Biologically-Inspired Hierarchical Architectures for Object Recognition

Ali Alameer

Newcastle University

Newcastle upon Tyne, UK

A thesis submitted for the degree of

Doctor of Philosophy

June 2018

This thesis is dedicated to my loving parents.

Declaration

NEWCASTLE UNIVERSITY

SCHOOL OF ENGINEERING

I, Ali Munthr Abdulkareem Alameer, declare that this thesis is my own work and it has not been previously submitted, either by me or by anyone else, for a degree or diploma at any educational institute, school or university. To the best of my knowledge, this thesis does not contain any previously published work, except where another persons work used has been cited and included in the list of references.

Signature:

Student: Ali Munthr Abdulkareem Alameer

Date:

SUPERVISOR'S CERTIFICATE

This is to certify that the entitled thesis "Biologically-Inspired Hierarchical Architectures for Object Recognition" has been prepared under my supervision at the School of Engineering / Newcastle University for the degree of PhD in Electrical and Electronic Engineering.

Signature:

Supervisor: Dr. Kianoush Nazarpour

Date:

Acknowledgements

First, I sincerely thank my supervisor Dr Kianoush Nazarpour for his valuable guidance, support and huge amounts of involvement in my journey over the past four years. I have benefited tremendously from his knowledge, enthusiasm and important feedback on my papers and reports. He was always available, replying my emails on weekends and on his own holidays. It is my privilege and honour to have been his research student. His support went beyond my PhD, including valuable suggestions for my future career. I would also like to thank Dr Patrick Degenaar for his guidance and assistance during my PhD. He has been very positive and supportive, providing valuable feedback and suggestions that have enhanced my PhD dramatically.

I would also like to thank my colleagues and friends, Ghazal Ghazaei, Federico Angelini, Haider Abbas, Safaa Awny, Carolina Silveira, Hannah Jones, Emma Brunton, Matthew Dyson, Waqas Rafique, Zaid Abdullah, and Michael Burn for their support during all stages of my PhD studies.

In addition, I would like to thank the Higher Committee For Education Development in Iraq (HCED) for providing the opportunity and the fund to complete my PhD degree and their continuous support and encouragement. Next, I would like to express my thanks to my country Iraq, and the Ministry of Higher Education and Scientific Research (MOHSR). I also would like to thank Iraqi Cultural Attache/London for their support.

Lastly, but most importantly I express my deepest gratitude to my parents, my sister, and my brother. I really cannot find suitable words to show my gratitude to them for their engagement throughout my education. In particular, my mother as she is the most important person in my life.

Abstract

The existing methods for machine vision translate the three-dimensional objects in the real world into two-dimensional images. These methods have achieved acceptable performances in recognising objects. However, the recognition performance drops dramatically when objects are transformed, for instance, the background, orientation, position in the image, and scale. The human's visual cortex has evolved to form an efficient invariant representation of objects from within a scene. The superior performance of human can be explained by the feed-forward multi-layer hierarchical structure of human visual cortex, in addition to, the utilisation of different fields of vision depending on the recognition task. Therefore, the research community investigated building systems that mimic the hierarchical architecture of the human visual cortex as an ultimate objective.

The aim of this thesis can be summarised as developing hierarchical models of the visual processing that tackle the remaining challenges of object recognition. To enhance the existing models of object recognition and to overcome the above-mentioned issues, three major contributions are made that can be summarised as the followings

1. building a hierarchical model within an abstract architecture that achieves good performances in challenging image object datasets;
2. investigating the contribution for each region of vision for object and scene images in order to increase the recognition performance and decrease the size of the processed data;
3. further enhance the performance of all existing models of object recognition by introducing hierarchical topologies that utilise the context in which the object is found to determine the identity of the object.

Statement of Originality

The contributions of this thesis have been supported by different journal and conference papers, which have been generated during the journey of my study. They can be listed as follows:

Alameer, A., Ghazaeil, G., Degenaar, P., and Nazarpour, K. (2015, December). An elastic net-regularized HMAX model of visual processing. 2nd IET International Conference on Intelligent Signal Processing (ISP). London.

Alameer, A., Ghazaei, G., Degenaar, P., Chambers, J. A., and Nazarpour, K. (2016). Object recognition with an elastic net-regularized hierarchical MAX model of the visual cortex. IEEE Sig. Process. Lett., vol.23, no.8, pp.1062-1066.

Alameer, A., Degenaar, P., and Nazarpour, K. (2016, October). Biologically-inspired object recognition system for recognizing natural scene categories. IEEE International Conference for Students on Applied Engineering (ISCAE), (pp. 129-132). Newcastle.

Ali Alameer, Patrick Degenaar, and Kianoush Nazarpour. Processing occlusions using elastic-net hierarchical MAX model of the visual cortex. (2017, July). IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA). (pp. 163-167). Gdynia.

Contents

List of Figures

List of Tables

Nomenclature

Symbols

$(\cdot)^{-1}$	Inverse
$(\cdot)^{T}$	Transpose
$\mathbf{I}$	Identity Matrix
$\|\cdot\|_F$	Frobenius Norm
$\|\cdot\|_2$	Euclidean Norm
C_1	Complex 1
C_2	Complex 2
S_1	Simple 1
S_2	Simple 2
V_1	Primary Visual Area
V_2	Secondary Visual Area

Acronyms/Abbreviations

AI	Artificial Intelligence
ANOVA	Analysis of Variance
AUC	Area Under the Curve
CNN	Convolutional Neural Network
En-HMAX	Elastic net Hierarchical MAX
FFA	Fusiform Face Area

HMAX	Hierarchical MAX
ICA	Independent Component Analysis
LASSO	Least Absolute Shrinkage and Selection Operator
MFS	Mid-fusiform Sulcus
PCA	Principal Component Analysis
PDF	Probability Density Function
PPA	Parahippocampal Place Area
RBF	Redial Basis Function
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SIFT	Scale-invariant Feature Transform
SPP	Spatial Pyramid Pooling
SURF	Speed Up Robust Features
SVMs	Support Vector Machines

Chapter 1

Introduction

1.1 Introduction

Vision is the process of observing the world by interpreting the environment using light reflected by the objects and accordingly extracting a meaningful interpretation [1]. Researchers of computer vision are continuously attempting to write computer programs to extract objects from images [2, 3]. Some of the approaches are inspired by the human visual system. To enable computers to infer the identity of the objects in the images, a model that extract formative features from the images is required in which each object has a unique signature [4].

1.2 Challenges Involved in Object Recognition

Object recognition is considered one of the main unsolved dilemmas in the field of Machine vision [5]. It can be explained as recognising an object represented in the form of an image captured from the real world. The objects in the real world are labelled by humans. Object recognition involves associating these priory known objects using a computer. Recognising an object in an image involves decoding the object of interest from its background. The background may also contain other objects. However, it is still the task of the object recognition model to identify objects of interest from its background. The model is expected to select the object true label for each frame image. In the past decades, object recognition was massively studied as will be explained thoroughly in this Chapter.

The task of recognising three-dimensional objects using two-dimensional images

from a particular view is extremely complicated. Each three-dimensional object may be represented in multiple images from different viewing angles.

Other difficulties include the variations in object scale, pose, illumination, location in the image, viewing angle, geometry and occlusions. Also, incomplete data is one of the most common problems in the field. However, in all the above cases, prior knowledge about the objects is given to the model. The models need to generalise to the known objects in novel transformative forms. Other challenges for object recognition involve intra-class variabilities, where some objects within the same class may vary dramatically as shown in Figure (1.1). This figure shows examples of objects with self-occlusions, for instance, as a result of their geometrical properties.

The visual processing in the primates' visual cortex was modelled using hierarchical models. Informative representation of the objects is extracted through a hierarchy of simple and complex cells of the developed models [2–4]. These models are based on hierarchical MAX operations, therefore, they are called hierarchical models. Recently, these models have shown an increased performance for recognising objects and solving the above dilemmas.

1.3 Applications of Object Recognition

Object recognition was recently used in several fields, due to the capabilities it offers for image understanding. Below is a discussion of the main applications that it can be utilised. However, it has potential applications in several different fields.

In robotics, object recognition can be developed to equip robots with a cognitive capability in which they are able to identify the objects being conceived. The robot can understand the environment by understanding the nature of objects in it, for instance, in an office, it is likely to observe a notebook, laptop, coffee mug and computer desk. This will enable robots to understand their surroundings and therefore become more equipped to handle the allocated task. In industrial applications, object recognition is essential for robotics, for instance, in consumers good industry, robots are required to handle different objects with different sizes and shapes. Recently, robotic vision was applied to a bionic hand for grasp recognition [7]. In this work, an artificial hand was equipped with a camera. The camera passes snapshots of images which is then processed to determine the type of the grasp for each object.

For mobile applications, object recognition is increasingly becoming an essential

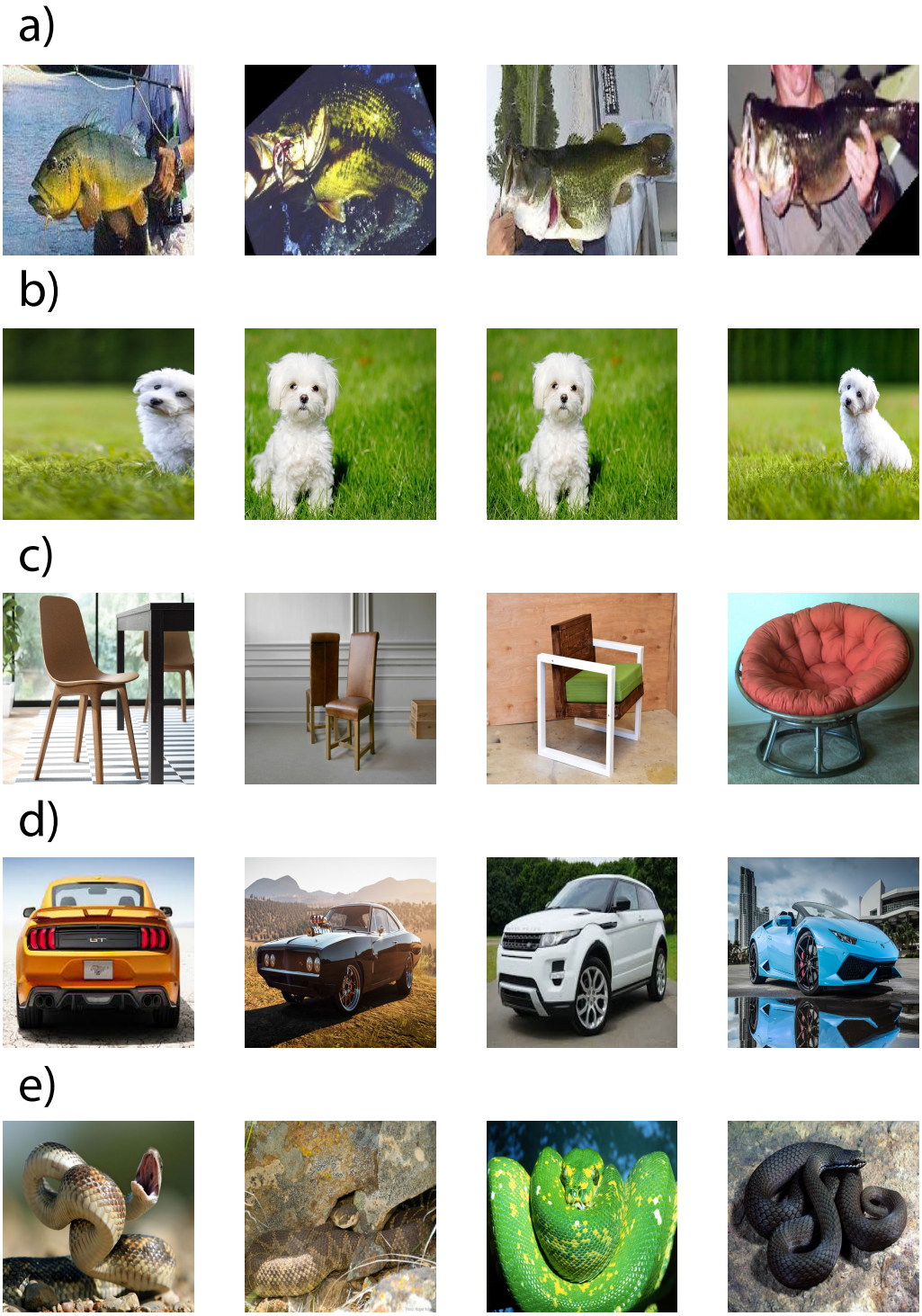


Figure 1.1: The real world scenarios for object recognition. a) Background clutter: backgrounds can obstruct the vision to make correct decisions. b) Object location: the location of the object within the image can alter the way that the model conceives the object. c) Intra-class variability: objects with similar classes can be extremely different in terms of structure and appearance. d) Orientation variance: the manifestations of objects may differ depending on the pose the image is taken from. e) Self-occlusion: objects may appear self-occluded due to their geometric properties. Images were collected from Caltech 256 dataset [6]

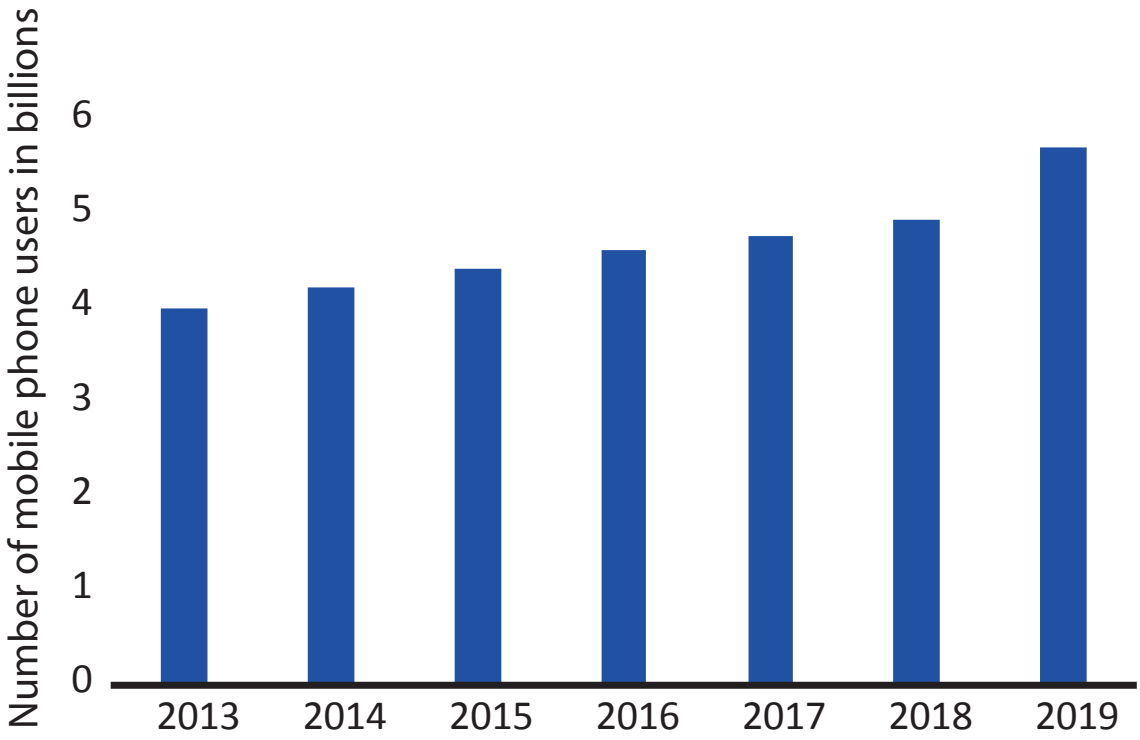


Figure 1.2: Estimated numbers of mobile phone users in the worldwide from 2013 to 2019 [8].

feature due to the availability of mobile phones with high computational capabilities. In 2016, studies have found that 62.9% of the world population are mobile users [8]. As shown in Figure 1.2, this number is expected to dramatically increase to reach 5 billion mobile phone users in 2019, for instance, in India, around 142 million mobile contracts were registered in 2011 [1]. This number is expected to reach 813.3 million mobile contractors in 2019 [9]. Therefore, mobile applications that provide an easier way to search the physical world is becoming more accessible, for instance, Google Goggles [10] and CamFind [11]. These applications help users to identify objects and scenes using the mobile camera and Cloud computing, for instance, identifying a film poster in a street, or recognising a famous landmark when travelling abroad without needing a text-based search.

