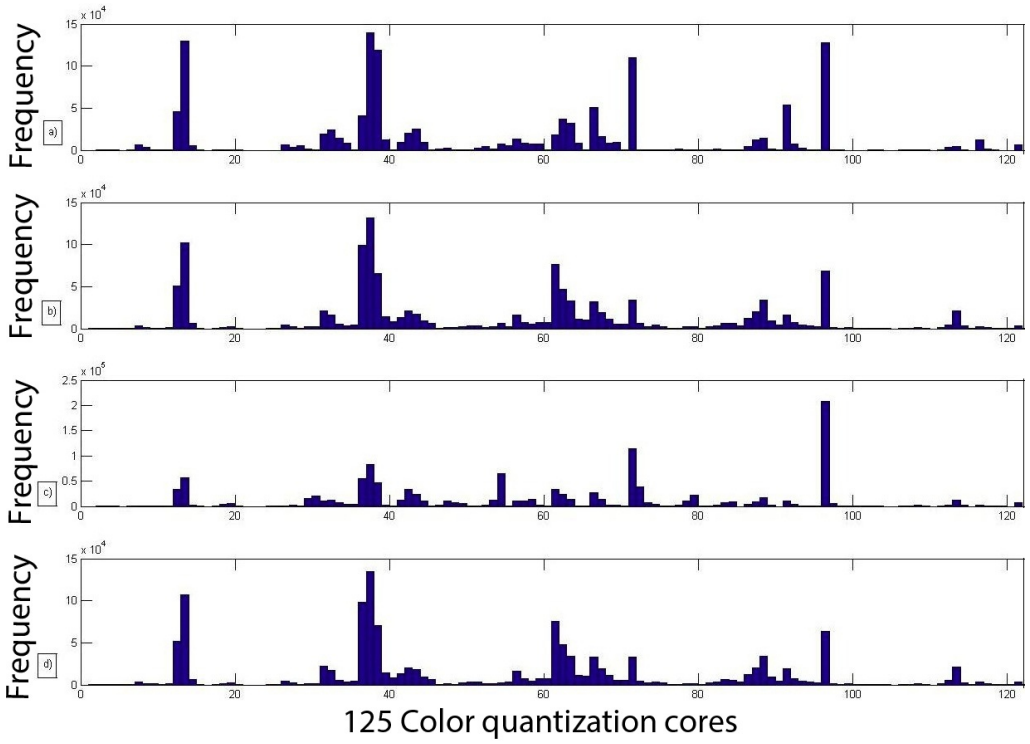


Figure 6.3: Histograms of color cores using a one-vs.-rest classification scheme in Flowers dataset. Accuracy for categories 1 and 2 are 43.3% and 100% respectively. a. Category 1. b. Average of all categories except category 1. c. Category 2. d. Average of all categories except category 2.

of all classes except category 1 (Figure 6.3b) are quite similar while the color distribution for category 2 (Figure 6.3c) and the average of all classes except category 2 (Figure 6.3d) are quite different. This suggests that our hard-coded quantization scheme should be further optimized in a dataset and class-specific manner. Further discussion is provided in Section 6.4. Our classification results are as good as using hue/SIFT model of (Zhang et al., 2012) when using both shape and color information but not as good as 83% of their SODO-HMAX.

Soccer Dataset

The Soccer team data set is introduced in Chapter 5.5.1.

As can be seen in Table 6.2, the use of CQ-HMAX model provides significant improvements over using shape based HMAX model. Since the images from different classes share similar shapes, using color results in better performance than shape.

Underwater Images Dataset

We also evaluated CQ-HMAX on the Underwater Images dataset Jalali et al. (2013c). This dataset is made of 1664 images of around 740 x 420 pixels from 13 different categories and sample images are shown in Figure 5.8. We used 30 randomly selected images per category for training and the rest for testing. These underwater images contain small objects of various shapes and color against a varied seabed background. The main challenge with these images is in light absorption by the water, and the existence of particles that limit visibility and result in scattering and reflection of light. In this experiment, we created a set of images using both grayscale and color cameras and compared the performance of CQ-HMAX on color images and HMAX on grayscale images. The classification accuracy on this dataset using HMAX is 92.93 %, CQ-HMAX is 94.03 % and combination of HMAX and CQ-HMAX results in 96.23 %.

Direct Use of $S1$ Features

As can be seen in Figure 6.2 (the final model), we have also experimented using only $S1$ features directly which makes the model simpler and faster but does not result in better classification performances over all datasets. In this extension, $S1$ features are used directly (the $C1$ layer is eliminated) and the dictionary of features is created by randomly sampling from $S1$ features and the $S2$ and $C2$ layers are created using these features. Since the $S1$ features are more selective and $C1$ features provide more invariance, the performance of $S1$ layer is slightly better than $C1$ features in datasets that fine-grained discrimination among categories with relatively similar shapes and colors such as Flowers. However, in datasets such as Scenes and Caltech101, the $C1$ performance is better than the $S1$ level. The performance of $S1$ and $C1$ models is equally good in Soccer dataset.

6.4 Discussions

(Zhang et al., 2012) proposed a new biologically inspired color descriptor that encodes color information in a low-level manner. In their model, they create 8 channels of opponent colors: R^+G^-, R^-G^+, R^+C^-, R^-C^+, Y^+B^-, Y^-B^+, Wh, Bl and used these channels to calculate the Gabor filters on different orientations and used them to create Single-Opponent and Double-Opponent channels. We explored the use of these color channels as inputs to HMAX and CQ-HMAX on Soccer and Flowers dataset and but use of these color channels as input to HMAX did not per se result in

any significant improvement to the classification results. The first row of results in Table 6.3 show the classification accuracy over different datasets using HMAX. Using CQ-HMAX vectors (second row) along with HMAX $C2$ vectors resulted in a better performance, but below the state-of-the-art of the (Zhang et al., 2012) as can be seen in the third row. In order to evaluate the combination of the SODO-HMAX model of (Zhang et al., 2012) with our model, which is a high-level color model, we concatenated their SODO-HMAX features with our CQ-HMAX features. In SODO-HMAX, Single-Opponent features encode color regions and Double-Opponent features encode color edges. The last row in Table 6.3 shows the performance of concatenating the $C2$ features from CQ-HMAX with SODO-HMAX features from (Zhang et al., 2012) which results in the best performance using bottom-up approaches on Soccer and Flowers dataset.

This significant improvement is encouraging and this motivates us to evaluate different combinations of these two models (not only concatenating features, but to merge these two models in a more principled manner) over other datasets which will be further evaluated in future work. The use of SO features for Soccer dataset in this experiment resulted in a better performance (2%) than use of $SODO$ features as Soccer dataset is more sensitive to color information and the addition of more shape information results in lower performance (e.g. similar patterns on players' shirts but in different colors). The use of SODO-HMAX features for Flowers dataset resulted in better performance (3%) in-line with the results in Table 6.2 where the use of shape and color works better than using each one individ-

ually.

Model	**Soccer**	**Flowers**
HMAX on RGYBRCWB	53.33% ±2.2	62.14% ±1.3
CQ-HMAX	77.14% ±0.9	77.64% ±0.8
SODO-HMAX (Zhang et al., 2012)	87.61% ±1.5	83.13% ±1.2
CQ-HMAX + SODO-HMAX	**93.33**% ±0.9	**90.14**% ±0.3

Table 6.3: Classification accuracy on the Soccer and Flowers datasets using different color channels and Single Opponent and Double Opponent features of (Zhang et al., 2012).

As shown in Figure 7.2, the $S1$ and $C1$ layers resemble a segmentation of the images. The use of these layers in the middle layers along with bottom-up and top-down interactions is a prospective extension to this model. In this extension, we will compute a set of clusters based on the $S1$ and $C1$ layers and use these clusters to confirm the similarity of an image to a category cluster after the SVM has classified an image (to double confirm the classifier output) which adds top-down, bottom-up interactions in the model. Further details of this method will be explored in future work.

As explained in Section 6.3, when the colors in different classes are similar, a lower classification accuracy is achieved. In order to optimize the color core selection, a learning system can also be used in which color cores are defined based on an unsupervised clustering in which more frequent colors in each dataset are chosen as color cores. This will be further explored

in future work.

Currently, our model quantizes the YIQ color space into arbitrarily-spaced cubed-shaped "color cores" at the $S1$ layer. Following the work of Shahbaz Khan et al. (2012) and Van De Weijer and Schmid (2006), learning the color values that correspond to semantic color names such as "orange", "brown", could also further improve performance.

In this chapter, we introduced a new biologically inspired approach to image classification that uses color in a manner consistent with high-level visual cortex by incorporating insights from cognitive psychology and neuroscience. We implemented the use of Max and Mean pooling operators and color information and showed that the use of color on some datasets outperforms shape information significantly and on some other datasets it helps achieve a better performance when added to shape information. We ran this model on several datasets such as Caltech101, Soccer, Flowers and Outdoor Scenes. The combination of our color features with (grayscale) shape features leads to double-digit average increases in performance over shape features alone. Using our model, performance is significantly higher than using color naively, i.e. concatenating the channels of various color spaces. Among approaches that use bottom-up information only, the combination of three sets of biologically-inspired features (our high-level color features with existing low-level color features of Zhang et al. (2012) and HMAX grayscale shape features of Mutch and Lowe (2008)) achieves the best performance to date on the Soccer and Flowers datasets.

Chapter 7

Applications of Proposed HMAX and CQ-HMAX Models

In this chapter we introduce a few of the relevant applications of our modified HMAX model, which includes the use of modified HMAX introduced in Chapter 5 and the use of CQ-HMAX introduced in Chapter 6. In this chapter, we use HMAX model for detecting mitosis in histopatholgy images and compare the performance of our modified HMAX and CQ-HMAX model with SIFT method.

[0]A part of the models and experiments presented in this chapter are published in workshop on Histopathology Image Analysis (HIMA), MICCAI 2012 (Humayun et al., 2012) and published in Journal of Histopathology Image Analysis (Humayun et al., 2013). Other parts are accepted as two different publications in proceedings of the International Joint Conference on Neural Networks IJCNN 2013 (Jalali et al., 2013b,a).

7.1 Automated Mitosis Detection Using Texture, SIFT Features and HMAX Biologically Inspired Approach

Researchers in histopathology appreciate the importance of qualitative analysis of histopathological images. These analyses are used to confirm the presence or the absence of disease and also to help in the evaluation of disease progression. Being important in diagnostic pathology, this qualitative assessment is also used to understand the realities for specific diagnostic being rendered like specific chromatin texture in the cancerous nuclei, which may indicate certain genetic abnormalities. In addition, quantitative characterization of pathology imagery is important not only for clinical applications (e.g., to reduce/eliminate inter- and intra-observer variations in diagnosis) but also for research applications (e.g., to understand the biological mechanisms of the disease process (Gurcan et al., 2009)).