Finally, object recognition models are being increasingly utilised in self-driving cars. Equipping cars with such technologies enable them to process frequencies beyond human capabilities, for instance, in order for the car to make the right decision in an emergency situation, it needs to know whether an obstacle is a person or not. Therefore, recognising objects in a moving vehicle is essential. While a car is driving a software can provide labels of objects that the car is encountering.

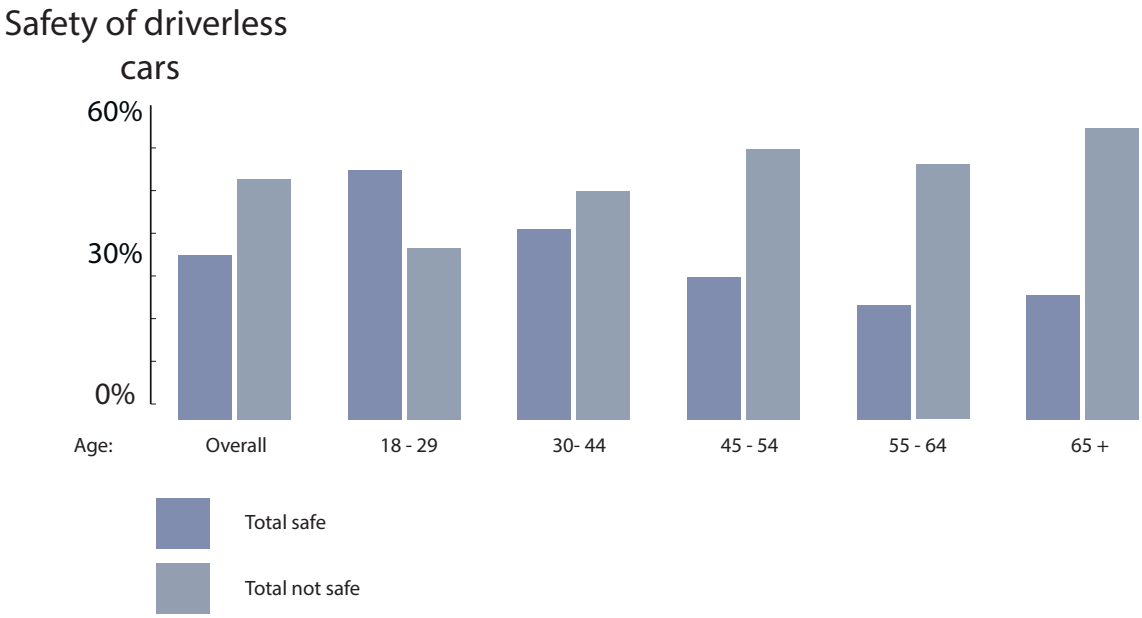


Figure 1.3: An estimation of the general public perspective regarding the safety of driver-less or autonomous cars [12].

Driver-less cars are still in the testing phase and recently there have been fatalities involving fully autonomous vehicles, for instance, one of the Uber autonomous cars has been involved in a fatal accident in Arizona and Toyota has suspended tests of self-driving cars on public roads [12]. This has changed the course of the general public perspective of the extent of the safety of driver-less cars. Figure 1.3 shows the perspective of the general public with respect to the safety of autonomous cars. The study also categorised the results within different age bands. The polls have shown that a majority of 43% thinks that self-driving cars are totally not safe. Therefore, the technologies of self-driving vehicles require further development.

1.4 Problem Statement

The main problem that was addressed in this thesis involved recognising objects under all form of transformations described in section (1.2) within a hierarchical model such that

- it can be used for multi-class recognition, i.e., not designed for binary classification;
- it is robust to the challenges mentioned in section (1.2);

- it is formed within an abstract architecture where formative features are extracted within a reduced number of layers;
- it has similar mechanisms to the human visual cortex, i.e., the feed-forward hierarchical architecture;
- it can achieve high recognition accuracy.

1.5 Aim and Objectives of this Thesis

The purpose of this thesis is to further the research on hierarchical architectures inspired by the human visual cortex to achieve better performances.

The particular objectives of this thesis are:

- Objective 1: to build a hierarchical model inspired by the visual cortex that address the issues discussed in section (1.2) and achieve higher accuracies and efficiency.
- Objective 2: to investigate the effectiveness of each region of vision for object and scene image dataset.
- Objective 3: to further enhance the recognition performance, a number of hierarchical topologies were formed, such that the recognition task considers the scene perspective for obtaining the identity of the objects.

1.6 Statement of Originality

The major contributions of this thesis can be summarised as follows:

- In Chapter 2, a comprehensive view is provided on the advantages and disadvantages of the available techniques of object recognition. These techniques were categorised into hierarchical models, histogram-based models and deep learning models. A survey was made to identify the problems and benefits of each technique. Furthermore, a survey was made with regard to the effective regions of vision on human subjects. It summarises the fact that human vision has biases toward the peripheral vision to recognise scenes and the reverse is true for objects.

- In Chapter 3, a novel model of object recognition is proposed, namely, the En-HMAX model. The En-HMAX model is hierarchical and feed-forward. It summarises basic facts of the ventral stream of the primates visual cortex. The En-HMAX model utilises the elastic-net regulariser for dictionary learning. It also uses learned filters for feature extraction. The model was tested with different datasets. The performance of the En-HMAX model was compared with other hierarchical models from the literature.

- In Chapter 4, the effective regions of vision are investigated using the En-HMAX model. The rationale of the experiments of this chapter was to quantify the contribution of the peripheral image content and the central image content to recognise scenes and objects using the En-HMAX model. To study the biases of computational models for recognition, two datasets were utilised. Also, along with the En-HMAX model, four computation models were used. This includes the classic HMAX model and state of art neural networks, such as GoogLeNet, AlexNet and VGG net. To quantify the contribution of each region of vision, the experiments involved modelling two paradigms, namely, scotoma and window.

- In Chapter 5, topologies that comprises shallow and deep models are formed. In order to enhance the performance in object recognition, it is proposed to change the order of the recognition process by using an initial stage to give an indication of the nature of the objects. Three topologies were proposed for this task. The topologies provide a trade-off between the decision sensitivity and the computational complexity. In summary, a top-level stage is used to categorise the nature of the scene that the object found, mainly formed using a shallow network. Then, another deeper stage is used to recognise the object.

- The thesis provides a comprehensive evaluation of hierarchical architectures for image processing and paves the way for future research to tackle issues reported in this thesis.

1.7 Thesis Organization

This thesis is composed of seven chapters, where the main challenges linked to hierarchical architectures for object recognition will be discussed in Chapter 3. While

in Chapter 5, the effective regions of vision will be covered. Chapter 6 will discuss optimised topologies for object recognition.

Chapter 1 "Introduction" presents the introduction, motivations, objectives and structure of the thesis.

Chapter 2 "Background and Literature Review" provides comprehensive background knowledge on models of object recognition. In particular, hierarchical models. It reviews the literature of hierarchical architectures inspired by the visual cortex in detail and presents datasets relevant to this thesis.

Chapter 3 "Object Recognition with an Elastic Net-regularised Hierarchical MAX Model of the Visual Cortex" presents a novel hierarchical model of object recognition. The newly developed model solve the problems of highly correlated images. It utilises the same hierarchy of the visual cortex. Additionally, it utilises techniques rooted in Neuroscience that help to provide better performances.

Chapter 4 "Objects and Scenes Classification with Selective Use of Central and Peripheral Image Content" studies the effective regions of vision using the developed model in Chapter 3. It highlights the significant difference in recognising object images and scene images. It proposes foveation to reduce the size of image data. It also discusses the potentials of using these techniques on Cloud processing.

Chapter 5 "Object Recognition Based on Understanding The Real World: Indoor Versus Outdoor Environments" proposes three hierarchical topologies to reshape the existing scheme of object recognition. It highlights the importance of understanding the scene as part of the recognition process. Additionally, it provides a trade-off between decision sensitivity and classification accuracy.

Chapter 6 "Conclusions and Future Work" summarises the advances achieved in this study with regard to hierarchical models for image processing and object recognition. It also shows the challenges that the available techniques still need to address. Furthermore, this chapter lists potential future work which can further the development of hierarchical architectures for object recognition.

Chapter 2

Background Theory And Literature Review

2.1 Introduction

Object recognition has been extensively studied and applied using many different approaches [13,14]. It is becoming one of the most important fields in image processing and computer vision [15]. It allows artificial intelligence programs to identify objects from inputs such as still camera images and videos. In this chapter, the main components of recent research methods of object recognition are covered. In particular, feature-based methods such as feedforward hierarchical models, deep learning models and histogram-based models. Feature-based methods are based on extracting informative features from the appearance of the object in the form of a two-dimensional image. These methods statistically describe the visual data in the real world using numbers. An excellent model of object recognition is able to represent different objects distinctively, for instance, bicycles and motorcycles. This chapter will also discuss the common methods of feature extraction and classification, for instance, principal component analysis, Independent component analysis, dictionary learning, elastic net regulariser and support vector machine. It will also show the advantages and the limitations of these approaches.

2.2 Choice of Coordinate System

The initial stages of this research involved determining a suitable coordinate system for object recognition. To perform object recognition, there are two approaches to determine the coordinate system: object-centred approach and viewer-centred approach [16]. The object-centred approach can be defined as representing the object as a three-dimensional entity. On the other hand, the viewer-centred approach is simply defined as representing an object in a natural way from a viewer perspective or a camera perspective. The rationale for selecting the appropriate approach was to extract more unique features that efficiently represent the objects for recognition. The choice of the approach allows only certain techniques and methods to be used at the cost of the others [17]. Therefore, the below subsections discuss the characteristics of both approaches in more details.

2.2.1 The Object-Centred Approach

The object-centred approach represents the object using a specific three-dimensional coordinate system [18]. This approach describes the object regardless of the camera location [19]. It depends more on the description of the shape of the object. However, to extract the object of complete data, sophisticated techniques are required for camera parametrisation and adjusting viewpoints [19]. Object-centred representation is based on understanding the object geometry, i.e., the remaining information when the object scale, orientation and position are removed from its description. The object is retrieved using the non-overlapping regions of the object from the three-dimensional space. Examples of these approaches are tetrahedral decomposition [20], octree [21] and voxel representation [22].

2.2.2 The Viewer-Centred Approach

The alternative method to the object-centred is the viewer-centred, which describes objects from a camera perspective. In the object-centred method, the three-dimensional objects are described using two-dimensional images within specific viewpoints. This approach is considered more accessible and computationally efficient than object-centred approaches. The matching in this approach is also more efficient as the comparison with the description is performed using only two dimensions.

In this method, the projection process is not required when performing the matching process [20]. Furthermore, this approach solves the problem of objects located far away from the camera system that exists in the object-centred approach. However, the viewer-centred approach has a number of drawbacks, for instance, to represent one object in the two-dimensional space, a large number of views of the same objects need to be stored for the matching process. Therefore, the object matching process becomes more complicated as the number of features is increased dramatically.

There is a strong evidence in neuroscience literature suggesting that the human visual system utilises the viewer-centred representation to perform object recognition [23–25]. Experiments have shown that the human visual system is able to recognise objects accurately and rapidly from a single viewpoint. This shows that those views of the object are already stored in the memory in its two-dimensional forms.

In the viewer-centred approach, the object appearance varies considerably from one view to the other. One approach to solving this problem is extracting invariant features from different viewpoints [26]. The relationship between angles, lines and the ratio of the lengths of the lines can provide invariance over the viewpoint. An efficient use of the above techniques can minimise the number of the object viewpoints required to solve the recognition task [26].

2.3 Methods for Object Recognition

There are two main methods to perform object recognition on images: image-based method and feature-based method [19]. The image-based method uses a direct representation of the images for recognition [27,28]. On the other hand, the feature-based representation depends on the shape information [2, 29]. More detailed description of each type of method is provided in the below subsections.

2.3.1 Image-Based Methods

The main characteristic of the image-based method for recognising objects is that the stored images are compared directly with the new images using the intensity of the images [19]. In this method, no features are extracted from the object. Only the object appearance characteristics are learned. In the feature-based approach,

however, the features of the images are used to describe the objects for recognition.

The image-based methods involve utilising a large image dataset that consists of images of objects taken from many different poses and with different lighting conditions. The similarity of images is calculated either by using low-level descriptors or by a whole image-based similarity measure [30].

The recognition process can take other forms in the image-based method. These processes can be categorised into rigid methods and flexible methods [30]. In the rigid methods, a template of the target photometry (or shape) of the object is primarily defined [19]. The template may be represented by an image. The image data is then compared to the template using different metrics, for instance, dissimilarity or the similarity measure [30]. The matching process is made when the metric optimal point is reached, i.e., the shortest distance from an image to the template.

This approach is considered effective in restricted object search where the search is limited to certain object types. These methods do not perform efficiently in the following scenarios:

- when the exact object shape is not known;
- when multiple object shapes are involved at the same time;
- when an unknown object background is used.

When the above scenarios are met, the flexible methods can be more appropriate for the recognition task. These methods are based on imposing many constraints on the appearance of the object, for instance, the object symmetry, smoothness and homogeneity. It also utilises a mathematical optimisation to determine the best fit for the input image data. Other flexible methods allow more flexibility that may cause the matching process to be computationally heavier [19]. Therefore, to conduct these methods, the initialisation point may need to be reinforced to be close to the correct solution.

The major drawback of the image-based approach is the variation of the object background. It also has shown limited performance in recognising objects in cluttered and partially occluded scenes [30]. It performs well when the objects are segmented from their backgrounds [19]. In the literature, there are several attempts to enhance these methods performance against occlusion, for instance, the small

Eigen-windows [31, 32]. However, these methods are considered computationally expensive in their search process [31].

In the image-based method, the performance is less influenced by the increasing geometric complexity as a result of performing the recognition task directly in the image domain [19].

2.3.2 Feature-Based Methods

This method for object recognition is based on understanding the object attributes from its appearance. It utilises stored objects attributes and corresponds to them when similar features are detected in the scene. This approach is successful when the background of the object is unknown. It can extract only the important features of an object within a background and accordingly create a code for this object for recognition. The recognition process is generally bounded by an error function that shows to what extent the newly observed features differ from the stored features [33].

The lack of scalability is one of the most common problems in object recognition, for instance, a large number of objects, or an extreme variation in objects appearance. The feature-based method has moved forward in a purposeful way to mitigate the lack of scalability. In particular, developing techniques that minimise the number of features that correspond to an object [34]. Additionally, minimising the possible number of matches [35, 36]. Similarly, a method of indexing was developed to limit the inappropriate matches using a priori information of the recognition task [37].

To achieve invariance to objects pose, several paradigms were developed to match an object within a particular viewpoint with a reference viewpoint [19,38]. However, this approach requires a huge amount of memory to understand the views of a large number of objects. For the feature-based method, there are many aspects and problems that will be discussed and addressed in this thesis. This includes the type of object features, methods for extracting these features reliably, the classification mechanisms and the computational complexity.

To summarise, the feature-based methods are considered more successful than the image-based methods for object recognition, especially for the cluttered environment [30]. Therefore, feature-based methods for object recognition will be primarily used for all tasks of object recognition in this thesis.

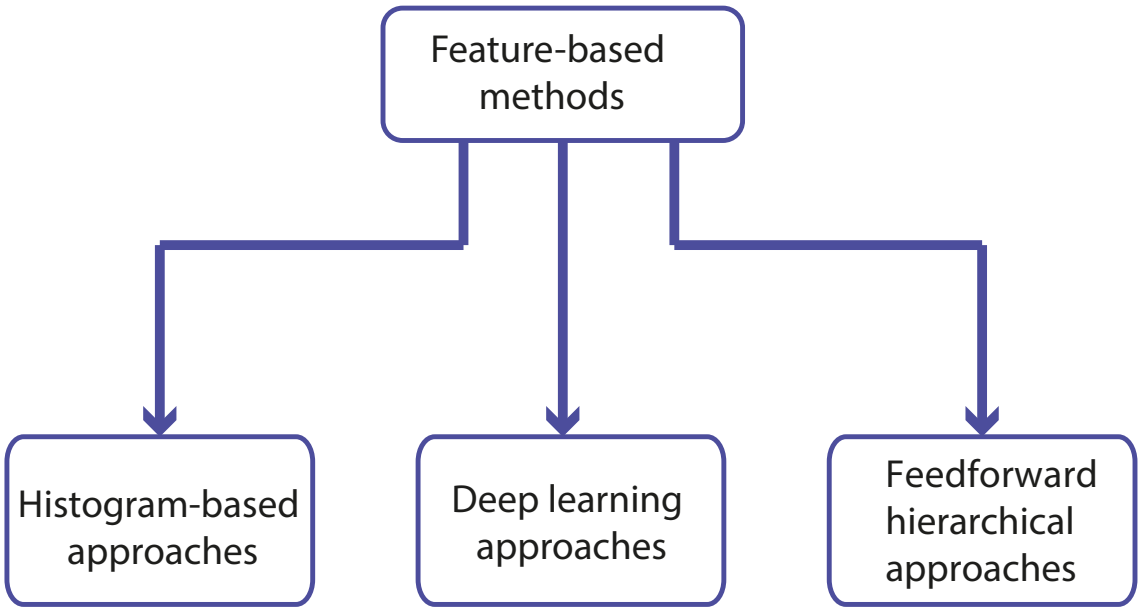


Figure 2.1: Main approaches of feature-based methods for object recognition.

The below sections describe the main existing approaches of feature-based methods for object recognition. These approaches were categorised into histogram-based approaches, deep learning approaches and feed-forward hierarchical approaches (as shown in the diagram in Figure 2.1)

2.3.2.1 Histogram-Based Approach

Histogram-based approaches have proved successful to generate invariant features for transformations, in particular, object translation [39]. The most common features are the scale-invariant feature transformation (SIFT) features [29]. The SIFT descriptor is based on converting the image into many feature vectors. These features are invariant to object transformations, local geometric distortion and illumination changes. Only the dominant features are extracted, small edges along stone edges features are discarded. This enables the SIFT descriptor to be robust to local distortion.

The mechanism of the SIFT descriptors for extracting features from the input images can be summarised as the following:

1. a set of key points are extracted from reference images of objects;

2. the extracted key points are then stored in a database;

3. a novel object in a new image is recognised by separately comparing each

feature of this new image to the features stored in the database. The Euclidean distance is used as metric to determine the matching between features;

4. a set of matches with the stored database is formed and stored;

5. the object is then identified depending on the number of matches scored with the stored database.

Comparing to other histogram-based approaches [40, 41], the SIFT features are considered the most robust to the transformations of objects. Many examples in the literature are also based on such descriptors, for example, the speeded up robust features (SURF) and Harris corner detector [40, 41]. However, experimental results [42] have shown that such descriptors may not perform well on a generic object recognition task, due to the limited degree of invariance they provide. Many other histogram descriptors, such as the square patch of an image [43], are incapable of capturing discrepancies after object transformation [44].

2.3.2.2 Deep Learning Approach

The deep learning is a class of machine learning algorithms. It has recently shown excellent performance on object recognition [45–48]. Convolutional neural networks (CNNs) were shown to function well especially for large object image dataset. Generally, deep learning methods comprise the following common elements [45]:

1. a series of non-linear layers of image feature processing. Each layer of the cascade receives the input from the previous layer's output;

2. a bank of filters in each layer are formed either using unsupervised learning methods or supervised learning methods;

At each layer of the cascade, deep learning models learn a different type of features. Advanced layers learn more complex patterns of the objects, while the first layers learn simple features, such as edges and lines. Deep learning methods attempt to use the same structure of artificial neural networks.

The ultimate aim of deep learning methods is to find the optimum representation of the image data depending on the structure of the model [46]. The higher level features of the deep model are defined based on the composition of low-level features in the previous layers. Learning features in many stages of a model allow for utilisation

of complex features essential for object recognition. It also allows the model to be independent of human-crafted features. Deep learning methods are now considered as one of the best platform to handle large datasets [48]. The performance of other object recognition models, however, levels out at a certain point. This is due to the finite size of other models where more data can cause such model to over-fit.

Deep networks depend massively on parameter optimization and tuning [49]. Additionally, training deep networks can be time-consuming. In order to avoid the massive time-consuming training stage for CNNs, state of art pre-trained CNNs were used in Chapter 6 of this thesis to form highly optimised topologies for object recognition.

2.3.2.3 Feed-Forward Hierarchical Approach

Hierarchical feed-forward models are inspired by the visual cortex of the primates [50]. In particular, the ventral stream pathway, a hierarchy of layers responsible for rapid object categorisation, see Figure 2.2. Hierarchical models provide a robust platform for object recognition using flexible and trainable features. The main examples of hierarchical feed-forward models are the hierarchical-MAX model [2, 3] and Fukushimas multilayer perception (Neocognitron) [51].

Experimental experiments on human subjects showed that humans are able to categorise objects in less than 150 ms [3]. Hierarchical feed-forward models utilise the following mechanisms of the ventral visual stream of the primates visual cortex:

1. Inspired by the primary visual area (V1) simple cells, these models achieve selectivity using excellent filters for feature extraction.

2. Inspired by V1 complex cells, invariance can be attained by down-sampling the response using the pooling operation [52].

Hierarchical feed-forward models have an abstract structure. This means that they can extract high-level features from the input images using few layers in their hierarchy. However, they perform well in terms of classification accuracy, especially in a cluttered environment. The neurons in the low areas are sensitive to low-level features such as edges and lines. The activation of these neurons is reflected in the next layer of the hierarchy, which extracts higher-level features. By the end of the hierarchy (V4), these models become more selective to object's shapes, regardless

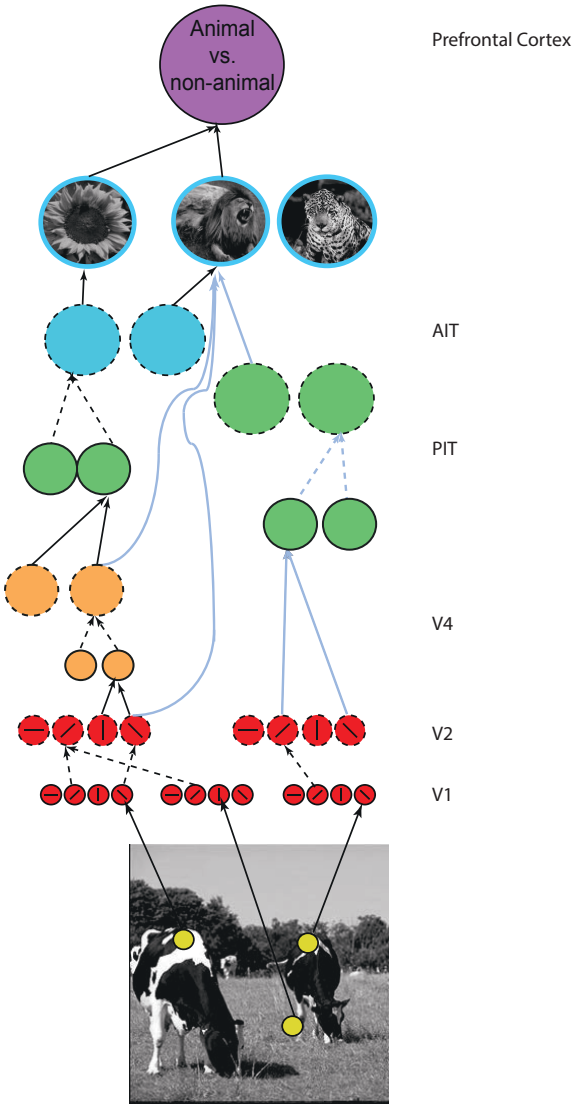


Figure 2.2: Illustration of the hierarchical architecture inspired by the visual cortex (taken from [2]).

of the viewpoint of the observed object. The processing of these networks is feed-forward, hierarchical and local in each layer. Furthermore, the processing in each layer depends on the processing at the previous layer. These models do not involve the mechanisms of perception and top-down processing of the mammal's visual system. Top-down processing is the cognitive process that flows down to lower-level functions and accordingly making decisions [53].

Feed-forward hierarchical models have the following advantages over the previously explained approaches for object recognition:

1. They can learn formative , i.e., class-specific and rich hierarchical features through their layers of hierarchy [3].

2. They perform the visual recognition task accurately with a high speed and

using an abstract structure [2].

3. They operate with a reduced computational complexity [54].

4. They efficiently generalise to objects with different backgrounds, orientations, scale and position [4].

In this thesis, the main tool that will be used for object recognition is hierarchical feed-forward model.

2.4 Common Methods of Feature Extraction and Classification

The main methods of feature extraction used in this thesis are explained in this section. In particular, principal component analysis (PCA), independent component analysis (ICA), dictionary learning, elastic net regulariser and support vector machine (SVM).

2.4.1 Principal Component Analysis

The PCA algorithm is a statistical approach that attempts to convert a set of observations of correlated variables into uncorrelated variables. It uses orthogonal transformation to extract the principal components for the input data. It can also be used to shrink the input data. The number of principal components represents the size of the output data. The first principal component corresponds to the highest possible variance of the input data. The succeeding component has the largest variance with the condition that it is orthogonal to the previous components. Therefore, the produced vectors are uncorrelated and orthogonal [55].

The PCA method is based on the eigenvector multivariate analyses. The internal representation of the PCA data provides a good description of the variances of the original data. Moreover, it provides a projection or view of the data from the most instructive viewpoint using the first few principal components [56].

Whitening Transformation

PCA whitening transformation is based on scaled PCA. It is a linear transformation of an input vector into new variables with a covariance of an identity matrix [57].

Each of the new variables has a unity variance, therefore, uncorrelated with one another [58]. The process transforms the input vector into a white noise vector, i.e., the components of the vector are statistically uncorrelated and have a probability distribution with zero mean and finite variance. Therefore, it is named whitening.

In two-dimensional images framework, the whitening transformation includes the following two steps:

1. Project the input data into the eigenvectors. As a result, the components of the dataset become uncorrelated.

2. Normalise the dataset so that each component has a variance of 1. This is achieved by dividing all the components by the square root of the eigenvalues.

For an input image data $\mathbf{X} \in \mathbb{R}^{m \times n}$ that contain m features and n data points, the covariance matrix $\mathbf{C} \approx \mathbb{E}[\mathbf{X}\mathbf{X}^T]$ can be estimated from the data matrix as follows:

$$\mathbf{C_X} \approx \frac{\mathbf{X}^T\mathbf{X}}{n}. \tag{2.1}$$

The covariance matrix ($\mathbf{C_X}$) has eigenvectors in the columns of $\mathbf{E}$ and eigenvalues on the diagonal of $\mathbf{D}$, as the following

$$\mathbf{C_X} = \mathbf{E}\mathbf{D}\mathbf{E}^T. \tag{2.2}$$

The desired PCA whitening transformation matrix $\mathbf{W_{PCA}}$ is given by

$$\mathbf{W}_{PCA} = \mathbf{D}^{-1/2}\mathbf{E}^T \tag{2.3}$$

The whitening transformation is considered as a decorrelation process, scaled by the inverse of the square root of $\mathbf{D}^{-1/2}$. The covariance of the transformed input data is the identity matrix, meaning that they are uncorrelated and each has variance 1.

2.4.2 Independent Component Analysis

Independent component analysis (ICA) is a well-known tool in signal and image processing used to separate signals into subcomponents [59]. It is an important statistical technique for extracting hidden variables from an observed random measurement. It can be considered as an extension to the PCA method [60].

The ICA method was applied in many fields, for instance, digital images, economic indicators, psychometric and document databases. However, the main application of the ICA method is the blind source separation. Blind source separation can be explained as a mixture of signals recorded simultaneously. Then, the signals are decomposed [59]. The input measurements of the ICA method are usually given as a time series or a set of parallel signals.

The independent components of the ICA method are assumed to be statistically independent [61]. The two random variables s_1 and s_2 are said to be statistically independent if the information in s_1 does not provide any information about s_2 and the reverse is also true. Mathematically, however, the above two variables are considered independent only if the joint probability density function (pdf) is factorisable to the product of the marginal distribution as the following:

$$P(s_1, s_2) = p_1(s_1)p_2(s_2), \tag{2.4}$$

where p_j denotes the joint probability. In a compact form, the equation can be rewritten as:

$$P_j(\mathbf{s}) = \prod_{i=1}^{N} p_i(s_i). \tag{2.5}$$

The distribution of the independent components of the ICA method is generally non-Gaussian. Also, it is assumed that both the independent components and the mixture variables have zero mean. The blind source separation scenario is used here to mathematically formulate the ICA method. The observed signal $\mathbf{x}$ (whose elements are the linear mixture $x_1, ..., x_n$) can be decomposed using a vector of the independent components $\mathbf{s}$ (whose elements are $s_1, ..., s_n$) and the mixture matrix $\mathbf{A}$. This can be formulated as the followings

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.6}$$

where $\mathbf{A}$ and $\mathbf{s}$ are both unknown.