Co-occurrence features, run-length features and SIFT features were extracted and used in the classification of mitosis. We evaluate the performance of the proposed framework using the modified biologically inspired model of HMAX and com-pare the results with other feature extraction methods such as dense SIFT.

7.1.1 Introduction

Nottingham Grading System (Bloom and Richardson, 1957) is an international grading system for breast cancer recommended by the World

Health Organization. It is derived from the assessment of three morphological features: tubule formation, nuclear pleomorphism and mitotic count. Several studies on automatic tools to process digitized slides have been reported focusing mainly on nuclei or tubule detection.

Mitosis detection has diagnostic significance for some cancerous conditions. Indeed, mitotic count provides clues to estimate the proliferation and the aggressiveness of the tumor (Elston and Ellis, 2002) and is a critical step in histological grading of several types of cancer. In clinical practice, the pathologists examine proliferated area and determine mitotic count after a tedious microscopic examination of hematoxylin and eosin (H& E) stained tissue slides at high magnification, usually $40X$. The area visible in the microscope under a $40X$ magnification lens is called a high power field (HPF). This mitotic counting process is cumbersome and often subject to sampling bias due to massive histological images. This results in considerable inter- and intra-reader variation of up to 20% between central and institutional reviewers in tumor prognosis (Teot et al., 2007). In histopathological image analysis, the accuracy of mitosis detection is crucial in order to identify the severity of the disease.

Mitosis detection is a difficult task having to cope with several challenges such as irregular shaped object, artifacts and unwanted objects because of slide preparation and acquisition. Mitosis has four main phases and each phase has different shape and texture. It is also observed that artifacts produce objects which look similar to mitosis. As a result, there is no simple way to detect mitosis based on shape and pixels values. However,

the major problem is the very low density of mitosis in a single HPF. It is not unusual to have an HPF without any mitosis.

7.1.2 Framework

We propose a color image processing-based strategy for mitosis detection in H& E images. The aim is to improve the accuracy of mitosis detection by integrating the color channels that better capture the texture features which discriminate mitosis from other objects. Two main stages are involved in the proposed methods as shown in Figure 7.1. In the first stage, we perform detection of candidate mitosis. The input RGB images are transformed into blue-ratio images (Chang et al., 2012). We perform Laplacian of Gaussian (LoG), thresholding and morphological operations on blue-ratio images as in Chang et al. (2012) to generate candidate mitosis regions.

In the second stage, we compute co-occurrence features, run-length features and SIFT features for each region, and select those features having better discrimination of mitosis regions from others. Finally a classification is performed to put the candidate region either in the mitosis class or in the non-mitosis class. Three different classifiers have been evaluated: decision tree, linear and non-linear kernel SVM. We also evaluate the performance of the proposed framework using the modified biologically inspired model of HMAX and compare the results. Modifications made to the original HMAX model for this experiment, include removal of $S2$ inhibition in Mutch and Lowe (2008), Calculating the max in $C2$ layer over a ± 1 in scale and $\pm 10\%$.

However once the features fall on the max or min scales or are close to the borders, the ±2 scale and/or ±20% position invariance is considered.

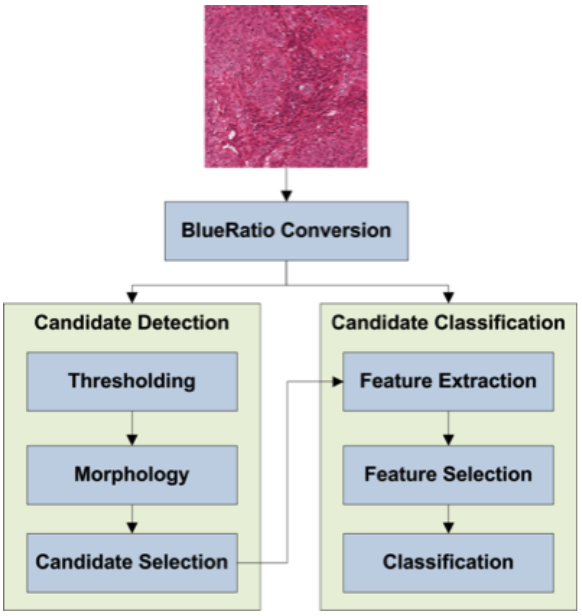


Figure 7.1: Framework for mitosis detection.

7.1.3 Experimental Results

We evaluated the proposed framework on MITOS dataset (http://ipal.cnrs.fr/ICPR2012, 2012), a freely available mitosis dataset. This dataset consists of 35 HPF images at $40X$ magnification. A HPF has a size of 512×512 m^2 (that is an area of 0.262 mm^2), which is the equivalent of a microscope field diameter of 0.58 mm. Each HPF has a digital resolution of 2084×2084 pixels. These 35 HPFs contain a total of 226 mitosis. The pathologists have annotated mitosis manually in each HPF images. 25 HPFs containing 154 mitosis will be used for training purpose, the remaining 10 HPFs containing 72 mitosis being used for testing.

A comparison of all different classification methods is presented in Table

7.1. One of the parameters that affect our experiments is existence of no balance between the number of mitosis and non-mitosis regions. When we used this dataset for training the classifier, then most of the classifiers are biased toward non-mitosis which resulted high number of false positives.

Method	**TP**	**FP**	**FN**	**TPR**	**PPV**	**F-Measure**
Texture with linear SVM	183	636	43	0.81	0.22	0.35
Texture with non-linear SVM	174	358	52	0.77	0.33	0.46
Texture with Random Forest (Tree)	185	**47**	42	0.82	**0.80**	**0.81**
SIFT with SVM	203	647	23	0.90	0.24	0.38
HMAX	205	151	21	0.91	0.57	0.71
HMAX (generative features)	213	171	**13**	0.94	0.56	0.70
CQ-HMAX	217	92	2	**0.96**	0.63	0.76

Table 7.1: Results of different Classifiers (Ground Truth = 226).

7.1.4 Discussion

In the first method, we used linear and non-linear SVM and random forest classifier on texture features. As compared with linear kernel, the experiments with non-linear kernel resulted in better performances in terms of less false positives but less true positives as well. When we used selected texture features with random forest, an ensemble classifier consisting of many decision trees, we achieved classification with low false positives and high PPV and f-measure. The random forest classifier has better results as

compared to other classifiers because of balancing error in class population unbalanced datasets. [1]

SIFT features are also examined in this study, but due to the lack of balance between number of mitosis and non-mitosis regions, the SIFT method does not perform as good as other methods. As can be seen in Table 7.1, we have also used HMAX model to train a dictionary of features from local max on Gabor filter responses over 12 orientations which resulted in high true positives but high false positives as well. The dimensionality of features in HMAX model is directly related to the size of the dictionary of features and we evaluated different sizes over several runs and used the optimum numbers. A global dictionary of features from generative images (Caltech101) was also used in another experiment to evaluate the performance of different dictionaries on these images and interestingly achieved almost the same results. It is because of the nature of this model in which the statistics of natural images are en-coded. However, using a non-linear kernel for SIFT and HMAX, in which the features dimensions are high, (order of 10000) results in over-fitting which resulted in lower classification accuracy. We also used the RGB images and fed them to the CQ-HMAX structure described in Chapter 6 and very high classification results were achieved. However since the PPV and F-Measure are directly related to the number of FP and FN, the final accuracy of these methods

[1]The material in this section, regarding introduction of mitosis dataset, image segmentation and selecting candidate patches are carried out by Humayun Irshad, in collaboration with Image and Pervasive Access Lab (IPAL) and different classification methods such as SVM, SIFT, HMAX and CQ-HMAX models are carried out by Sepehr Jalali.

is not the best. In a new dataset of color images extracted from mitosis dataset, HMAX accuracy is about 10% where CQ-HMAX outperforms HMAX significantly and results in classification accuracy of 30%. Further investigation of CQ-HMAX model is in progress.

7.2 Classification of Marine Organisms in Underwater Images using CQ-HMAX

In many coastal environments, particularly in tropical zones, coral reef ecosystems have exceptional biodiversity, contribute to coastal defense, provide unique and important habitats and valuable commercial resources. Assessment of environmental impacts on biodiversity in such areas are increasingly important to mitigate potential adverse effects on specific ecosystems. Visual classification of marine organisms is necessary for population estimates of individual species of corals or other benthic organisms. In this chapter, we introduce a new image dataset of benthic organisms that are of different colors, shapes, scales, visibility and are taken from different viewpoints. We evaluate several different classification approaches on this dataset, and show that CQ-HMAX, results in better classification results in comparison with existing computational models such as support vectors machines, SIFT based approaches and the HMAX biologically inspired approach. We show that concatenating our model which encodes color information with the HMAX model which encodes grayscale shape information results in the highest classification accuracy.

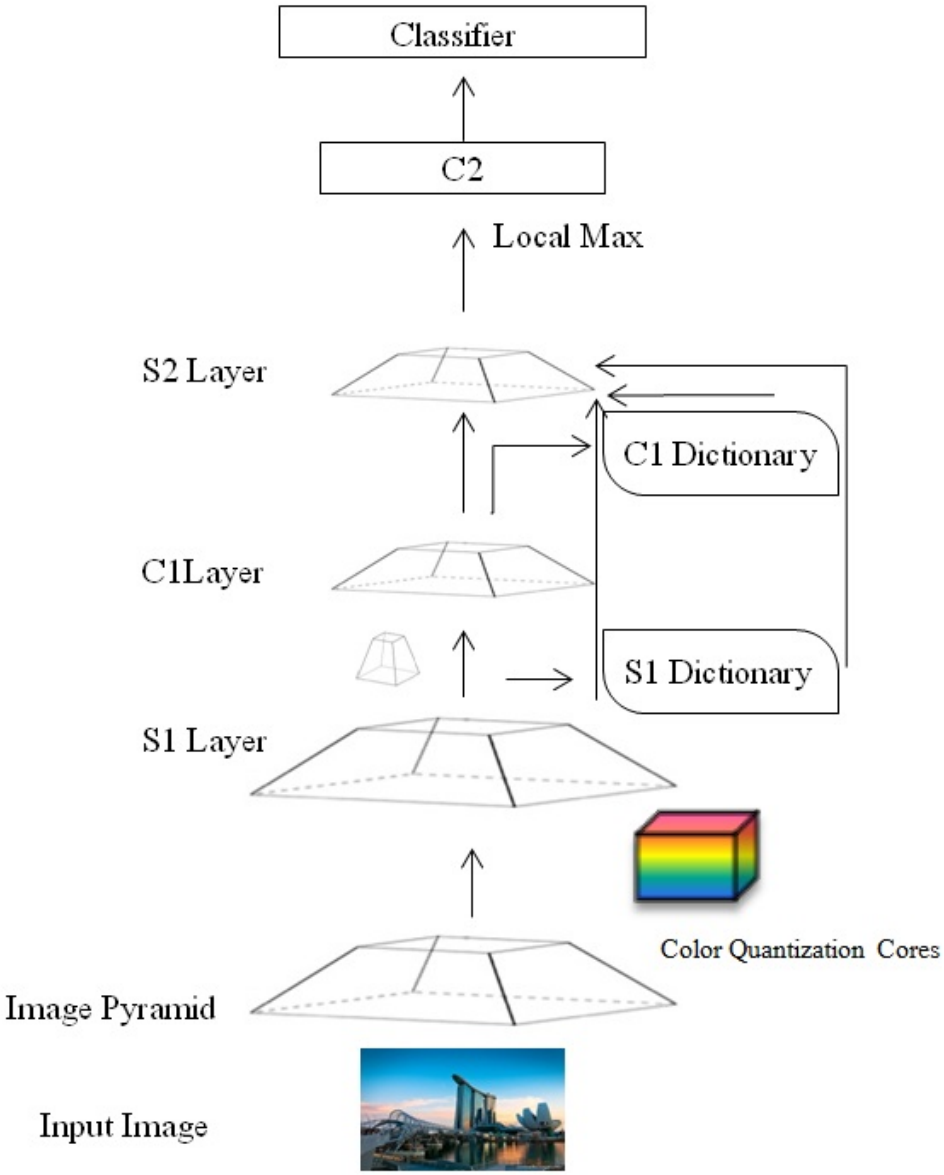


Figure 7.2: The hierarchical structure of integrated HMAX and CQ-HMAX models.