The task of the ICA method is to recover the signal of the original sources, i.e., the independent components $\mathbf{s}$. To further simplify the above formulation, the number of voices are assumed to be similar to the number of the observed mixtures. Therefore, the coefficient matrix $\mathbf{A}$ is always square. As a result, the matrix $\mathbf{A}$

is invertible, producing the unmixing matrix $\mathbf{W}$. The signal $\mathbf{s}$ can, therefore, be computed as follows

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{2.7}$$

The ICA method describes how the observed measurements are generated by the process of mixing the components of $\mathbf{s}$. Therefore the ICA model is considered as a generative model. In this thesis, the ICA method is used to generate filters from natural scene images. Whitening transformation is first applied to random patches of the natural scene images. Then, the ICA method is used to extract the independent component of these patches, creating excellent filters used to extract low level features from the input images.

2.4.3 Dictionary Learning

Dictionary learning is a method that aims to find a sparse representation to the input data [62]. It is also known as sparse coding. The process involves generating a linear combination of elements to reconstruct the input data [63]. The combination of these elements form a dictionary and each element is called an atom. There are several important features in dictionary atoms, for instance, they maybe over-complete set (i.e., the dimension of the produced signal may be larger than the dimension of the input signal) and unlike the ICA method, they are not necessarily orthogonal [64]. The above features allow dictionary learning to have multiple representations of the input signal. It also provides more flexibility and sparsity in its representation.

The main rationale of dictionary learning is to reconstruct an input data using a dictionary and a sparse code, i.e., an efficient representation that capture the pattern and structure of the input data [62]. It can reproduce the input data with the minimum possible number of atoms. Prior to dictionary learning, the common approach was using hand-crafted dictionaries, for instance, wavelet transformation and Fourier. Ideally, when a dictionary fits the input data, the sparsity can efficiently increase. This phenomenon has many applications in data compression, decomposition, image de-noising, audio and video processing, image fusion, in-painting and image analysis classification.

In this thesis, dictionary learning is used to process image data. Given an image data $X = [x_1, ..., x_k], x_i \in \mathbb{R}^d$, the objective of the dictionary is to compute a dictio-

nary $\mathbf{D} \in \mathbb{R}^{d \times n}$. Both the dictionary matrix $D = [d_1, ..., d_n]$ and the representation matrix $R = [r_1, ..., d_k]$, are minimised such that each representation vector $r_i \in \mathbb{R}^n$ in $\|\mathbf{X} - \mathbf{DR}\|_F^2$ is sparse enough for reconstructing the images. This can be formulated as follow

$$\begin{aligned} \underset{\mathbf{D},\mathbf{R}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{DR}\|_F^2 + \lambda_1 \|\mathbf{R}\|_0 \\ \text{subject to} \quad & \|\mathbf{d}_i\|_2 \leq 1, \quad i = 1, \cdots, n, \end{aligned} \tag{2.8}$$

where $\|.\|_F$ denotes the Frobenius norm and λ_1 is the regularization coefficient. The constraints on d_i are introduced to prevent the atoms from extending into high values for lower values of r_i.

The above minimisation problem is non-convex due to the existence of the ℓ_0 norm. In other applications, the ℓ_1 norm regulariser or the least absolute shrinkage and selection operator (LASSO) regulariser is used to introduce more sparsity [65]. As a result, the above minimisation problem becomes convex with respect to the variables $\mathbf{D}$ and $\mathbf{R}$ individually, i.e., fixing $\mathbf{D}$ and solving for $\mathbf{R}$. However, as a whole, the minimisation problem is not jointly convex. Similar facts can be observed when introducing the ℓ_2 norm regulariser or ridge regression.

In Chapter 3 of this thesis, a combination of both LASSO and ridge regression that form the elastic net regulariser is used to train dictionaries. The below subsection will discuss the elastic net regulariser in more detail.

2.4.4 Elastic Net Regulariser

The elastic net regulariser is a hybrid of LASSO and ridge regression. In particular, it combines both the ℓ_1 norm and the ℓ_2 norm penalties [66]. In-line with LASSO regulariser, the elastic net regulariser can shrink the input data by introducing zero-valued coefficients [67]. Research studies have suggested that the elastic net regulariser can outperform the LASSO regulariser for processing highly correlated predictors [68]. The Elastic net regulariser is also considered as a robust platform that encourages the grouping effect. In particular, for highly correlated predictors. Unlike the LASSO regulariser, the elastic net regulariser is robust when the number of observations is considerably less than the number of predictors.

In image processing, the elastic net regulariser can be used in dictionary learning. A dictionary $\mathbf{D} \in \mathbb{R}^{m \times p}$ and sparse weighting matrix $\mathbf{S} \in \mathbb{R}^{p \times n}$ can reconstruct an

input image dataset $\mathbf{X} \in \mathbb{R}^{m \times n}$. In the matrix notation, sparse coding is formulated as $\mathbf{X} = \mathbf{DS}$. To learn the dictionary $\mathbf{D}$ and the sparse weighting matrix $\mathbf{S}$, elastic-net regularization can be formulated as the following:

$$\begin{aligned} \underset{\mathbf{D},\mathbf{S}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 + \lambda_2 \|\mathbf{S}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{d}_i\|_2 \leq 1, \quad i = 1, \cdots, p, \end{aligned} \tag{2.9}$$

where $\|.\|_F$ denotes the Frobenius norm and $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$ are the regularization coefficients that regulate the trade-off between sparsity and the sensitivity of basis selection. When $\lambda_1 = 1$ and $\lambda_2 = 0$, equation (2.9) reduces to the ℓ_1 coding method described in [4, 69], hereafter called the LASSO and when $\lambda_1 = 0$ and $\lambda_2 = 1$, equatioin (2.9) reduces to another extreme case, called ridge regression.

Figure 2.3(a) illustrates the solutions of LASSO and ridge regression. It visualises the solutions of them by plotting their loss function (sum of squares) equicontours using the least square solution. It displays the points where the equicontour touches the edge of the regularisers function [68]. For a large value of λ, the area of the penalty constraint become larger. The example shown in Figure 2.3(a) is performed using only two dimensions (β_1 and β_2). The optimum estimate of the model parameter occurs at the point where both contours intersect with each other. The LASSO regulariser function is an ℓ_1-ball centred at the origin, while the ridge regression function is an Euclidean ℓ_2-ball. As a result of the sharp corners of the function of LASSO, the solution of the whole minimisation problem is likely to touch its function at an axis point. Therefore, one of the two dimensions will always be zero. For a large dimensional scenario, the solution using the LASSO regulariser will, therefore, be extremely sparse.

The penalty region of the elastic net regulariser is shown in Figure 2.3(b). It can be observed that the elastic net penalty region is a mix between the ℓ_1 norm and the ℓ_2 norm regularisers. The elastic net regulariser is more robust to multicollinearity due to its curved constraint region. It shows sharp corners due to the utilisation of the LASSO regulariser in its penalty function. Therefore, it is equipped with an aggressive variable selection property.

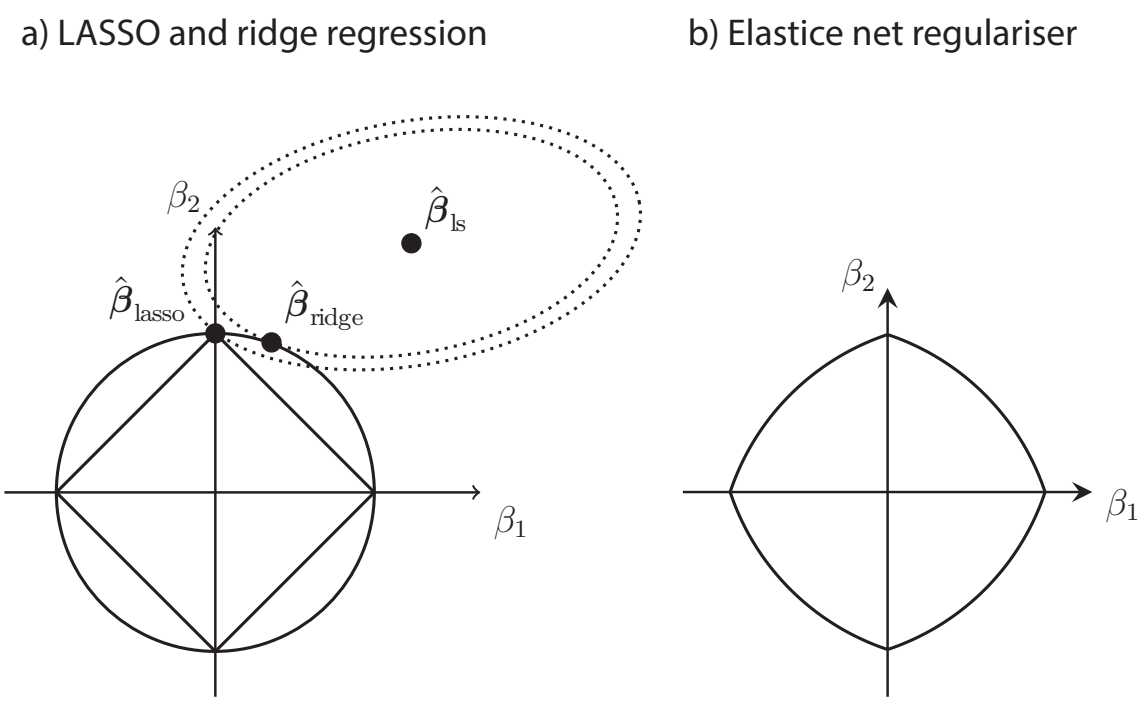


Figure 2.3: a) The visualisation of the two-dimensional solutions of the least square problem for LASSO regulariser and ridge regression. b) The constraint region of the elastic net regulariser.

2.4.5 Support Vector Machines

Support vector machines (SVM) is becoming more popular in many applications including object recognition [70]. Here, for completeness, a brief description of the SVM method mechanisms is mentioned. The SVM is a supervised learning method for regression analysis and classification. For each set of training data within a specific class, the SVM method can build a model that allocates the examples of data to one class or the other [71]. The SVM model represents the examples as mapped points in space, as such, the points of different categories are separated with the largest possible gap. Therefore, the new predicted examples are converted to the same space and a prediction take place to decide to which class they belong [70].

The SVM classifier utilises a hyperplane to linearly separate the training data in order to reduce the error for classifying the unseen test data. The optimal hyperplane is estimated using the training data by selecting a weighted combination called the support vectors. In this thesis, a multi-class linear SVM [70, 72] implemented within the LIBLINEAR library [72] was selected as the main classifier due to its computational simplicity.

The SVM method was applied to efficiently process large-scale data to tackle

problems regarding the face detection [73]. In the wavelet domain, the SVM method was also used for detecting pedestrians faces [74]. In [26], the SVM method was used to recognise the pose of a face image, this study involved a large image dataset of faces, the face recognition process was performed using the eigenfaces method. The SVM was widely used for object recognition. In [75], SVMs were used to recognise objects without feature extraction. Furthermore, a combination of the HMAX model and the SVM classifier was developed to recognise unseen images of objects [2].

2.5 Related Work

This section will present the literature on object and scene recognition. It will also present a literature of the common challenges, such as the occlusion and the used methods to overcome them. It will then present a thorough literature regarding the importance of the different regions of vision for recognising objects and scenes. Finally, it will provide a brief literature in reference to the environment of the object and its importance in the recognition process.

Machine vision has become an essential component of many human-computer interaction applications [15, 76]. By augmenting computers and robots with artificial vision, they have become capable of observing and (partially) understanding surrounding environments [77, 78]. Yet, reliably distinguishing objects and animals in arbitrary and cluttered backgrounds has remained challenging. This is because representations can differ considerably depending on position, orientation and scale [79]. The recognition performance of many computer vision algorithms, however, declines when the object is rotated or shifted excessively [40, 80].

In contrast, biological systems can recognise an object with different positions, orientations and scales following a single observation [29]. In addition, they can generalize to identify new objects, within the same category. Machine vision systems should, therefore, be able to similarly recognise and classify novel objects.

Neuroscience experiments in rodents, e.g. [81], non-human primates, e.g. [3] and humans [82] support the hypothesis that the visual cortex can be approximated with a feed-forward multi-layer structure. This architecture has inspired the multi-layer hierarchical MAX (HMAX) model [3]. Recently, the HMAX model, shown in Figure 2.4 was implemented for use in real-time object classification applications [83, 84].

The primary visual cortex of the brain uses sparse coding to encode input data

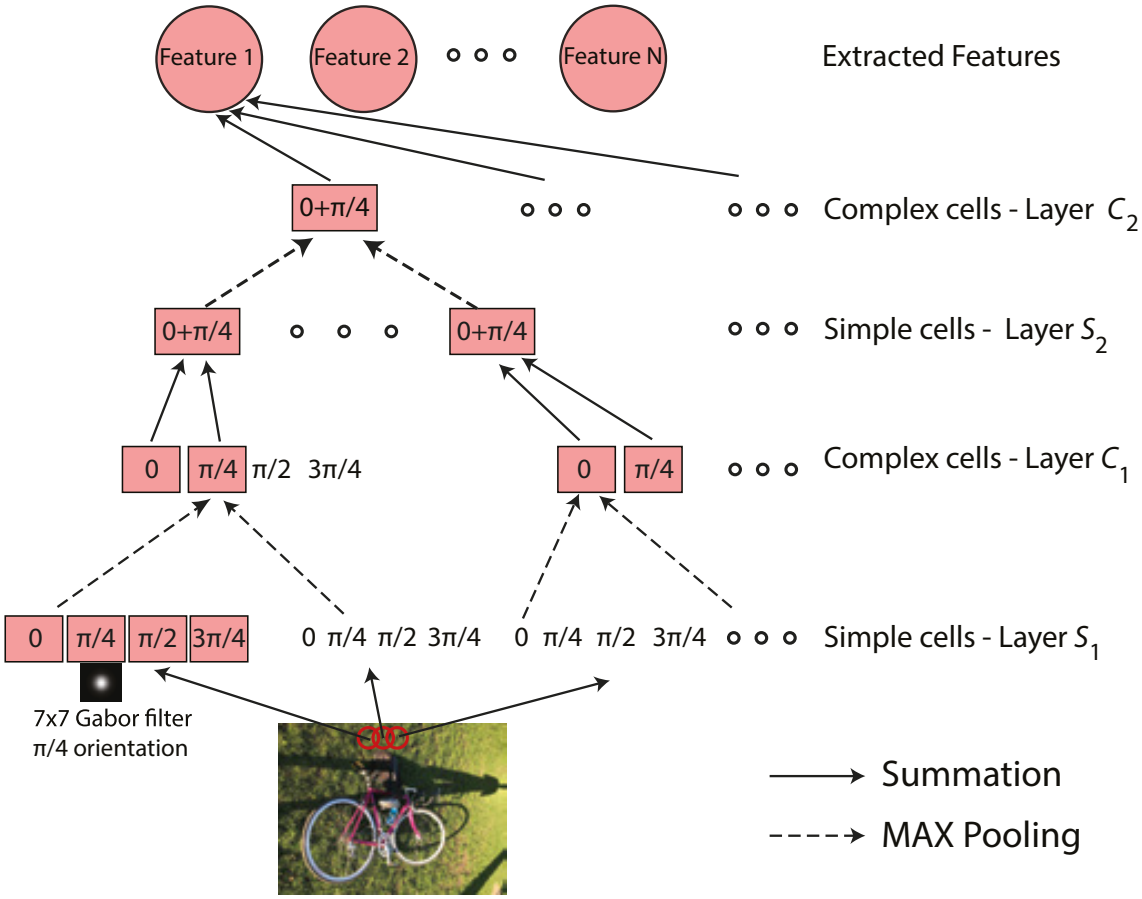


Figure 2.4: The basic structure of the original HMAX model.

by strong activation of a relatively small set of neurons [79, 82, 85]. Whilst, this method improved classification performance, the LASSO regulariser discarded the grouping effect in the higher layers of the model [69].

Despite progress towards more accurate object recognition, partial occlusion remains the main challenge even to state-of-the-art object recognition models. Recognising partially occluded objects is an essential problem in visual processing. Researchers have introduced different types of solutions to address this problem [86,87]. However, a common factor in their approaches was modelling the presence of occlusion. Typically, these methods depend on the occluder, where the training data comprises both the occluder and the occludee. The key limitation of these methods is their limited generalisation capabilities as they are application specific.

For example, when recognising a car parked in a road, the visible part of the car and the occluder form a pattern that continuously repeats itself. These patterns were sufficient to achieve correct classification.

One other approach to handle occlusion was utilising the statistical inconsistency

between occluder and occludee. Girshick et al. [88] developed a grammar model to detect occlusions. It is based on representing objects using many segments within a defined structure, to perform recognition. In [89], three-dimensional sensors have been utilised to determine the depth inconsistency, whereby occlusions are located and isolated. Similarly, In [90, 91], local similarity has been used to decorrelate partial occlusions. However, the above approaches may not be practical in all classification scenarios. The performance decreases drastically whenever an image feature does not match the ideal classification scenarios.

The human visual system processes the peripheral and central information of the visual field using a sophisticated retinotopic mapping. The peripheral and central representation of the visual scene can be found in low- and mid-level areas of the visual cortex, for example, V1-V4 [92]. Recognising objects depends more on details associated with central data while recognising buildings and scenes are associated with peripheral data [93]. The fMRI records showed more brain activity in fusiform face area (FFA) when recognising centre-based data such as faces [94] and words [95]. However, more activity was registered in the parahippocampal place area (PPA) during the recognition of peripheral-based data, such as buildings [96, 97]. The mid-fusiform sulcus (MFS) segregate the peripheral-biased pathway and the central-biased pathway in order to enable parallel processing [98].

The human visual system provides a compromise between fields of vision and its resolution [99]. It reduces the size of the processed visual data by using lower neural resolution in peripheral vision [100]. The retina needs to compress information from 100 million photoreceptors to only 1 million ganglion cells, suggesting a compression ratio of at least 100:1 [101]. The compression objective is presumably to reduce the energy cost of both the transmission and subsequent processing. Much can be learned from this hierarchy in the next generation of machine vision systems.

The angle formed by the two extremities of a viewed object is referred to as the visual angle. Behavioural research shows that the highest level of recognising objects in a particular form within a certain environment can be achieved within a range of visual angle from 1° to 2° of the fixation point [102, 103]. Object perception drops gradually as the eccentricity, that is anatomical segregation of peripheral versus central visual field bias, increases. The concentration of both photoreceptors and ganglion (communication) cells decreases with eccentricity from the fovea. As such, resolution and perception decreases into the peripheral vision. Therefore, the highest

visual acuity is perceived, when a human observer focuses on the central fovea [104]. However, the speed of visual processing is greater in the periphery [105].

Recently, several computer models of human vision have been developed [76,106–108]. Such models are based on coding the real-world visual data into numerical values, that enables the computer to meaningfully interact with the natural environment [77]. Despite progress towards a more accurate object and scene recognition, machine-based vision still falls short of the human recognition capability. Developing systems which mimic the visual cortex is a challenging, yet possible, path to achieving comparable performance.

In a similar context, studies have shown that models that function well in an indoor environment, perform poorly in an outdoor environment and the reverse is also true [109]. This is due to the stark difference in local and global properties of both environments. The daily life environment, such as living-rooms and city streets, comprises a large number of objects. The nature of these objects depends on the context in which they can be found. Current algorithms of object recognition are trained to recognise objects regardless of their context, dismissing all the information in the backgrounds. This poses a great difficulty for these models to make logical decisions.

Scene understanding is a necessary stage that provides important information about the possible object's identity. Identifying the scene can dramatically reduce the probabilities of the object identity and therefore increasing the recognition chance level. For example, outside in a desert, it is more likely to expect a camel than a microscope. Context-based recognition that depends on the environment characterises the object recognition process.

2.5.1 Limitations and Challenges

Several issues and challenges remain untackled for the recent models of object recognition which affect their performances. These are:

- The transformations of objects, for instance, orientation, scale, position and surrounding background. The performance of current models of object recognition decreases dramatically due to the introduction of the above factors.

- Recognising objects that are highly correlated, for instance, motorcycles and

bicycles. Models of object recognition show decreased performance to classes with a high degree of correlation. In particular, local features and backgrounds.

- Recognising objects under partial collisions because the performance of the existing object recognition models drops dramatically in the presence of occlusions.

- The increased number of layers in recently developed object recognition models makes their implementation more complicated, computationally expensive and slow to fine-tune.

2.6 Datasets

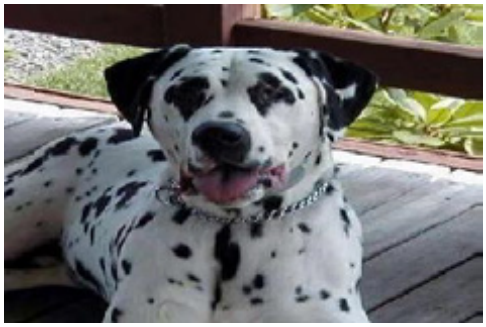
This section describes in detail the main datasets employed in this thesis. Also, it explains their properties and challenges. There are also other smaller image datasets used in this thesis. More description of these will be provided in separate sections in the following chapters.

2.6.1 Caltech 101 Dataset

The Caltech-101 image dataset [110] is a well-known object image database. It is considered as one of the main benchmarks in the field of machine vision. It comprises 9144 different images in 102 classes including the Google background category. The Google background category consists of 468 images randomly selected from Google images. Instead of forcing an object recognition model in making a wrong class label in classification scenarios where the model cannot identify an object, this class is added to the dataset to solve this problem. It corresponds to the uncertain decision that a model might predict when trained with Caltech 101 dataset. The size of each image is approximately 300×200 pixels. It has a wide range of classes, for instance, bonsai, dolphin, leopards and accordion. In addition, It is considered diverse in terms the shapes, sizes, scales and orientations of the objects. However, the number of images per class is inconsistent, where some categories have 31 images per class and other categories have up to 400 images per class. This dataset is tremendously challenging because images are of different size, illumination, appearance, viewing angle and orientation. Some of them are portrait and the others are landscape.


Brain


Crab


Dalmatian


Euphonium


Barrel


Leopard

Figure 2.5: Samples of the images in the Caltech 101 image dataset [110].

The Caltech 101 dataset was used in most of the experiments in this thesis, for instance, in Chapter 3, Chapter 4 and Chapter 5. Some examples of Caltech 101 dataset are shown in Figure 2.5.

2.6.2 Fifteen Scene Categories Dataset

Fifteen scene categories [111] dataset is a well-known scene image dataset. It consists of fifteen classes of scene images. The dataset was first collected by Fei-Fei and

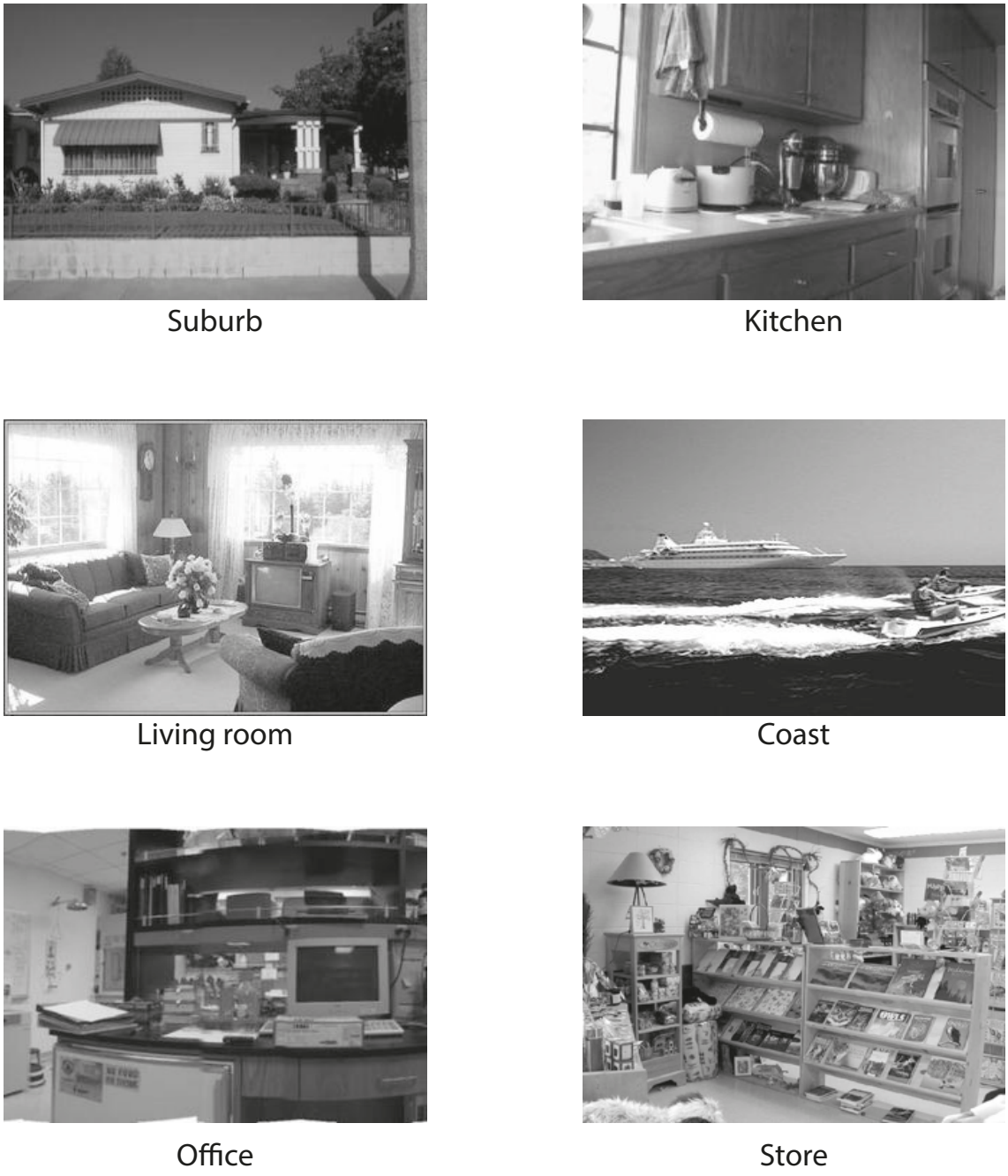


Figure 2.6: Examples of the images in fifteen scene categories dataset [111].

Perona [112]. It was then upgraded to include fifteen image classes in [111]. They have collected the dataset from personal photographs, Google image search and other datasets.

The dataset contains a plethora of scene images, for instance, natural scenes (such as forest and beach), man-made indoor scenes (such as kitchen and store) and man-made outdoor scenes (such as streets and buildings). Each class consists of 200 to 400 images, with an average image size of 300 $\times$ 250 pixels. The classes were: bedroom, CAL suburb, industrial, kitchen, living room, MIT coast, MIT forest,

MIT highway, MIT inside city, MIT mountain, MIT open country, MIT street, MIT tall building, PAR office and store. This dataset is considered as one of the complete scene category datasets in the literature thus far.

The dataset was used in Chapter 3, Chapter 4 and Chapter 5 of this thesis. Some examples of fifteen scene category dataset are shown in Figure 2.6.

2.7 Chapter Summary

In this chapter, an introduction to the available methods of object recognition systems was presented. It has included the main coordinate systems used for recognition, such as object-centred approach and viewer-centred approach. Then, an overview of the available methods of the viewer-centred approach was given. This has involved image-based methods and feature-based methods. Next, the main approximations of the feature-based methods were discussed, for instance, histogram-based methods, deep learning method sand feed-forward hierarchical methods. The main processes that were used in this thesis were also highlighted. The common techniques for feature extraction were discussed. A relevant literature of the recent methods was then provided. Also, the main unsolved limitations and challenges in object recognition were highlighted. Finally, a synopsis of the datasets which will be used in the proposed contributions in chapters 3, 4, 5 and 6 was introduced.

The next chapter will provide a developed feed-forward hierarchical model of the visual cortex for object recognition. It will describe the limitations of the recently developed hierarchical models. It will then discuss the main features of the newly developed model. It will also show comparisons with other hierarchical models on two image datasets.

Chapter 3

Object Recognition with an Elastic Net-regularised Hierarchical MAX Model of the Visual Cortex

3.1 Introduction

This chapter focuses on object and scene recognition in applications where the environment is cluttered, for instance, recognising a cup from a different point of view, different colours and different scale and position. Therefore, a new machine vision model for object recognition is introduced.

Since mammals achieve superior performances for observing the environment and intelligently classify objects [2, 113], the designed model focuses on replicating basic facts about the mammal's visual cortex. In particular, the ventral visual stream, a hierarchy of the visual cortex areas responsible for object recognition in the brain [3]. In order to improve the performance of the currently available models of object recognition, an elastic net-regularised dictionary learning approach was developed for use in the HMAX model. The model was termed as the En-HMAX model. The En-HMAX model developed in this chapter is designed to mitigate the limitation points mentioned in section (1.4). Therefore, the major contributions of this chapter are

- providing features invariant to transformations, for instance, object orientation, scale and rotation.
- learning formative features from the highly correlated data.

- performing the recognition process using an abstract architecture.

This chapter discusses the steps and the methods used for designing the En-HMAX model. This includes the architecture of the original HMAX model. Also, the features and methods introduced in the En-HMAX model are explained. The datasets used to test the En-HMAX model is presented. The results were then presented and the statistical analysis was conducted. With the En-HMAX model, the sparsity-grouping trade-off was exploited, such that correlated but informative features are preserved for object classification. As a result, the developed model was robust to the challenging daily-life environment, for instance, occlusions. A comparison was made with other available models. Finally, a summary of the chapter is provided.

3.2 The HMAX Model

The HMAX is a well-known object recognition model that attempts to mimic the same mechanisms of the primates visual cortex, a hierarchy of visual areas that mediate object recognition in the brain [2, 3]. It summarises the basic facts of the ventral visual stream of the visual cortex with similar hierarchal structure [3]. It was first developed to achieve invariances with similar shape tuning properties of the neurons in the mammalian's inferotemporal cortex, the highest visual area in the hierarchy of the visual cortex [2,113]. The model has proved to be successful in many datasets, also providing interesting perspectives. It provided a simple computational platform that is physiologically plausible to explain the cognitive processing. The below section explains the HMAX model with more explicit details.