Our final model is based on the concatenation of CQ-HMAX and HMAX models as shown in Fig. 7.2.

7.2.1 SIFT Features

Scale Invariant Feature Transform (SIFT) feature extraction method (Lowe, 1999) is a well-known method which has produced promising results in classification tasks. Here we investigate its application in classification of marine organisms. In SIFT methods, a series of features are calculated using difference of Gaussian (DoG) methods over different scales. Once a set of features are selected, features from new images are compared with these candidate regions using their Euclidean distance and from the full set of matches. A subset of key point features which agree on the object, its scale, orientation and location in the new image, are identified to filter out good matches. Finally a histogram of features is calculated and the final histograms are sent to a SVM classifier. In this experiment we use PHOW features (dense multi-scale SIFT descriptors (Bosch et al., 2007)), Elkan k-means for fast visual word dictionary construction, spatial histograms as image descriptors, a homogeneous kernel map to transform a Chi^2 support vector machine (SVM) into a linear one and finally an internal SVM for classification using VLFeat toolbox Vedaldi and Fulkerson (2010).

7.2.2 Marine Organisms Dataset and Experimental Results

The marine benthic organisms dataset includes 19 classes of marine organisms that grow on or are closely related to the benthos (the seabed). Each class contains between 60 to 300 color images of different sizes. The size of each image is approximately 5000×3000 pixels. However, due to

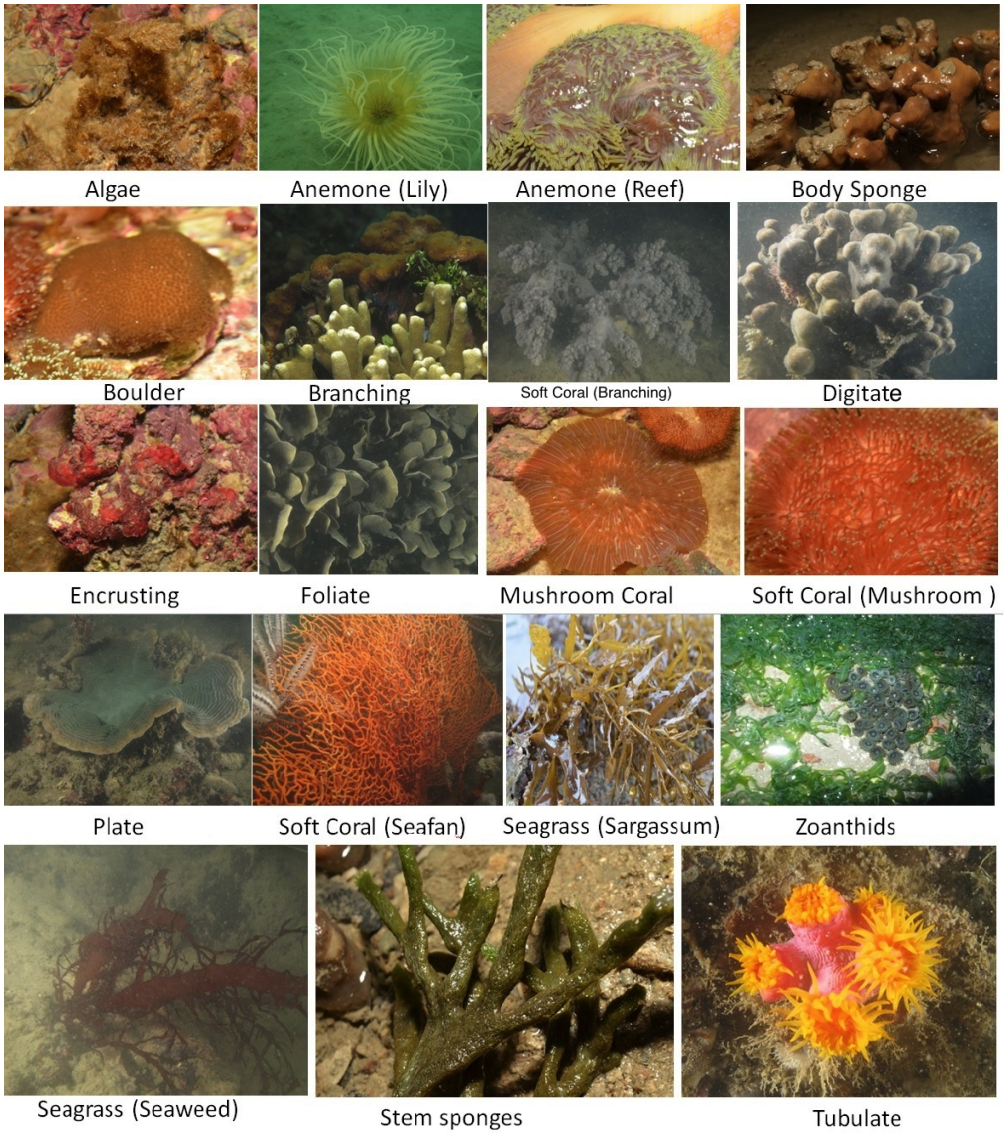


Figure 7.3: Sample images from the marine organisms dataset.

high computational costs, images are resized to approximately 500×300 pixels, while maintaining the aspect ratio. We used 30 randomly selected images for training from each class and the rest of the images were used in the test phase. Sample images of every class of this dataset are shown in Figure 7.3.

Many benthic marine organisms have several distinguishing factors which set them apart from each other. Some visual characteristics of some of the classes used are as follows: boulder or submassive corals are easily differentiable from others by their roughly spherical shape which is similar in all dimensions except the base which is flattened. The foliate organisms have a leaf-like appearance with folded plates or spires extending upwards. Branching corals have an outward growth of branches which have primary and secondary branchings, unlike digitate forms which do not have secondary branches. The plate-like corals have laminar and flattened sheets which may be vertical or horizontal. Mushroom soft corals have a unique appearance of a flat uneven circle or oval. Anemones have a single body with tentacles radiating in all directions. With all these distinguishing factors and more, the model recognizes and is able to differentiate them, hence providing us with useful identifications. In this dataset, we have 19 categories (and sub-categories) of benthic marine organisms: 1- Algae, 2- Anemone (Lily), 3- Anemone (Reef), 4- Body Sponge, 5- Boulder, 6- Branching, 7- Branching (Soft), 8- Digitate, 9- Encrusting, 10- Foliate, 11- Mushroom Coral, 12- Mushroom (Soft Coral), 13- Plate, 14- Seafan (Soft Coral), 15- Seagrass (Sargassum), 16- Zoanthids, 17- Seagrass (Seaweed),

18- Stem Sponges, 19- Tubulate.

We evaluated different classification methods on this dataset such as naiive use of SVM in which images are resized to the size 160×90 pixels and sent to a linear kernel classifier (in all HMAX and CQ-HMAX experiments, image resolution is reduced to $140 \times Si$ where Si is dependent on the aspect ratio). In other experiments, we used HMAX and SIFT methods. As can be seen in Table 7.2, the use of CQ-HMAX model provides significant improvements over using shape based HMAX model and when this model is concatenated with HMAX model, the highest accuracy is achieved. Feeding images directly to a support vector machine results in a very low classification accuracy of 20.5 %. When HMAX model is applied on this dataset, and scale invariant features are extracted, a boost in classification is achieved to enhance it to 40%. A SIFT based method results in about 52 % accuracy on this dataset. The use of CQ-HMAX results in better classification accuracy than SIFT and HMAX and reaches 56.2 %. Concatenation of $C2$ vectors of HMAX model (shape features) with CQ-HMAX model (color features) results in the highest classification accuracy of 61.2 % on this dataset. These results are inline with the previous experiments carried out on the CQ-HMAX model in Jalali et al. (2013d) where concatenation of HMAX and CQ-HMAX models results in a better classification accuracy in several datasets such as Flowers, Soccer, Caltech101 and Scenes. In this section, concatenation of CQ-HMAX with SODO-HMAX model of Zhang et al. (2012) results in classification accuracy better than the state of the art in Soccer (Van De Weijer and Schmid,

2006) and Flowers (Nilsback and Zisserman, 2006) datasets.

Classification Model	**Performance**
SVM	20.54 ± 1.8%
HMAX (Grayscale)	39.61 ± 1.2%
HMAX (Red Channel)	40.94 ± 1.6%
HMAX (Green Channel)	39.92 ± 2.3%
HMAX (Blue Channel)	41.13 ± 1.7%
HMAX (RGB Channels)	48.22 ± 1.8%
SIFT	52.11 ± 1.1%
CQ-HMAX	56.23 ± 0.5%
HMAX + CQ-HMAX	**61.18 ± 0.7**%

Table 7.2: Classification accuracy on the marine benthic organisms dataset using different methods.

7.2.3 Discussion

Using color information alone, we could achieve a higher classification accuracy than using shape information alone. Combining color and shape information in this classification task results in the best performances achieved. As can be seen in Figure 7.4, CQ-HMAX (red bars) outperforms HMAX model (blue bars) in almost all classes.