3.2.1 Gabor Filter and Other Operators

In the standard HMAX model, Gabor filters [114] with different scales and orientations were used to filter the input images in the first S_1 layer. Gabor filters have received significant consideration in image processing due to their excellent capabilities for extracting features. The outline of simple-cell receptive fields in the mammalian cortex can be modelled by implementing two-dimensional Gabor function [114, 115]. The frequency and orientation illustrations of Gabor filters are similar to those of the human visual system [116].

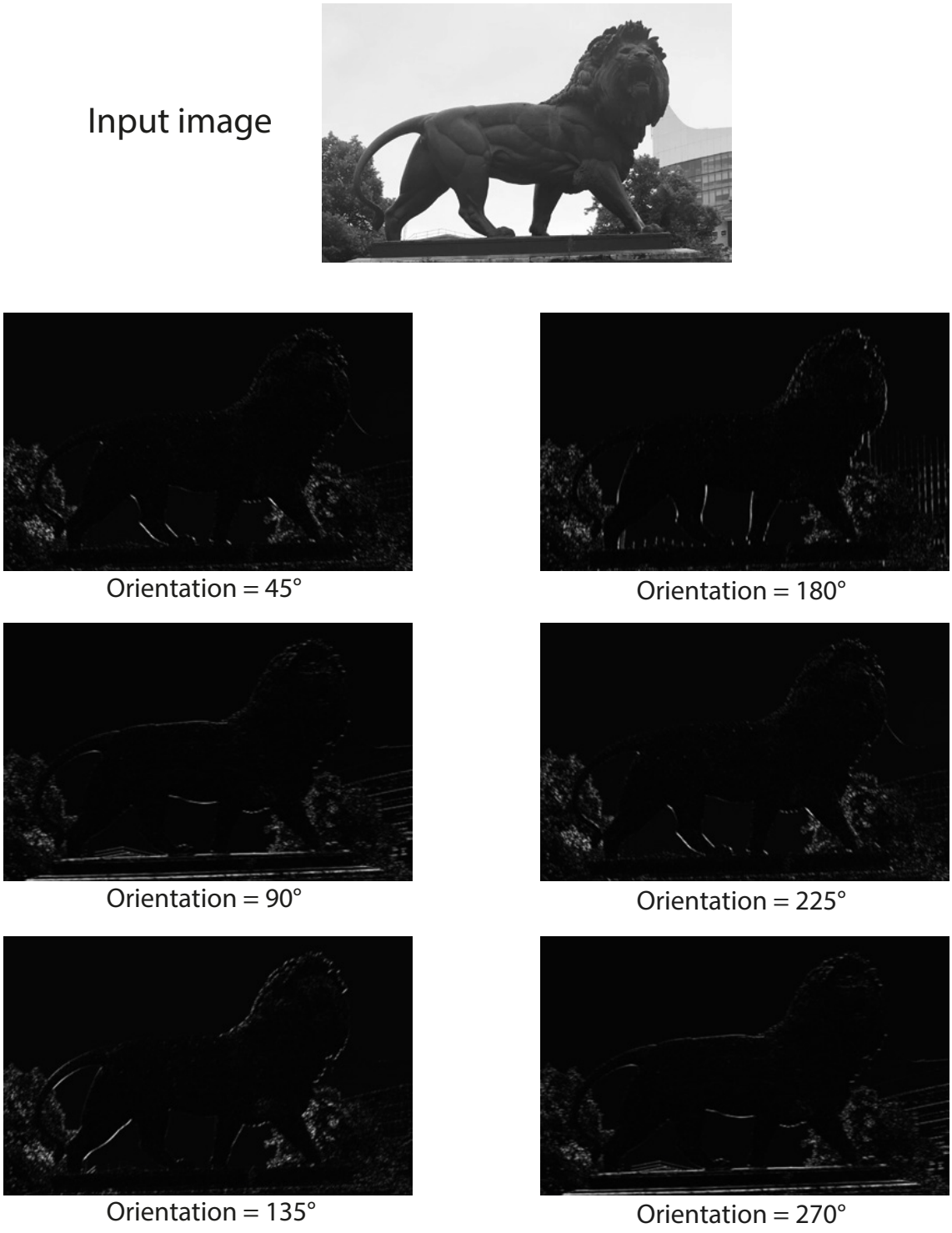


Figure 3.1: Response of the input image above to Gabor filters with six orientations.

Gabor filters are able to obtain textures of the entities within the images [117]. Therefore, they are considered an excellent feature extraction platform for two-dimensional images. The filters coefficients can be generated by multiplying the Gaussian kernel function by a sinusoidal wave [118].

There are numerous applications that use Gabor filters, such as for fractal di-

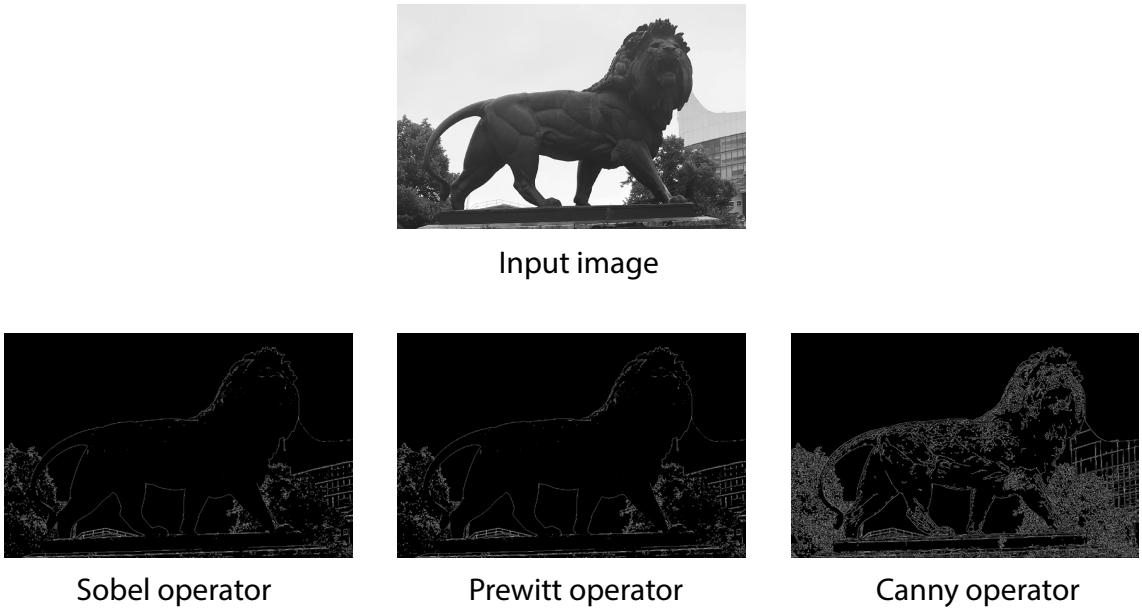


Figure 3.2: Classic edge detection operators applied to an image.

mension management, image coding, image representation, texture representation, target detection, edge detection, discrimination, retina identification and document analysis [114–120]. Figure 3.1 shows the original image (above) and the corresponding Gabor filter results of different orientations (below). The parameter values of Gabor filters could be modified depending on the task. Figure 3.1 shows the results of using Gabor filters with different orientations.

The first stages of mammalian vision system involve extraction of the edges and local features [119]. Edges are considered as an important feature of the images. Edge detection operators outline the surfaces and objects of the scene and disregard unimportant details. The followings are other well-known operators that perform edge detection: Sobel, Robert, Prewitt and Canny operator [121]. Figure 3.2 shows examples of using the three of these operators applied on the above input image. The main difference between these operators can be summarised by the kernel type and the smoothing mechanism. The mechanism of computing the gradient in two-dimensional images is different from one operators kernel to the other [120].

The features of the S_1 layer of the HMAX model are found by a bank of Gabor filters, resembling the cortical simple cell receptive fields which respond to the input activation within a particular orientation, scale and position. These filters can be represented with:

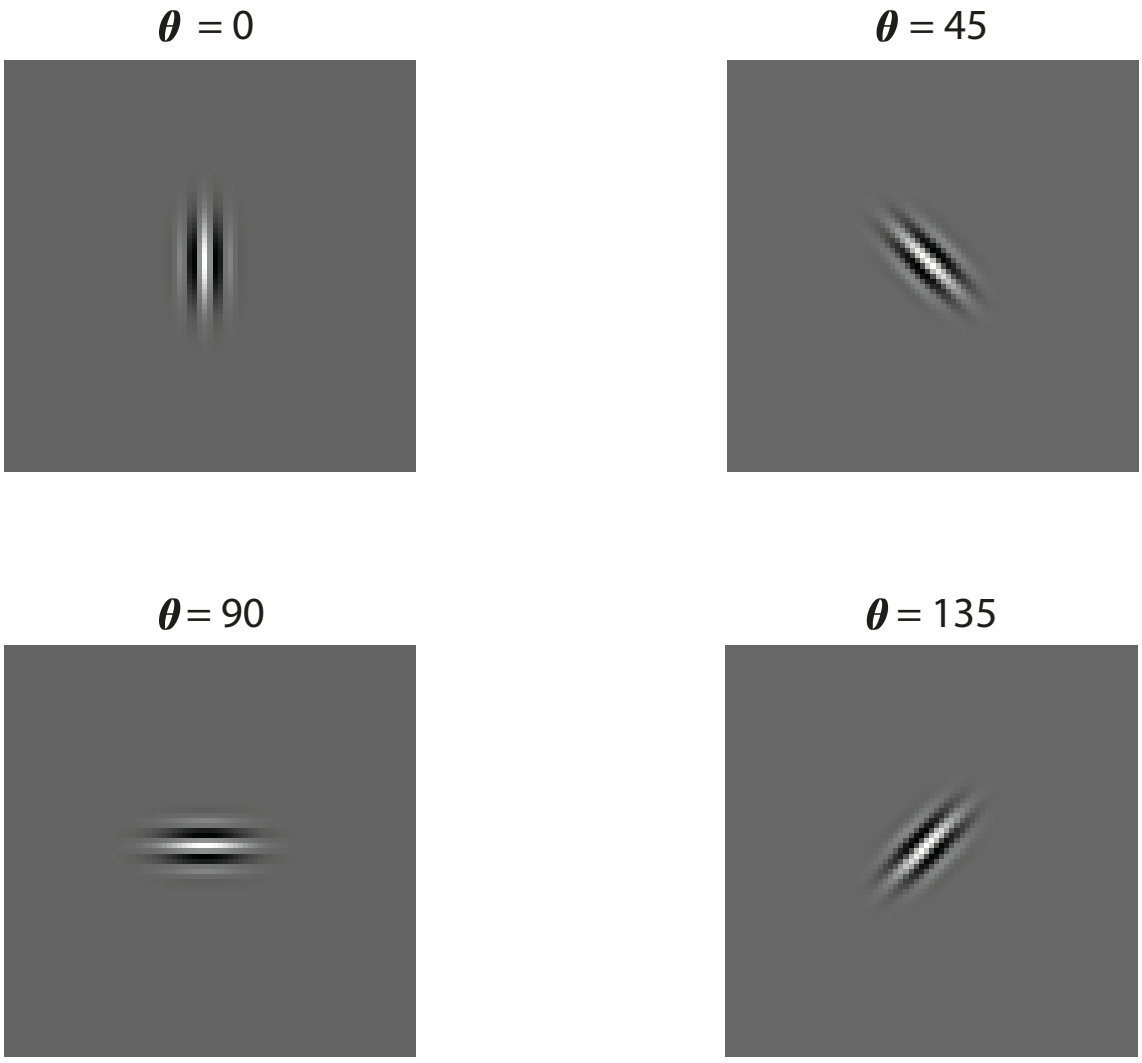


Figure 3.3: Gabor filters with different combinations of orientation.

$$F(x,y) = \exp\left(-\frac{(x_0^2 + \gamma^2 y_0^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} x_0\right) \tag{3.1}$$

where

$$\begin{aligned} x_0 &= x\cos(\phi) + y\sin(\phi), \\ y_0 &= -x\sin(\phi) + y\cos(\phi). \end{aligned}$$

In equation (3.1), ϕ is the orientation of the stripes in a Gabor function, λ is the wavelength of the sinusoidal factor, γ is the spatial aspect ratio and σ is the standard deviation of the Gaussian envelope. Figure 3.3 shows examples of Gobor filters with different combination of orientations.

3.2.2 The HMAX Model Architecture

The feed-forward construct of the HMAX model can simulate the function of the early stages of the visual cortex in recognising objects [2, 113]. In each stage of the HMAX model, two distinct groups of cortical cells are modelled [3]:

1. Simple cells S, to achieve selectivity;

2. Complex cells C, to offer invariance.

Therefore, the original HMAX model (Figure3.4A) comprises two stages. A set of Gabor filters [122] forms the first stage and the second is a template matching mechanism. Each stage of the HMAX model has two sub-stages containing simple and complex cells, namely Simple 1 (S_1), Complex 1 (C_1), Simple 2 (S_2) and Complex 2 (C_2) [3].

The HMAX model uses the classic scheme of convolution/pooling as reported in [3]. The convolutional layers generate selective feature maps and the pooling layers provide invariance.

The input image is first filtered with the above Gabor filters. This results in S_1 feature maps on which the MAX pooling operation is applied (Figure 3.4B). MAX pooling is performed according to scale and orientation to achieve the sub-sampled layer C_1 feature maps. The complex C_1 units obtain the maxima of neighbouring square patches $\mathbf{u}_{i,j}$ of S_1 feature maps as follows:

$$\mathbf{C_1}(i,j)_{response} = \max \mathbf{u}_{i,j}. \tag{3.2}$$

The S_2 layer behaves as a redial basis function unit (RBF). To build the S_2 layer, a set of prototype random patches is extracted from the C_1 layer. All patches from the C_1 layer are then compared with these prototypes using a radial basis function or a Euclidean distance metric. The response of the comparison is inversely proportional to the distance. The response of S_2 units rely on the Euclidean distance between the stored prototypes and the new input image in a Gaussian fashion. That is, for an image patch $\mathbf{Y}$ within the C_1 layer, the distance r of the corresponding S_2 unit is given by:

$$r = \exp(-\beta \left\| \mathbf{Y} - \mathbf{P_i} \right\|^2), \tag{3.3}$$

where β defines the sharpness of the metric and $\mathbf{P_i}$ is one of the N features at the

Table 3.1: The selected parameters for the S_1 and C_1 layers of the HMAX model.

S_1 layer			C_1 layer		
Filter size s	Gabor σ	Gabor λ	Scale band S	Spatial pooling grid ($N_S \times N_S$)	Overlap Δ_S
7×7	2.8	3.5	Band 1	8×8	4
9×9	3.6	4.6			
11×11	4.5	5.6	Band 2	10×10	5
13×13	5.4	6.8			
15×15	6.3	7.9	Band 3	12×12	6
17×17	7.3	9.1			
19×19	8.2	10.3	Band 4	14×14	7
21×21	9.2	11.5			
23×23	10.2	12.7	Band 5	16×16	8
25×25	11.3	14.1			
27×27	12.3	15.4	Band 6	18×18	9
29×29	13.4	16.8			
31×31	14.6	18.2	Band 7	20×20	10
33×33	15.8	19.7			
35×35	17.0	21.2	Band 8	22×22	11
37×37	18.2	22.8			

centre of the RBF units. For each of the eight scale bands and across all positions, the S_2 maps are calculated based on the above.