One of the classes in which CQ-HMAX significantly improves over HMAX are class 17 and 18 (Seagrass/ seaweed and Stem Sponges) where

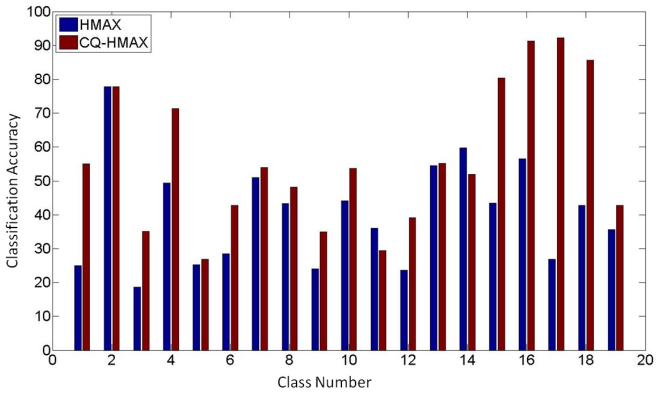


Figure 7.4: Comparison of HMAX and CQ-HMAX classification accuracy.

images have similar colors, but are of different viewpoints, scales and orientations. A few samples of this class are shown in Figure 7.5a,c. On the other hand, in classification of images of Seafan category (Class 14) shown in Figure 7.5b, HMAX performs slightly better than CQ-HMAX and this is due to the variety of colors in images of Seafan category and the consistency of oriented edges. In most of the other classes, CQ-HMAX performs as well or better than HMAX. Class 2 (Lily Anemone) has the highest classification accuracy in HMAX model, and the accuracy is equally as good as CQ-HMAX. This is due to consistency in the oriented bars, and having enough training samples from different colors of this class.

As shown in Table 7.2, the use of SVM on the raw pixels does not result in a high classification accuracy and this is due to the different intraclass viewpoints, scales and varieties in the colors. Hence a model that is more invariant to the viewpoints and scales would match this dataset better. SIFT based methods and HMAX model, both provide invariance to scale and position of the features, however they are also selective to the intra-class variations when the orientations of the edges are very random among

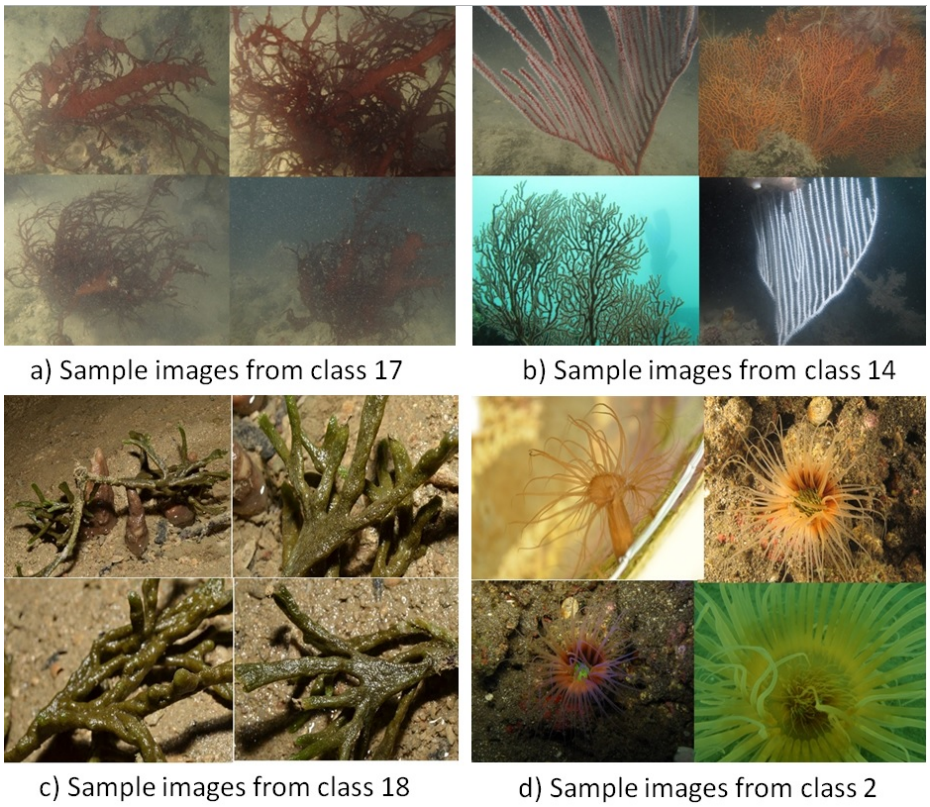


Figure 7.5: Sample images from different classes to compare the classification accuracy of HMAX and CQ-HMAX. a) Seagrass (Seaweed) where CQ-HMAX significantly outperforms HMAX. b) Seafan soft coral, where HMAX has a slightly higher classification accuracy than CQ-HMAX. c) Stem Sponges, where CQ-HMAX significantly outperforms HMAX. d)Lily Anemone, where HMAX and CQ-HMAX have equal classification accuracy.

images in the same class.

Since this is a dataset of live organisms which generally do not have a firm rigid structure, and the images are taken from different viewpoints, these models do not provide the highest classification accuracy. On the other hand, CQ-HMAX model is very invariant to changes in the rotations of the edges in the images and encodes the color information which is an important characteristic in this model. As Table 7.2 shows, the use of R, G and B channels separately does not result in any significant improvement but their combination results in a better classification accuracy which is below the CQ-HMAX model as HMAX model is more orientation based. As Figure 7.4 shows, the classification accuracy achieved with HMAX model is highest in Class 2 (Fig. 7.5d) where the orientation of lines in the images are consistent.

Another interesting characteristic of these two models is that despite having a similar accuracy in many classes, concatenation of the two models results in a better classification.

One of the future applications of this model is to enable an automatic system for classification of marine organisms underwater in certain areas to investigate their abundance. Since the classification accuracy achieved with our model is very reasonable (about 61%) and in real life situations the number of classes in a specific area are often fewer than the classes covered in this dataset, this accuracy will probably improve and hence this model could be used for segmenting the images and classifying the marine organisms into broad classes.

7.3 The Use of Optical and Sonar Images in the Human and Dolphin Brain for Image Classification

In this section we propose a new biologically inspired model which simulates the visual pathways in the human brain used for classification of matching optical and sonar derived images. Marine mammals, such as dolphins, that live in waters with poor optical clarity and low light levels such as littoral zones, use a combination of optical vision and biosonar to navigate and hunt for prey. Given that dolphins have evolved a synergistic combination of optical visual input and acoustic/sonar input, the primary focus of this section is on reaching a similar level of synergy for a diver or Autonomous Underwater Vehicle (AUV) platform equipped with a system to extend the range and resolution of vision in poor ambient visibility. We propose a biologically inspired model that combines and processes visual images acquired via optical and acoustic pathways and show that the combined model enhances the accuracy of automatic classification of target objects in underwater images.

7.3.1 Similarities between Auditory and Visual System in Mammals

In this section, we review the similarities between the organization of the auditory system and the visual system in mammals. While there are im-

portant differences between the two systems (King and Nelken, 2009), there are also important broad similarities, such as hierarchical organization, organization into parallel streams and topographic mapping (Rauschecker, 1998; Rauschecker and Scott, 2009). In this section, we take the view that these gross similarities may be sufficient for a model of visual processing to be also used for auditory processing. In the case of sonar, actual pixel images can be produced (using techniques described elsewhere in this section). As such, using the HMAX model of visual processing on such "auditory" data is not as outlandish as it sounds.

Broadly, the auditory system seems to be geared toward producing a semantic auditory scene from raw sound intensities, similar to the way the goal of visual processing is to perform semantic scene analysis from raw pixels. Like the visual system, the auditory system is hierarchically organized (Okada et al., 2010; Chevillet et al., 2011; Talkington et al., 2012). In addition, the system is organized into two separate, but interacting, streams (Romanski and Averbeck, 2009). The 'where' stream handles spatial processing, while the 'what' stream handles non-spatial processing (Rauschecker and Tian, 2000; Kuśmierek et al., 2012). In both cases, hierarchical processing in the 'what' stream gradually produces outputs at the higher levels that correspond to semantic features (e.g. objects) (Nelken, 2004) that are invariant to low-level sources of variation such as object location and intensity. Analogous to the increase in spatial receptive field size up the hierarchy of the visual system, the temporal receptive field size similarly increases up the hierarchy of the auditory system (Lerner et al.,

2011).

Interestingly, there is also some evidence that at the level of individual neurons and small scale neuronal circuitry, both systems could in fact be implementing similar computations. For ferrets in which outputs of the retina are rewired to feed to the primary auditory cortex (instead of primary visual cortex), these rewired auditory cells develop a number of properties found in primary visual cortex (Roe et al., 1990, 1992; Sharma et al., 2000; Sur and Leamey, 2001). These results suggest that the same basic computations underlie processing in both the visual and auditory systems, and the key differences could really just be due to the difference in inputs.

Our biologically inspired model of sonar and image classification is similar to the one described in Mutch and Lowe (2008). This model which is an extension of HMAX model, has a hierarchical structure as shown in Fig. 7.6.

7.3.2 Combination of Optical and Sonar Images

Our model is a combination of HMAX model for both optical images and sonar images as shown in Figure 7.6. Two parallel structures of HMAX are provided and the final $C2$ vectors of image and sonar features are concatenated to be fed to the classifier.

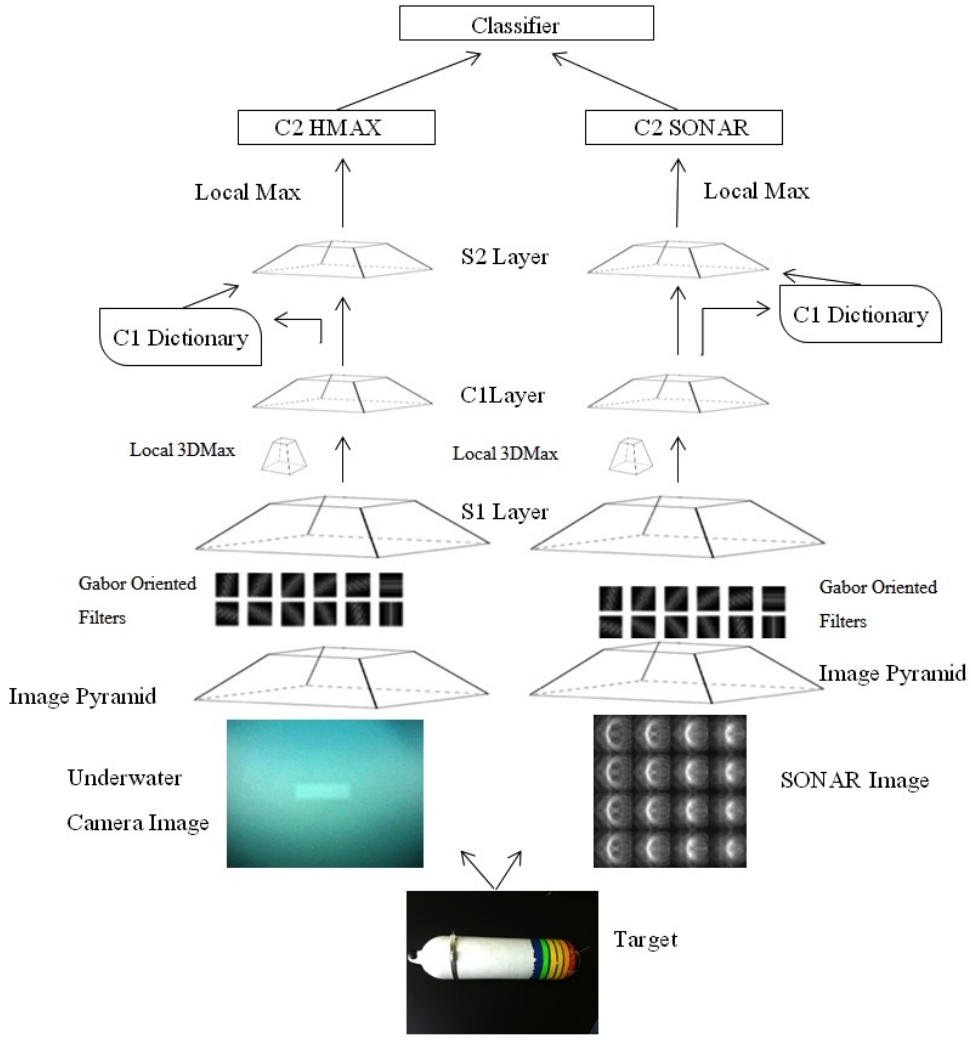


Figure 7.6: The hierarchical structure of our dual model.

7.3.3 Experimental Model and Dataset

In this section, we introduce our dataset to evaluate our model. This dataset consists of two sub-sets: optical and sonar images.

7.3.4 Diver Sonar and Optical Images

In order to explore the synergy between biosonar and optical images a data set was generated based on the physics of the acquisition of optical

images and sonar images in the underwater environment. The theoretical situation is that a diver has an optical visual system, and is equipped with a hand-held sonar unit which outputs a sonar image. We explore classification schemes such as HMAX to classify known target shapes in the optical images and the sonar images, and compare each of them to the synergistic combination of the optical and sonar images. In order to have greater control over the input in the first iteration of this approach, a simulated data set was generated. The parameters for the the model were the target shape, range to the target, the angle of the target to the diver, and the type of visual noise environment.

Optical Images

The optical images of the targets were synthesized based on target shape, range and angle to the viewer. The brightness of the target shape was reduced as a function of the range squared, representing the spherical light loss. The target image was merged with a random still frame taken from one of three videos. The videos were taken from one of three video recordings made during a variety of lighting conditions: in the late afternoon, dusk, and at night.

The video camera was hand held by a diver at mid-water column pointing horizontally. The Secchi depth, the distance beyond which a standard target (a Secchi disk marked with two black and two white quadrants) could no longer be seen was measured to be 2 meters) was measured to be 2 meters. The model was calibrated to match this distance.

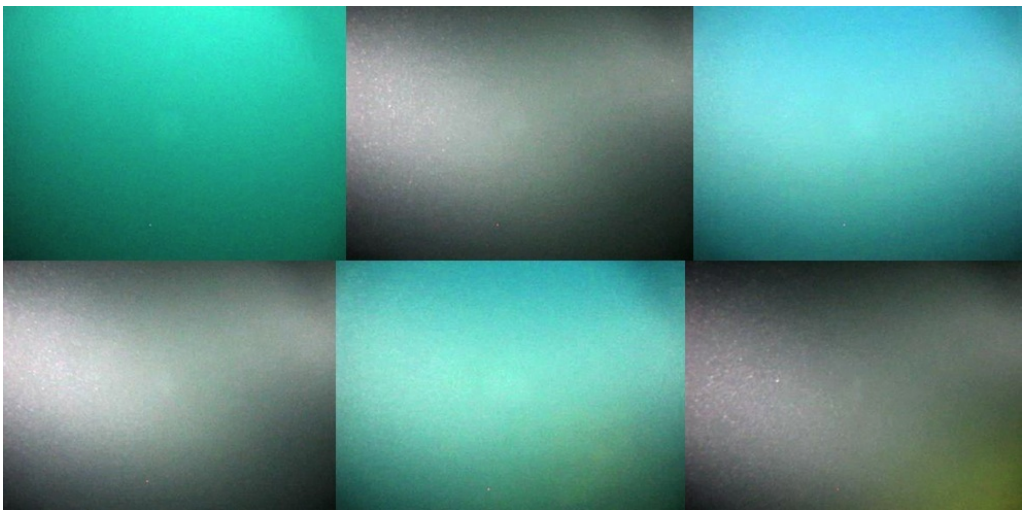


Figure 7.7: Target visibility reaches zero at farther ranges. Sample images of targets at range 3 meters.

At night a dive torch was used to illuminate the scene. The particulate matter in the water column is particularly apparent at night where there is considerable backscatter from the particulates. Figure 7.7 shows all targets (description of shapes) used in this experiment at a range of 3 meters. As this figure shows, none of the targets are visible at this range.

Sonar Images

Imaging sonars work by emitting a short pulse of acoustic energy and receiving any target reflections on an array of spatially distributed sensors. Then a process known as beamforming is employed to derive an image from the acoustic data.

We modeled the reflection of dolphin-like, Gaussian windowed pulses, 4 cycles in duration with center frequency of 130 kHz from the different target shapes, and consequent reception using an array of spatially distributed

sensors. The acoustic reflection from a target was modeled by considering it to be made up of acoustic emitters forming the shape of the target. Since the dolphin-like click is short and well defined in time, the time that it arrives on sensor can be determined. By triangulation of the time of arrival on each sensor the direction of the source can be determined. It is important that the array is large enough such that arrivals are unique to a given point in space. Technically, this is a sparse array, since the spacing of the sensors is much greater than the wavelength of 130 kHz. The targets were all 25 cm in width and height. The receiver array was formed by using 64 sensors arranged in an 8 by 8 regular square grid, over a length and breadth of 0.6 m.

To form an image, sparse array beamforming is carried out in 3D space. If the echo came from point x,y,z then the theoretical time of arrival on each sensor can be calculated from geometry. At each corresponding arrival time in the timeseries for each sensor, a window is centered at the delay, and the windows are averaged across sensors to determine how much energy came from that position. The window length determines the dimensions of the volume pixel (Voxel) around each point in 3D space. Typically the window length is chosen to be the same length as the source signal. Due to refraction and the long wavelength of acoustic waves in water, wavefronts tend to be curved, especially at shorter ranges, hence reflections from the edges of the target come back later than those from the center. Since time and range are interchangeable, several range slices have to be considered to represent the target shape. In this simulation the times series was beamformed at

16 ranges centered around the target. The range slices are rearranged and presented as a 2D image suitable for input to the HMAX model.

At 120 to 130 kHz the average peak-to-peak dolphin click level has been observed to be in the order of 220 dB re 1 μPa @ 1 m. In Singapore waters ambient noise in the same frequency band is typically 60 dB re 1 μPa @ 1 m. Taking into account the two way range dependent spreading losses given by $2 \times 20 \times log_{10}3 = 20db$ at 3 meters and target strength losses, typically 10 db for an aluminum target, the Signal to Noise Ratio (SNR) = 130 dB over the range of the simulation. This has very little effect on the sonar system. It also implies that maximum range of a dolphin bio-sonar system is of the order of a few hundred meters. However, with increasing range, the angles and hence the time differences get progressively smaller between the array sensors, hence accuracy/resolution will drop.

7.3.5 Dataset

In this experiment, we created 6 classes of objects (rectangle, triangle, vertical target, horizontal target, circle and cross) and recorded their images underwater in average visibility in the daytime and at night. Images were taken from 5 different viewpoints (15, 30, 45, 90, 115, 130 and 145 degrees) and were taken at different ranges from 1-5 m in steps of 0.25 m. Sample images of this dataset and their equivalent sonar images are shown in Figure 7.8. The dataset includes 155 images of each category at the short range, 60 images at the mid range and 60 images at the long range for the visual images, and the same numbers for the sonar images. Fifty training images

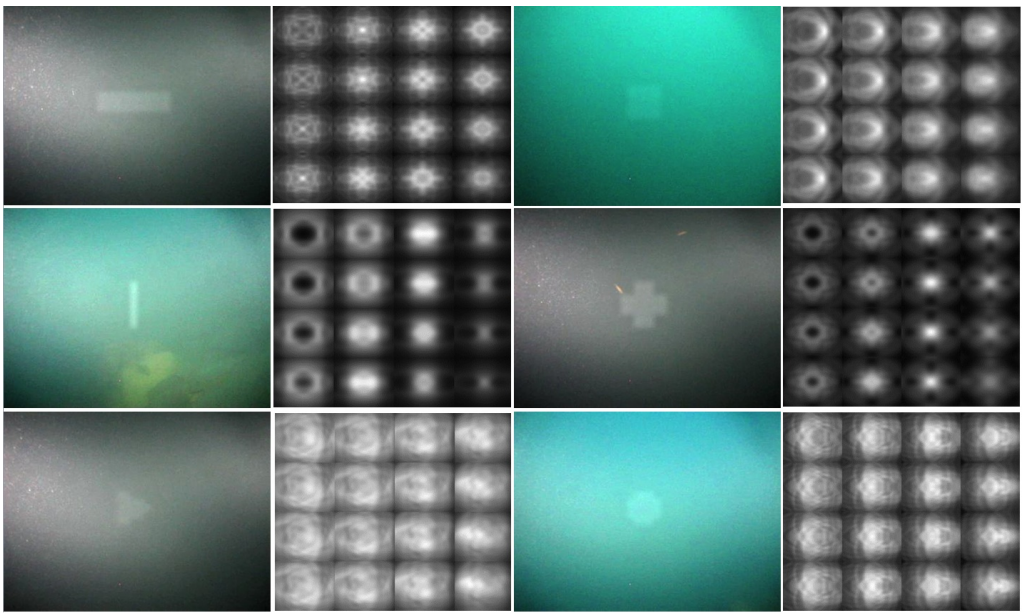


Figure 7.8: Sample pairs of images of camera and sonar taken at range 1.5m. The images on the left of each pair show a visual image of an object and those on the right are cuts from a 3D sonar image.

were randomly selected from each category in the short range subset and the rest of the images formed the test set. Thirty training images were randomly chosen from each category in the mid and long range subsets and the remaining thirty images were used for testing.

7.3.6 Experimental Results

Table 7.3 shows the classification rates for the optical and sonar images as a function of short (1 m to 2.5 m), mid (2.5 m to 3.5 m) and long range (3.5 m to 5 m). Rates of classification for the optical images falls off quickly with range: at long range the accuracy reaches 16% which is about random class selection (since there are 6 classes in these experiments,

chance is equal to 1/6). Sample images of targets at the long range are shown in Figure 7.7.