Finally, the C_2 layer is generated by MAX pooling of S_2 to obtain position- and scale-invariant feature maps for classification. For more details, the reader is referred to [2, 3].

The parameters that control the pooling operation were experimentally adjusted to achieve matching between the units of S_1 and the units of C_1 (see Table3.1) [2].

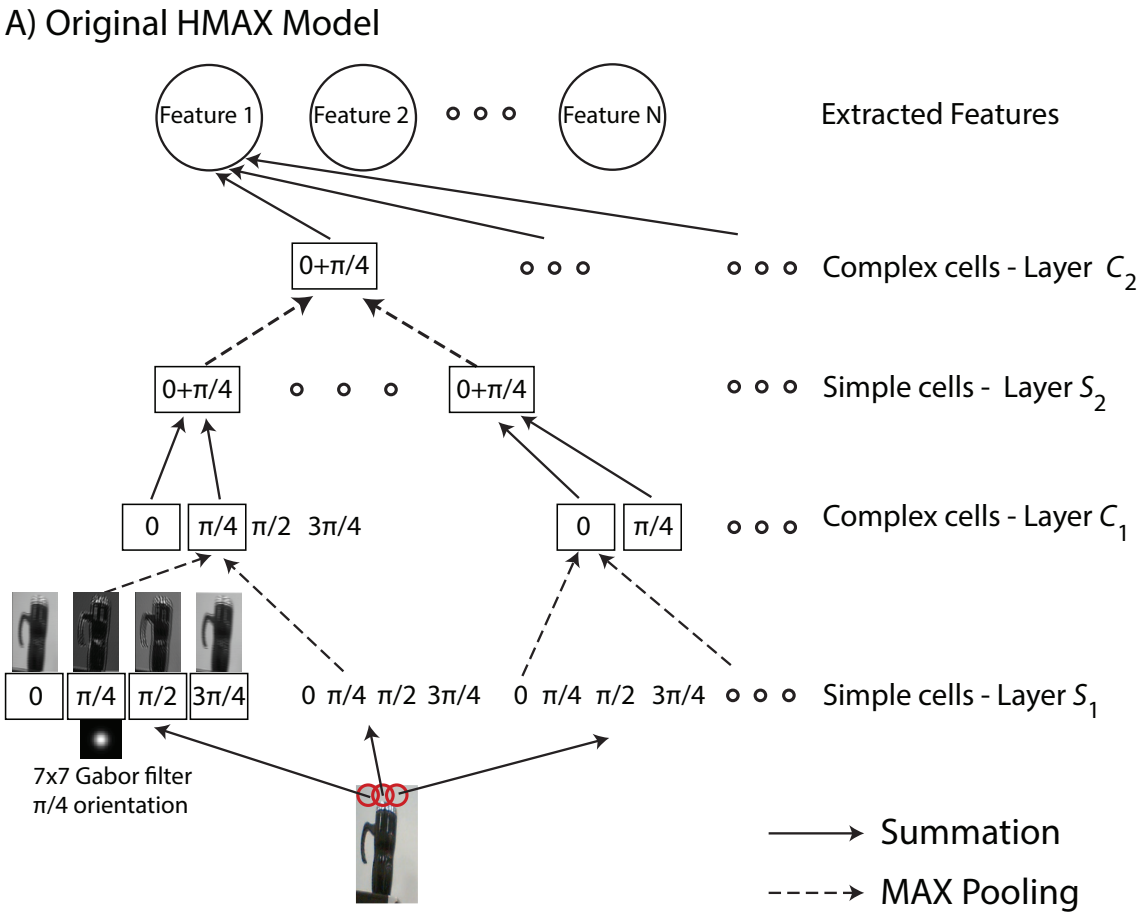


Figure 3.4: A) Schematic of the HMAX model. The basic model consists of a hierarchy of two stages each having S and C layers with S_1 simple-cell like response properties to the C_2 layer with shape tuning and invariance properties [3]. B) MAX pooling operation over non-overlapping windows.

3.3 The Proposed En-HMAX Model

In this section, the structure of the elastic-net regularised version of the HMAX model which is called the Elastic-net HMAX (En-HMAX) is reviewed. The En-HMAX model comprises three layers, each consisting of both simple S and complex C units. Instead of a Gabor filter in S_1, the independent component analysis (ICA)

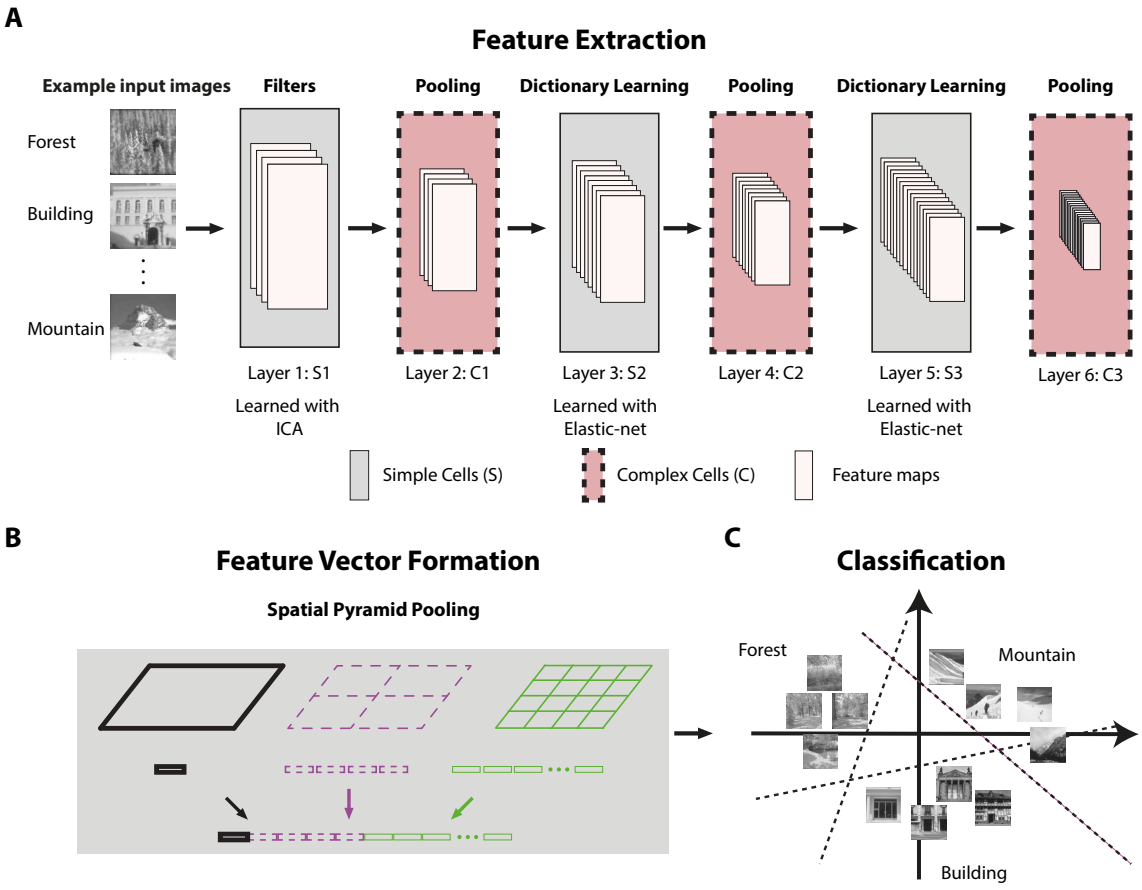


Figure 3.5: A) Schematic of the En-HMAX model with each block representing an S or C layer of the model along with their function. B) Spatial pyramid pooling layer with a grid resolution of $\{1, 2, 4\}$. C) The classification layer. Images shown in the figure are extracted from a scene category dataset [111, 112].

is used to generate filters that resemble the receptive fields of V1 simple cells in the visual cortex [123–125]. Extracting filters from natural images using ICA is believed to better model V1 receptive fields of the visual cortex. The S_2 and S_3 units of the En-HMAX model utilises an elastic-net regularised dictionary learning [67] to reinforce model sparsity and grouping effect, simultaneously.

The proposed En-HMAX model (Figure 3.5A) differs from the original HMAX model in the following aspects:

3.3.1 Number of Stages

The original HMAX model has only two stages, each comprising a simple and a complex layer, as shown in Figure 3.4A. However, Serre et al. [2], among others, showed that an HMAX model with 3 stages is more appropriate to model rapid categorization. Therefore, the designed En-HMAX model was designed with three stages. Nevertheless for completeness, both 2- and 3-stage En-HMAX models were

compared with the original 2-stage HMAX model. Adding a third stage can help the model to learn abstract features from different levels of the hierarchy. Generally, the first layer of the model extracts basic features, such as edges and lines. The second layer recognises advanced features such as shapes from a collection of the edges in the first layer. Therefore, adding a third layer to the classical HMAX model can enhance the model generalisation.

3.3.2 Elastic-Net Regularisation for The HMAX Model

Hu et al. [4] proposed the use of sparse coding in the HMAX model to better represent the visual cortex. They adopted independent component analysis (ICA) [123] in the first simple layer of HMAX (S_1) followed by an ℓ_1-regularised dictionary learning structure in the following S layers. Here, the same approach was followed using the ICA method in S_1 layer. Inspired by [66], the dictionary learning approach in S_2 and S_3 was augmented by using both ℓ_1 and ℓ_2 norms of the sparse coefficients matrix as penalizing terms.

Let $\mathbf{X} \in \mathbb{R}^{m\times n}$ contain m-dimensional image patches $\mathbf{x}$ in the S_2 or S_3 layers of the En-HMAX model, $\mathbf{D} \in \mathbb{R}^{m\times p}$ be a dictionary comprising p bases $\mathbf{d}$, and $\mathbf{S} \in \mathbb{R}^{p\times n}$ include n sparse vectors $\mathbf{s}$ in its columns. Then, in the matrix notation, sparse coding is formulated as $\mathbf{X} = \mathbf{DS}$. To learn the dictionary $\mathbf{D}$ and the sparse weighting matrix $\mathbf{S}$, elastic-net regularisation was used as the following

$$\begin{aligned} \underset{\mathbf{D},\mathbf{S}}{\text{minimize}} \quad & \|\mathbf{X}-\mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 + \lambda_2 \|\mathbf{S}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{d}_i\|_2 \le 1, \quad i = 1, \cdots, p, \end{aligned} \tag{3.4}$$

where $\|.\|_F$ denotes the Frobenius norm and $\lambda_1, \lambda_2 \in \mathbb{R}_{\ge 0}$ are the regularisation coefficients that regulate the trade-off between sparsity and the sensitivity of basis selection. When $\lambda_1 = 1$ and $\lambda_2 = 0$, equation (3.4) reduces to the ℓ_1 coding method described in [4, 69], hereafter called the LASSO-HMAX model and when $\lambda_1 = 0$ and $\lambda_2 = 1$, equation(3.4) reduces to another extreme case, which in this thesis referred as the Ridge-HMAX model. The notions of LASSO and Ridge regressions are borrowed from [69].

The En-HAMX structure is followed by a feature formation layer in which the spatial pyramid pooling (SPP) technique is adopted [126]. With the SPP method, with a grid resolution of {1, 2, 4}, each feature map in C_3 is transformed into

Table 3.2: Parameters of the proposed model

Model parameters	Stage 1	Stage 2	Stage 3
Sparse coding	ICA	Elastic net	Elastic net
No. of bases	8	256	1024
Patch size	8×8	$4 \times 4 \times 8$	$2 \times 2 \times 256$
Sample size	25×10^4	25×10^4	25×10^4
regularisation	-	$\lambda_1 = 0.15$, $\lambda_2 = 0.15$	$\lambda_1 = 0.15$, $\lambda_2 = 0.15$
Pooling method	$(\sum_{r=1}^{n} \lvert q_r \rvert)^{\frac{1}{2}}$	$(\sum_{r=1}^{n} \lvert q_r \rvert)^{\frac{1}{2}}$	Max spatial pyramid
Pooling size	2×2	1×1	$\{1, 2, 4\}$

a feature vector of length 21. Figure 3.5B illustrates the structure of the SPP operation. Finally, as illustrated in Figure 3.5C, we used a linear multi-class classifier to group the input images. Detailed information on classification and cross validation are presented in the following.

3.3.3 Pooling Method

The C_1 and C_2 complex layers were partitioned into small non-overlapping square patches, termed $\mathbf{q}$ in a vector form. The ℓ_1 pooling was then applied such that from each patch the ℓ_1-norm, that is $(\sum_{r=1}^{n} \lvert q_r \rvert)^{1/2}$ was extracted. In addition, for C_3 the spatial pyramid [126] pooling method was used.

A full description of the parameters of the proposed En-HMAX model is presented in Table 3.2. The same parameters and settings were used in both training and testing stages in all En-HMAX, Ridge-HMAX and LASSO-HMAX model setups.

3.4 Software Implementation

All models were implemented in Matlab on a dual-core i5 processor (3.4 GHz) PC with 32G RAM without GPU acceleration. The average time for one single-threaded operate within the standard architecture of the En-HMAX model was about 90 minutes. The timing calculations include parameter initialization, training and testing of the Caltech-101 image dataset. The dictionary learning stage of the proposed En-HMAX model remains a challenge to run on-line, where the filters are updated continuously depending on the type of the environment. The PC used for these experiments was a dual-core i5 processor (3.4 GHz) with 16 G RAM and all timings were calculated on a single thread.

3.5 Image Database

3.5.1 Object Dataset

Seven image classes from the Caltech 101 dataset [110] were selected. These classes were: bass (54 images), binoculars (33 images), brontosaurus (50 images), camera (50 images), chair (62 images), gerenuk (34 images, also known as Waller's gazelle) and grand piano (99 images). Figure 3.6 shows two examples in four classes of the dataset to reflect the richness of this dataset in terms of object size, orientation, position and background. The rationale for choosing these classes was that an ample number of images per class was available, which allowed tuning the model parameters effectively; whilst keeping the computations to a reasonable level. Some of the images of the Caltech 101 dataset were collected from Google Image search to include the highest number of images for each category. Minimum preprocessing was introduced to the images of the dataset. Some images in the dataset were manually flipped such that objects of the images face the same direction. Lastly, the images of the dataset were scaled roughly to approximately 300 pixels wide.