Input Images	**Short**	**Medium**	**Long**	**Average**
Optical	92.7% ±1.1	43.3% ±4.8	16.7% ±0.1	50.9% ±1.5
Sonar	89.1% ±1.7	**96.7**% ±0.8	94.3% ±1.7	93.36% ±1.4
Optical + Sonar	**97.8**% ±0.9	92.7% ±2.2	**94.3**% ±1.2	**94.9**% ±1.4

Table 7.3: Classification accuracy using different ranges of images and sonar. Short range is between 1 - 2.5m. Medium range is 2.5 - 3.5m and long range is between 3.5 - 5m.

Rates of classification for the sonar images remain reasonably constant at mid and long ranges, and clearly outperform the optical images. By combining the optical and sonar images the classification rates are higher than either optical or sonar images in short range. At far range, the classification accuracy of the combined model is as good as the performance of the sonar model alone, and this shows that the support vector machine classifier neglects the optical images as they do not contain any information.

Since the range to the target can be estimated from the sonar data stream the best model can be chosen. However, this estimation requires a separate setup and when the distance estimation is not available, the combined model may be implemented since generally the combined model of images and sonar performs as well or better than any individual model

(except for the mid-range).

7.3.7 Discussion

The sonar images are somewhat difficult to interpret visually, but can be classified accurately by the HMAX model, suggests that the sonar image could be replaced by a pictorial representation of the target. This would be far easier for a diver to interpret in difficult conditions such as encountered when working underwater. Not only would this be less ambiguous but it would reduce the time needed to interpret a sonar image, and anything that both increases accuracy and saves time when performing tasks underwater is important.

As Table 7.3 shows, the classification accuracy of the model at short range is best when both optical images and sonar images are used. At mid range, the sonar images outperform the combined model; however the difference is negligible. When we position the targets at the far range, the classification accuracy of the combined model is as good as the performance of the sonar model alone, and this shows that the support vector machine (SVM) classifier neglects optically derived images because they do not contain any useful information. Generally the combined model of images and sonar performs as well or better than any individual model at short and long ranges. In a real application, the distance to the target can be estimated by the signal acquired from sonar information and the more accurate model may be chosen. However, the combined model can be used without a significant loss in order to accelerate the classification process

when the range of the target range is unknown. In a lower level task, sonar information can also be used to simply detect the presence of an object - any undefined object - which might be of practical importance.[2]

We have demonstrated that, as the evolution of echolocation or biosonar shows, optical sensing of surface information of target shape produces higher resolution images than sonar/ultrasound imaging due to the wavelengths of the signals used. However, the situation is complicated when color is involved, and also when information on the material composition of a target or it's internal structure is important. In the former case, only optical images contain relevant information and we will explore this aspect through the CQ-HMAX model we have successfully used previously on various datasets Jalali et al. (2013d,b). It is also important to remember that ultrasonic imaging can give far greater 3D structural information than conventional optical imaging (confoccal imaging is not applicable in the current work) and this might be important both at short range even though resolution is higher in optical images, but also at longer ranges. The present work is the first stage towards producing the most informative image.

Finally with regard to bio-sonar in dolphins, it is not clear how the dolphin forms a mental 'image'. It is thought that the mandibles, particularly in the lower jaw, received echolocation clicks to the inner ears, and perhaps

[2]The part on providing the SONAR and Visual images was carried out in collaboration with Dr Paul Seekings, the section on similarities between audio and visual cortex is provided in collaboration with Dr Cheston Tan and the model proposal, and experiments were done by Sepehr Jalali.

the teeth act as complementary sensors. It would be interesting to repeat the experiment carried out here, but using an auditory system model on the timeseries received on a receiver array modeled on the jaw of a species of dolphin with known echolocation abilities, and compare the output with the results shown here.

Chapter 8

Conclusion

In this thesis, we reviewed several biologically inspired models for image classification in Chapter 2. We also touched on the most relevant computer vision approaches and provided a discussion and comparison on these models. We described the original HMAX model in Chapter 3 and presented the existing modifications and improvements to this model in detail. We proposed several modifications, enhancements and applications for the HMAX model in the rest of the chapters.

We investigated different methods for the creation of the dictionary of features and compared random and non-random sampling methods in Chapter 4. We introduced several pooling methods and encoding of occurrence and co-occurrence of features in Chapter 5, followed by our new biologically inspired color model in Chapter 6 and presented an application for this model in Chapter 7. In this chapter we summarize the previous chapters and our contributions, and suggest some directions for the future work.

8.1 Contributions

Image classification is a challenging problem in computer vision and it remains an open research area as there is no perfect solution to this task. Recently more scientists are looking into human (and other mammals) visual cortex for inspirations to find a better computational model. However due to the complexity of the brain, the exact process in which the brain carries out image classification and object recognition is far from being well understood. Based on the current findings, a hierarchical structure is proposed for simulating the human visual cortex. One of the models that resembles this structure well is HMAX hierarchical approach (Riesenhuber and Poggio, 1999) which models the first 150 ms of the bottom-up processes in the human visual cortex.

In HMAX model, there is a S-C (simple-complex) interleaving structure in which S layers provide selectivity and C layers add invariance to the features or filters. We explored several feature selection methods in the middle layers of this structure and showed that the use of random sampling from $C1$ layer in order to create a dictionary of features for $S2$ layer, performs as good as non-random clustering of more features in many different combinations when the spatial information of features and the occurrence of features is neglected. This suggests that random sampling is an effective fast method in comparison with clustering for creation of the dictionary of features. However when the spatial information of features is used with a higher number of clusters created and more repetitive ones in each class are chosen for creation of the dictionary of features, a better classification

accuracy is achieved.

This suggests that the use of this frequency information (similar to bag of features) encodes useful information and can be used in HMAX structure. Hence, we evaluated different pooling methods and showed that using MAX pooling along with Mean pooling, a better performance is achieved on several datasets. Furthermore, we explored encoding co-occurrence of features, and used the more frequent features as candidate features for this. Using co-occurrence of features, results in a boost in classification performance when a higher number of training images are available like in a subset of Caltech256 dataset. This encourages encoding co-occurrence of more than two features and using top-down information for selecting more meaningful features such as 'eyes, nose, mouth, etc.' for encoding co-occurrence rather than selecting features based on their frequency alone.

The use of color information also showed no significant improvement when naively added to the HMAX structure as three (RGB) parallel structures and concatenation of the final $C2$ features for feeding to classifier. However, we proposed a new hierarchical biologically inspired color model in which color quantization cores are used for quantizing the YIQ color space, and a new structure similar to HMAX was proposed which resulted in significant improvements on several datasets. Table 8.1, a summary of the best classification accuracy achieved by applying different models in comparison with HMAX model is provided.

In this thesis, we enhanced the performance of HMAX model on Cal-

Database	**HMAX**	**Improved Model**
Caltech101	54.7	64.3
Caltech256 subset (14 classes)	60.2	64.4
Soccer	24.7	77.1
Flowers	42.5	78.3
Scenes	71.4	86.5
Underwater (good visibility)	92.9	99.0
Underwater (bad visibility)	50.9	94.9
Mitosis	10.1	30.5
Benthic Marine Organisms	39.6	61.1

Table 8.1: Comparison of HMAX performance vs. the best performance achieved by a modified HMAX model on each dataset. The best performance is either CQ-HMAX, Co-Occurrence HMAX, HMean or a combination of them.

tech101 dataset by adding HMean, and CQ-HMAX information to it. We outperformed the state-of-the-art performance on Soccer and Flowers datasets by adding the HMAX, HMean and CQ-HMAX to the SODO-HMAX model and showed that the use of CQ-HMAX, which is a high-level color model to SODO-HMAX, results in better classification results which supports the use of color in both high and low levels of image processing. CQ-HMAX model is significantly faster than SOD-HMAX model as it does not need the Gabor or Gaussian filter convolution on the image pyramid and can be used in real-time applications. Using this combination (CQ-HMAX, HMAX, HMean and SODO-HMAX), we reached classification performances of about 93.3% on Soccer dataset and 90.1% on Flowers dataset

which are better than the state-of-the-art performances of all bottom-up computer vision and biologically inspired approaches and reached 86.5% classification accuracy on Scenes dataset (with the combination of CQ-HMAX, HMAX and HMean) which is on par with the classification score achieved by SODO-HMAX model.

We used our modified version of HMAX and CQ-HMAX on MITOS dataset for detection of mitosis in histopatholgy images and compared it with some state-of-the-art SIFT methods and showed that HMAX and CQ-HMAX models outperformed SIFT based models in this specific application. Our proposed model of CQ-HMAX also outperforms the SIFT, HMAX and SVM methods in classification of benthic marine organisms. We also proposed a new model for simulating the visual and auditory pathways by creating sonar images and feeding them to a HMAX structure in parallel with visual images.

In this thesis, we provided three main modifications to the HMAX model including different methods for creation of the dictionary of features, different pooling methods and encoding occurrence and co-occurrence of features. We introduced a new biologically inspired color model to image classification and showed that using these modifications, a higher classification accuracy is achieved in several benchmark datasets. We also created three different datasets for further evaluating our methods. Our modifications and our new biologically inspired color model can be merged into many other models and are not bound to HMAX model specifically.

8.2 Future Works

One possible extension is to use non-redundant features as discriminative features. In order to find the best features for encoding co-occurrence, we find the features with the highest frequency in each class regardless of their occurrence in other classes. One possible extension is to find the cross category occurrence of features and select the ones which make a discriminative difference between classes. Evaluating inter-class clustering to create the dictionary of features in CQ-HMAX model could be another future direction.

Another prospective extension is to encode co-occurrence of features in CQ-HMAX model. In this extension, the same approach used for encoding HMean features will be used and features with a higher occurrence in each category are found and used for creation of dictionary of co-occurrences.

Another possible extension to weighting features is to use a retinotopic mapping approach in which the center of each patch is given a higher weight similar to the projection of an image in the $V1$ as can be seen in Figure 8.1. The representation of the central 5 degrees (shaded areas) in the visual field occupies about 40% of the cortex (LeVay et al., 1985).

One other possible extension could be using Mean pooling in $S1$-$C1$ layer. Our HMean model performs mean pooling in $S2$-$C2$ layer.

The combination of visual and sonar images can be used in other applications such as robotics where the range of objects can be detected by the acoustic information. The advantage that the use of acoustic signals have over laser used for detection of the range (such as in Kinect), is the differ-

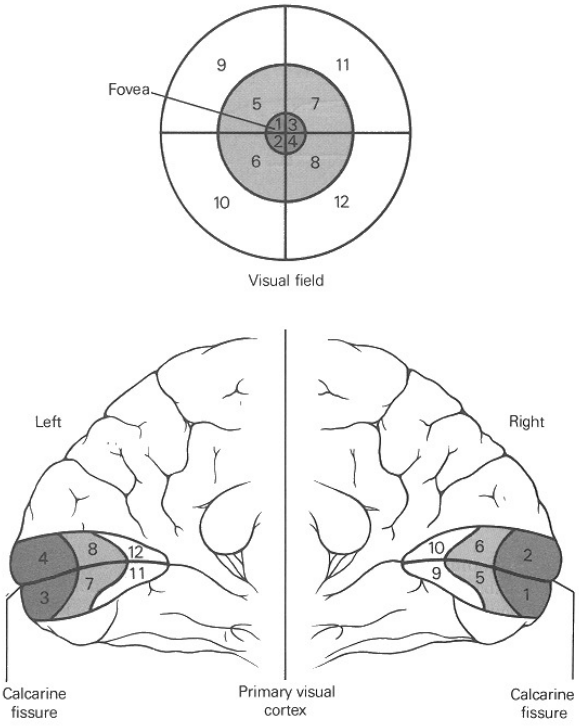


Figure 8.1: Retonotopic mapping in the fovea. The foveal area is represented by a relatively larger area in $V1$ than the peripharal areas.

ent responses (strengths) we receive when touching different materials. For instance the response taken from a soft tissue and the response from a hard object, have different strengths and this can be used to further differentiate objects. Hence this final model will use color, grayscale, and acoustic information for achieving a better understanding of the environment.

One interesting research area as future work of this thesis, could be evaluating combination of CQ-HMAX and SODO-HMAX models from lower levels.

Bibliography

Aslin, R. N. and Newport, E. L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. *Current Directions in Psychological Science*, 21(3):170–176.

Baker, C. I., Behrmann, M., and Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature neuroscience*, 5(11):1210–6.

Banno, T., Ichinohe, N., Rockland, K. S., and Komatsu, H. (2011). Reciprocal connectivity of identified color-processing modules in the monkey inferior temporal cortex. *Cerebral Cortex*, 21(6):1295–310.

Barlow, H. et al. (1975). Visual experience and cortical development. *Nature*, 258(5532):199–204.

Bart, E., Byvatov, E., and Ullman, S. (2004). View-invariant recognition using corresponding object fragments. *Computer Vision-ECCV 2004*, pages 152–165.

Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers Inc.

Bengio, Y. and LeCun, Y. (2007). Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*.

Bloom, H. J. G. and Richardson, W. W. (1957). Histological grading and prognosis in breast cancer a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, 11:359–377.

Bosch, A., Zisserman, A., and Muñoz, X. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–27.

Bosch, A., Zisserman, A., and Muoz, X. (2007). Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

Boynton, R. M. and Olson, C. X. (1987). Locating basic colors in the OSA space. *Color Research & Application*, 12(2):94–105.

Brown, M. and Susstrunk, S. (2011). Multi-spectral SIFT for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE.

Burghouts, G. J. and Geusebroek, J.-M. (2009). Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48–62.

Chang, H., Loss, L. A., and Parvin, B. (2012). Nuclear segmentation in h&e sections via multi-reference graph-cut (mrgc). In *ISBI*.

Chevillet, M., Riesenhuber, M., and Rauschecker, J. P. (2011). Functional

correlates of the anterolateral processing hierarchy in human auditory cortex. *Journal of Neuroscience*, 31(25):9345–52.

Conway, B. R., Chatterjee, S., Field, G. D., Horwitz, G. D., Johnson, E. N., Koida, K., and Mancuso, K. (2010). Advances in color science: from retina to behavior. *The Journal of Neuroscience*, 30(45):14955–63.

Conway, B. R. and Livingstone, M. S. (2006). Spatial and temporal properties of cone signals in alert macaque primary visual cortex. *The Journal of Neuroscience*, 26(42):10826–46.

Conway, B. R., Moeller, S., and Tsao, D. Y. (2007). Specialized color modules in macaque extrastriate cortex. *Neuron*, 56(3):560–73.

Conway, B. R. and Tsao, D. Y. (2006). Color architecture in alert macaque cortex revealed by FMRI. *Cerebral Cortex*, 16(11):1604–13.

de Heering, A., Houthuys, S., and Rossion, B. (2007). Holistic face processing is mature at 4 years of age: evidence from the composite face effect. *Journal of Experimental Child Psychology*, 96(1):57–70.

Dean, P. (1979). Visual cortex ablation and thresholds for successively presented stimuli in rhesus monkeys: II. Hue. *Experimental Brain Research.*, 35(1):69–83.

Derrington, A. M., Krauskopf, J., and Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357:241–65.

Desimone, R., Schein, S. J., Moran, J., and Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Research*, 25(3):441–52.

Edelman, S. (1991). Features of recognition. *CS-TR 91-10, Weizmann Institute of Science*.

Edelman, S., Yang, H., Hiles, B., and Intrator, N. (2002). Probabilistic principles in unsupervised learning of visual structure: human data and a model. *Advances in neural information processing systems*, 1:19–26.

Edwards, R., Xiao, D., Keysers, C., Földiák, P., and Perrett, D. (2003). Color sensitivity of cells responsive to complex stimuli in the temporal cortex. *Journal of Neurophysiology*, 90(2):1245–56.

Elston, C. W. and Ellis, I. O. (2002). Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer experience from a large study with long-term follow-up. *Histopathology*, 41:151–151.

Fidler, S., Boben, M., and Leonardis, A. (2008). Similarity-based cross-layered hierarchical representation for object categorization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Finn, I. M. and Ferster, D. (2007). Computational diversity in complex cells of cat primary visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(36):9638–48.

Fiser, J. and Aslin, R. N. (2001). Unsupervised statistical learning of

higher-order spatial structures from visual scenes. *Psychological science*, 12(6):499–504.

Fiser, J. and Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15822–6.

Freiwald, W. A., Tsao, D. Y., and Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature neuroscience*, 12(9):1187–96.

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130.

Gawne, T. J. and Martin, J. M. (2002). Responses of Primate Visual Cortical V4 Neurons to Simultaneously Presented Stimuli. *J Neurophysiol*, 88(3):1128–1135.

George, D. (2008). *How the brain might work: A hierarchical and temporal model for learning and recognition.* PhD thesis, Stanford University.

George, D. and Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits.

Geusebroek, J., Van den Boomgaard, R., Smeulders, A., and Geerts, H. (2001). Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350.

Gevers, T. and Stokman, H. (2004). Robust histogram construction from color invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):113–118.

Goldstein, E. (2009). *Sensation and perception*. Wadsworth Pub Co.

Gurcan, M., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., and Yener, B. (2009). Histopathological image analysis a review. *Biomedical Engineering, IEEE Reviews in*, 2:147–171.

Harris, Z. (1954). Distributional structure. *Word*.

Hawkins, J. and Blakeslee, S. (2005). *On intelligence*. Owl Books.

Hawkins, J. and George, D. (2006). Hierarchical temporal memory: Concepts, theory and terminology. *White paper, Numenta Inc.*

Heywood, C. A., Shields, C., and Cowey, A. (1988). The involvement of the temporal lobes in colour discrimination. *Experimental Brain Research.*, 71(2):437–41.

Hinton, G., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Hirabayashi, T. and Miyashita, Y. (2005). Dynamically modulated spike correlation in monkey inferior temporal cortex depending on the feature

configuration within a whole object. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(44):10299–307.

Horel, J. A. (1994). Retrieval of color and form during suppression of temporal cortex with cold. *Behavioural Brain Research*, 65(2):165–72.

http://ipal.cnrs.fr/ICPR2012 (2012). Mitosis detection contest website.

Hubel, D. and Wiesel, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574.

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106.

Hubel, D. H. and Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*.

Humayun, I., Sepehr, J., Roux, L., Racoceanu, D., Joo-Hwee, L., Gilles, L. N., and Frédérique, C. (2012). Automated mitosis detection using texture, sift features and hmax biologically inspired approach. In *Histopathology Image Analysis (HIMA), MICCAI*.

Humayun, I., Sepehr, J., Roux, L., Racoceanu, D., Joo-Hwee, L., Gilles, L. N., and Frédérique, C. (2013). Automated mitosis detection using texture, sift features and hmax biologically inspired approach.

Hurvich, L. M. (1981). *Color Vision*. Sinauer Associates Inc., Sutherland, MA.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1255.

Jalali, S., Lim, J., Ong, S., and Tham, J. (2010). Dictionary of features in a biologically inspired approach to image classification. In *Internationl Conference on Neural Information Processing*, pages 541–548. Springer.

Jalali, S., Lim, J., Tham, J., and Ong, S. (2012). Clustering and use of spatial and frequency information in a biologically inspired approach to image classification. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE.

Jalali, S., Seekings, P., Tan, C., Ratheesh, A., Lim, J., , and Taylor, E. (2013a). The Use of Optical and Sonar Images in the Human and Dolphin Brain for Image Classification. In *Proceedings of the International Joint Conference on Neural Networks (Accepted)*.

Jalali, S., Seekings, P., Tan, C., Tan, H. Z. W., Lim, J., , and Taylor, E. (2013b). Classification of Marine Organisms in Underwater Images using CQ-HMAX Biologically Inspired Color Approach. In *Proceedings of the International Joint Conference on Neural Networks (Accepted)*.

Jalali, S., Tan, C., Lim, J., Tham, J., Ong, S., Seekings, P., and Taylor, E. (2013c). Encoding Co-occurrence of Features in HMAX Model. In

Proceedings of the Annual Conference of the Cognitive Science Society (Accepted).

Jalali, S., Tan, C., Lim, J., Tham, J., Ong, S., Seekings, P., and Taylor, E. (2013d). Visual Recognition using a Combination of Shape and Color Features. In *Proceedings of the Annual Conference of the Cognitive Science Society (Accepted).*

King, A. J. and Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? *Nature Neuroscience*, 12(6):698–701.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69.

Komatsu, H. (1993). Neural coding of color and form in the inferior temporal cortex of the monkey. *Biomedical Research*, 14:7–13.

Komatsu, H. (1997). Neural representation of color in the inferior temporal cortex of the macaque monkey. In Sakata, H., Mikami, A., and Fuster, J. M., editors, *The Association Cortex*, pages 269 – 280. Harwood Academic Publishers, Amsterdam.

Komatsu, H. (1998). Mechanisms of central color vision. *Current opinion in Neurobiology*, 8(4):503–8.

Komatsu, H., Ideura, Y., Kaji, S., and Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *The Journal of Neuroscience*, 12(2):408–24.

Kuśmierek, P., Ortiz, M., and Rauschecker, J. P. (2012). Sound-identity processing in early areas of the auditory ventral stream in the macaque. *Journal of Neurophysiology*, 107(4):1123–41.

Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *Journal of neurophysiology*, 92(5):2704–13.

Le Grand, R., Mondloch, C. J., Maurer, D., and Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological Science*, 15(11):762–8.

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, pages 255–258.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–15.

LeVay, S., Connolly, M., Houde, J., and Van Essen, D. (1985). The complete pattern of ocular dominance stripes in the striate cortex and visual field of the macaque monkey. *The Journal of neuroscience*, 5(2):486–501.

Li, F.-F., Fergus, R., and Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004. In *Workshop on Generative-Model Based Vision*, volume 2.

Li, F.-F. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE.

Lim, J. (1999). Learning visual keywords for content-based retrieval. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, volume 2, pages 169–173. IEEE.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157.

Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.

Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19(1):577–621.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee.

Lyon, D. C. and Connolly, J. D. (2012). The case for primate V3. *Proceedings. Biological sciences / The Royal Society*, 279(1729):625–33.

Masquelier, T. and Thorpe, S. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology*, 3(2):e31.

Matuzawa, T. (1985). Colour naming and classification in a chimpanzee. *Journal of Human Evolution*, 14:283 – 291.

McClelland, J. and Rumelhart, D. (2002). An interactive activation model of context effects in letter perception. *Psycholinguistics: critical concepts in psychology*, 88(5):422.

Mishkin, M., Ungerleider, L., and Macko, K. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817–20.

Mladenić, D., Brank, J., Grobelnik, M., and Milic-Frayling, N. (2004). Feature selection using linear classifier weights: interaction with classification models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241. ACM.

Mollon, J. D. and Jordan, G. (1997). On the nature of unique hues. In

Dickinson, C., Murray, I., and Carden, D., editors, *John Dalton's Colour Vision Legacy*, pages 381 – 392. Taylor and Francis, London.

Mutch, J., Knoblich, U., and Poggio, T. (2010a). CNS: a GPU-based framework for simulating cortically-organized networks. Technical Report MIT-CSAIL-TR-2010-013 / CBCL-286, Massachusetts Institute of Technology, Cambridge, MA.

Mutch, J., Knoblich, U., and Poggio, T. (2010b). CNS: a GPU-based framework for simulating cortically-organized networks. *MIT-CSAIL-TR-2010-013*.

Mutch, J. and Lowe, D. (2006). Multiclass object recognition with sparse, localized features. *Computer Vision and Pattern Recognition (CVPR)*, 148(3):574.

Mutch, J. and Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80(1):45–57.

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology*, 14(4):474–80.

Nilsback, M.-E. and Zisserman, A. (2006). A visual vocabulary for flower classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1447–1454.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T., and Hickok, G. (2010). Hierarchical organization of

human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, 20(10):2486–95.

Oleksiak, A., Klink, P. C., Postma, A., van der Ham, I. J. M., Lankheet, M. J., and van Wezel, R. J. A. (2011). Spatial summation in macaque parietal area 7a follows a winner-take-all rule. *Journal of neurophysiology*, 105(3):1150–8.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Olshausen, B., Anderson, C., and Van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700.

Op de Beeck, H. P. and Baker, C. I. (2010). The neural basis of visual object learning. *Trends in Cognitive Sciences*, 14(1):22–30.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT Press, Cambridge, MA.

Ramanathan, K., Shi, L., Li, J., Lim, K., Li, M., Ang, Z., and Chong, T. (2009). A neural network model for a hierarchical spatio-temporal memory. *Advances in Neuro-Information Processing*, pages 428–435.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, 8(4):516–21.

Rauschecker, J. P. and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, 12(6):718–24.

Rauschecker, J. P. and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):11800–6.

Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 19(5):1736–53.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025.

Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, 3:1199–1204.

Roe, A., Pallas, S., Kwon, Y., and Sur, M. (1992). Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex. *Journal of Neuroscience*, 12(9):3651–3664.

Roe, A. W., Pallas, S. L., Hahm, J. O., and Sur, M. (1990). A map of visual space induced in primary auditory cortex. *Science*, 250(4982):818–20.

Romanski, L. M. and Averbeck, B. B. (2009). The primate cortical auditory

system and neural representation of conspecific vocalizations. *Annual Review of Neuroscience*, 32:315–46.

Rutishauser, U., Walther, D., Koch, C., and Perona, P. (2004). Is bottom-up attention useful for object recognition?

Sakai, K. and Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, 354(6349):152–5.

Salakhutdinov, R. and Hinton, G. (2009). Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.

Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. *Experimental Brain Research*, 77(1):23–30.

Serre, T., Oliva, A., and Poggio, T. (2007a). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424.

Serre, T. and Riesenhuber, M. (2004). Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. *Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory.*

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426.

Serre, T., Wolf, L., and Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 994–1000. IEEE.

Shahbaz Khan, F., Anwer, R. M., Van de Weijer, J., Bagdanov, A. D., Vanrell, M., and Lopez, A. M. (2012). Color attributes for object detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3306–3313. IEEE.

Sharma, J., Angelucci, A., and Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature*, 404(6780):841–7.

Sigala, N. and Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415(6869):318–20.

Sigala, R., Serre, T., Poggio, T., and Giese, M. (2005). Learning features of intermediate complexity for the recognition of biological motion. *Artificial Neural Networks: Biological Inspirations–ICANN 2005*, pages 241–246.

Stoughton, C. M. and Conway, B. R. (2008). Neural basis for unique hues. *Current Biology : CB*, 18(16):R698–9.

Sur, M. and Leamey, C. A. (2001). Development and plasticity of cortical areas and networks. *Nature Reviews Neuroscience*, 2(4):251–62.

Takechi, H., Onoe, H., Shizuno, H., Yoshikawa, E., Sadato, N., Tsukada,

H., and Watanabe, Y. (1997). Mapping of cortical areas involved in color vision in non-human primates. *Neuroscience Letters*, 230(1):17–20.

Talkington, W. J., Rapuano, K. M., Hitt, L. A., Frum, C. A., and Lewis, J. W. (2012). Humans mimicking animals: a cortical hierarchy for human vocal communication sounds. *Journal of Neuroscience*, 32(23):8084–93.

Tang, J., Miller, S., Singh, A., and Abbeel, P. (2012). A textured object recognition pipeline for color and depth image data. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3467–3474. IEEE.

Tarr, M. and ulthoff, H. (1998). Image-based object recognition in man, monkey and machine. *Cognition*, 67(1-2):1–20.

Taylor, G., Hinton, G., and Roweis, S. (2007). Modeling human motion using binary latent variables. *Advances in neural information processing systems*, 19:1345.

Teot, H. A., Sposto, R., Khayat, A., Qualman, S., Reaman, G., and Parham, D. (2007). The problems and promise of central pathology review development of a standardized procedure for the children oncology group. *Pediatric and Developmental Pathology*, 10:199–207.

Theriault, C., Thome, N., and Cord, M. (2011). Hmax-s : deep scale representation for biologically inspired image classification. In *International Conference on Image Processing*, number 3, pages 3–6.

Turk-Browne, N. B., Jungé, J., and Scholl, B. J. (2005). The automaticity

of visual statistical learning. *Journal of experimental psychology. General*, 134(4):552–64.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., and Johnson, M. K. (2009). Neural evidence of statistical learning: efficient detection of visual regularities without awareness. *Journal of cognitive neuroscience*, 21(10):1934–45.

Uchikawa, K. and Boynton, R. M. (1987). Categorical color perception of Japanese observers: comparison with that of Americans. *Vision Research*, 27(10):1825–33.

Valberg, A. (2001). Unique hues: an old problem for a new generation. *Vision Research*, 41(13):1645–57.

Van de Sande, K. E. A., Gevers, T., and Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–96.

Van De Weijer, J. and Schmid, C. (2006). Coloring local feature extraction. *Computer Vision–ECCV 2006*, pages 334–348.

Van de Weijer, J. and Schmid, C. (2007). Applying Color Names to Image Description. In *2007 IEEE International Conference on Image Processing*, volume 3, pages 493– 496. IEEE.

Van de Weijer, J., Schmid, C., Verbeek, J., and Larlus, D. (2009). Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–23.

Vedaldi, A. and Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, pages 1469–1472. ACM.

Walther, D. (2006). Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics.

Yasuda, M., Banno, T., and Komatsu, H. (2010). Color selectivity of neurons in the posterior inferior temporal cortex of the macaque monkey. *Cerebral Cortex*, 20(7):1630–46.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1):141–145.

Young, A. W., Hellawell, D., and Hay, D. C. (1987). Configurational information in face perception. *Perception*, 16(6):747–59.

Zhang, J., Barhomi, Y., and Serre, T. (2012). A new biologically inspired color image descriptor. In *Computer Vision – ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 312–324. Springer.

Zoccolan, D., Cox, D. D., and DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 25(36):8150–64.

Publications

- Jalali, S., Lim, J.H., Tham, J.Y, Ong, S.H.,, "Dictionary of features in a biologically inspired approach to image classification" in International Conference on Neural Information Processing (ICONIP10). Springer, 2010, pp. 541-548.

- Jalali, S., Lim, J.H., Tham, J.Y, Ong, S.H.,, "Clustering and use of spatial and frequency information in a biologically inspired approach to image classification". IJCNN, The 2012 IEEE International Joint Conference on Neural Networks, 2012.

- Irshad, H., Jalali, S., Roux, L., Racoceanu ,D., Lim, J.H., Gilles, N. and Frederique,C. "Automated mitosis detection using texture, sift features and hmax biologically inspired approach" in Workshop on Histopathology Image Analysis (HIMA), MICCAI, 2012.

- Irshad, H., Jalali, S., Roux, L., Racoceanu ,D., Lim, J.H., Gilles, N. and Frederique,C. "Automated mitosis detection using texture, sift features and hmax biologically inspired approach" accepted in Journal of Histopathology Image Analysis, 2013.

- Jalali, S., Seekings, P., Tan, C., Tan, HZW., Lim, J.H. and Talyor, E. Classification of Marine Organisms in Underwater Images using CQ-HMAX Biologically inspired Color Approach . IJCNN, The 2013 International Joint Conference on Neural Networks, 2013. (Accepted)

- Jalali, S., Seekings, P. , Tan, C., Ratheesh A., Lim, J.H., Taylor, E. The Use of Optical and Sonar Images in the Human and Dolphin Brain for

Image Classification . IJCNN, The 2013 International Joint Conference on Neural Networks, 2013. (Accepted)

- Jalali, S., Tan, C., Lim, J.H., Tham, J.Y, Ong, S.H., Seekings, P. and Taylor, E. Encoding Co-occurrence of Features in HMAX Model . (CogSci), the annual meeting of the cognitive science society, 2013. (Accepted)

- Jalali, S., Tan, C., Lim, J.H., Tham, J.Y, Ong, S.H., Seekings, P. and Taylor, E. Visual Recognition using a Combination of Shape and Color Features. (CogSci), the annual meeting of the cognitive science society, 2013. (Accepted)

- Jalali, S., Tan, C., Lim, J.H., Tham, J.Y, Ong, S.H., "CQ-HMAX, a New Biologically Inspired Color Approach to Image Classification". in progress for Pattern Recognition.

- Jalali, S., Tan, C., Lim, J.H., Tham, J.Y, Ong, S.H., "Occurrence and co-occurrence of features in HMAX biologically inspired approach". In progress for Journal of Neural Networks.