3.5.2 Scene Dataset

After the success of the eighth-scene dataset [127], emerged a need for a larger and more complete dataset. Researchers intended to enlarge the number of classes and the degree of complexity of the above dataset. To meet this need, fifteen scene

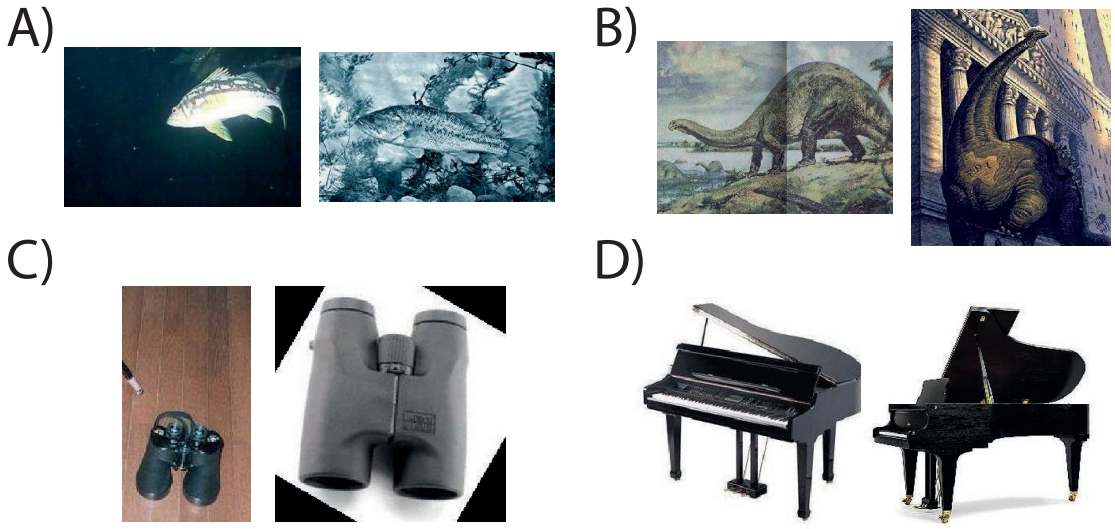


Figure 3.6: Example of 4 (of 7) image classes, A) bass, B) brontosaurus, C) binoculars and D) grand piano that were used in analysis. Samples illustrate the range of image sizes, orientations (portrait and landscape) and backgrounds [110].

category dataset was introduced. The fifteen scene dataset inherited the same classes of the eighth-scene dataset, including two more outdoor classes and five more indoor classes. As such, forming an integrated platform for scene understanding tasks, a mandatory benchmark in recent research. It is considered as one of most the complete scene category datasets [111].

The scenes dataset included man-made as well as natural scenes. Scene images were extracted from a scene categories dataset that was collated by Li and Perona [112] and augmented by Lazebnik et al. [111]. The images of the scene dataset have different dimensions but on average are of 300 $\times$ 250 pixels. All images of the fifteen scene dataset are converted to grey scale. The classes in the Scenes dataset are: bedroom (216 images), suburb (241 images), industrial (311 images), kitchen (210 images), living room (289 images), coast (360 images), forest (328 images), highway (260 images), inside city (308 images), mountain (374 images), open country (410 images), street (292 images), tall building (356 images), office (215 images) and store (315 images). Figure 3.7 shows three examples from each category.

3.6 Classification

Two classification scenarios were conducted of the object dataset: 15 or 30 images were selected randomly from each class to train the classifier. The remaining samples in each class were used for testing the classifier. However, for the scene dataset

Figure 3.7: Example images from the scene category database [111].

experiment, 100 images per class were used for training. The number of test images in each class was different, therefore to avoid bias, classification scores were averaged across all categories. Additionally, to ensure that the classification scores were not biased by the random choice of training samples, the classification was repeated for 20 independent runs in each condition (15 and 30 training samples). The average classification scores were reported together with the standard deviations. A

Table 3.3: The average sparsity achieved with different models

LASSO-HMAX		**En-HMAX**		**Ridge-HMAX**	
C_2	C_3	C_2	C_3	C_2	C_3
0.004	0.001	0.354	0.102	0.427	0.112

multi-class linear support vector machine (SVM) [70, 72] implemented within the LIBLINEAR library [72] was selected as the classifier due to its computational simplicity. The size of the output feature of the En-HMAX model is 21504 for each image. These features are fed into the SVM classifier for classification purposes.

3.7 Statistical Analysis

To test the statistical significance of using the En-HMAX model in improving the classification performance, a $3 \times 2 \times 2$ analysis of variance (ANOVA) was performed with repeated measures. The main factors were the choice of model (LASSO-, En- and Ridge-HMAX), whether classification was carried out at C_2 or C_3 layers, and finally the number of training samples, 15 versus 30. Following the main analysis, post-hoc comparisons were performed. Multiple comparisons were adjusted using Bonferroni correction. Furthermore, the F1-scores was calculated and reported to measure the En-HMAX model test's accuracy. The F1-score computes both the recall and the precision of the test data. It then obtains the average of the recall and the precision, where the maximum value that the F1-score can reach is 1.

3.7.1 Quantifying Sparsity

It was suggested that using two penalty terms in (3.4), ℓ_1 and ℓ_2-norms of $\mathbf{S}$, would lead to extraction of sparse C_2- and C_3-layer feature maps, which can retain second, and potentially higher, order correlation features. To support this, representative examples of C_2- and C_3-layer feature maps were provided. Feature maps were calculated with the En-HMAX and LASSO-HMAX ($\lambda_2 = 0$) model settings in Figure 3.8. In this figure, the responses of the C_2- and C_3-layers, calculated with the En-HMAX model, have several areas with class-specific strong activations that resemble the original image, e.g. the neck of the brontosaurus. The feature maps extracted

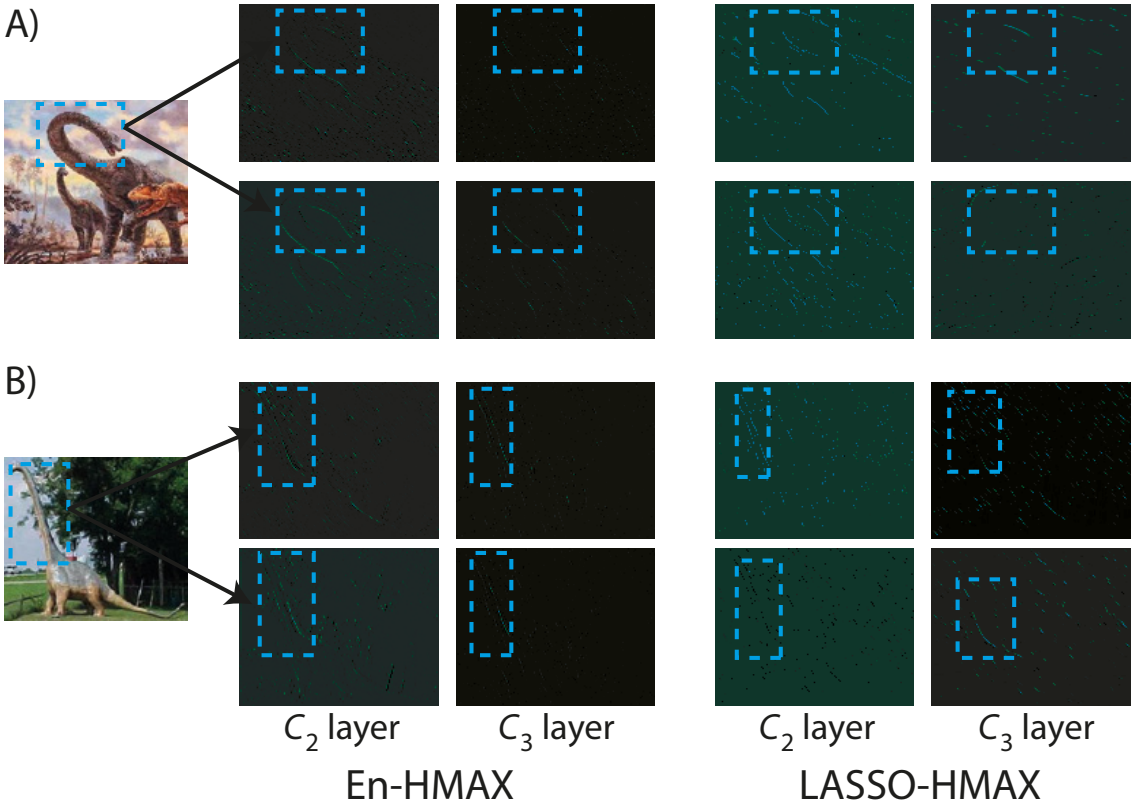


Figure 3.8: Higher order correlation in representative feature maps extracted by using the En-HMAX model from the two example images A and B. Feature maps obtained by the En-HMAX model extract the neck of the brontosaurus very clearly. On the other hand, feature maps calculated with the LASSO-HMAX model are too sparse to reveal any determining feature of these image classes. Feature maps are gray scale. For visualization only, color scaling was used and feature maps were enlarged to counterbalance size shrinkage due to norm-pooling. Images are taken from Caltech 101 dataset [110].

by the LASSO-HMAX model are, however, too sparse and although they can correspond to some of the important features of the input images, many of the other important details are missed.

Table 3.3 reports the average sparsity achieved when all images of all classes were introduced to the En-, LASSO- and Ridge-HMAX models. As predicted, using the En-HMAX model led to sparsity levels that fall between those achieved with the LASSO- and Ridge-HMAX models in both C_2 and C_3 layers.

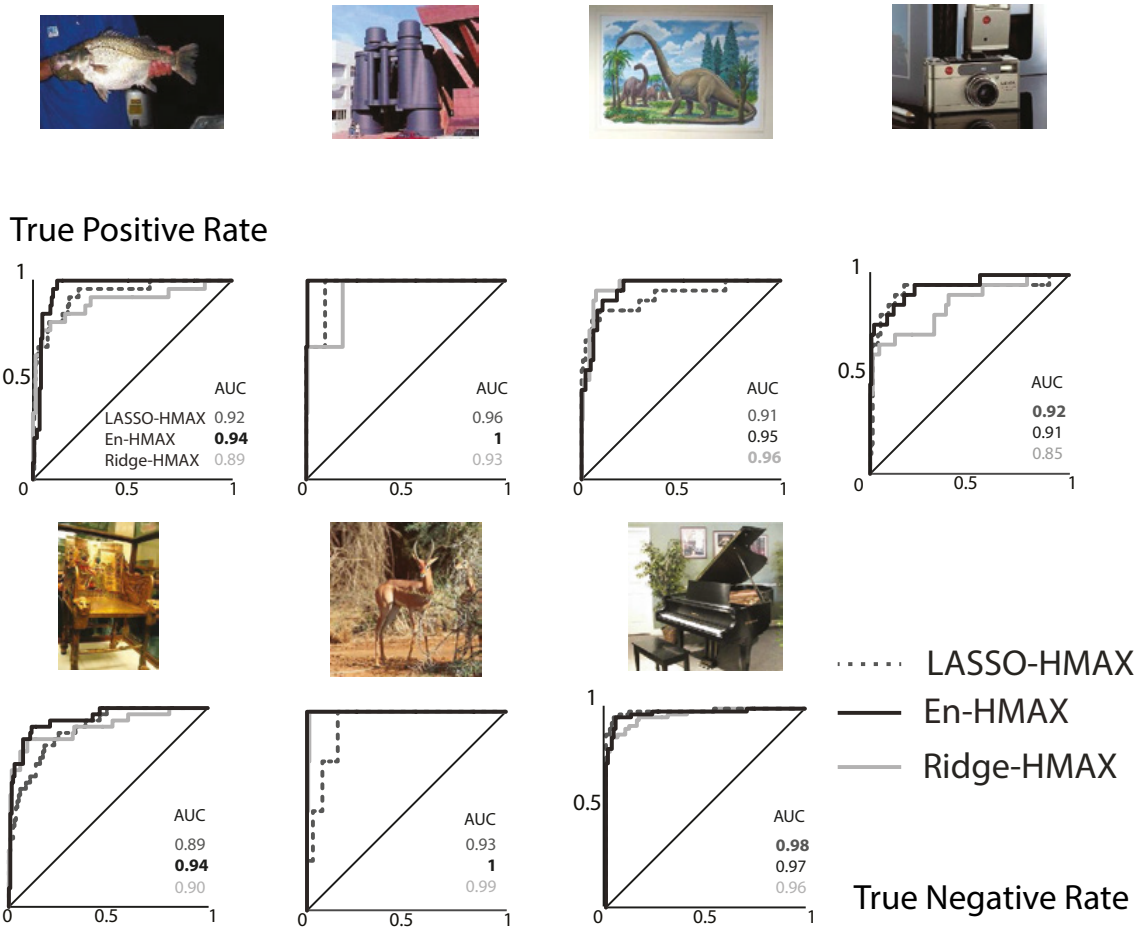


Figure 3.9: Performance comparison of the En-, LASSO- and Ridge-HMAX models with respect to the ROC and AUC measures; Top: Samples from images classes with different sizes and orientations; Bottom: The corresponding ROCs curves and the calculated AUC values for each image class. The highest AUC value is in a bold font. The vertical and horizontal axes denote the true positive and false positive rates, respectively.

3.8 Results

3.8.1 Object Classification Scores

A comparison was made between the En-, LASSO- and Ridge-HMAX models in terms of classification accuracy. For completeness, the classification scores achieved by the original 2-layer HMAX model [3] were included. Also, the results of the recent models of deep learning algorithms are reported using pre-trained neural network. Table 3.4 reports the classification results. Statistical analysis revealed the main effect of the model ($F_{2,18} = 266.59, p < 10^{-5}$), feature map selection ($F_{1,19} = 24.37, p < 10^{-5}$) and number of training data ($F_{1,19} = 115.83, p < 10^{-5}$). In both 2- and 3-layer structures and in both 15 and 30 training sample conditions, the En-HMAX model outperformed all other algorithms ($p < 10^{-5}$). The performance

Table 3.4: Average classification accuracy ± standard deviation (SD).

HMAX model configuration	2-layer Arrangement No. of training images		3-layer Arrangement No. of training images	
	15	30	15	30
HMAX [3]	35.014±0.09	40.587±0.08	-	-
LASSO-HMAX [4]	69.48±0.03	75.08±0.05	56.55 ±0.02	63.93±0.05
En-HMAX	75.14±0.02	80.37±0.04	78.71 ±0.01	82.72±0.04
Ridge-HMAX	66.14±0.02	71.45±0.05	67.27 ±0.02	73.30±0.06
Deep Learning Methods			**Higher Layers**	
AlexNet [47]	-	-	90.62±0.01	97.68±0.02
VGG19 [46]	-	-	95.37±0.01	97.87±0.01
GoogLeNet [48]	-	-	96.65±0.01	98.99±0.01

improvement in the 3-layer arrangement was considerably larger than that in the 2-layer setup ($p < 10^{-5}$). This is particularly interesting because in the experimental neuroscience literature, a 3-layer HMAX model setup is deemed more appropriate for modelling visual processing [2]. Finally, using 30 training images, instead of 15, improved classification scores significantly ($p < 10^{-5}$).

Theoretical analysis indicated that all forms of ℓ_p norm pooling can offer invariance [128]. However, in practice, different pooling mechanisms could lead to stark differences in recognition performance. It was found that the use of ℓ_1-norm pooling in the C_1 and C_2 layers offers much better performance than MAX (ℓ_∞-norm) pooling. The overall performance achieved by the use of ℓ_1- and ℓ_2-norm pooling in C_1 and C_2 were comparable.

Figure 3.9 shows the receiver operating characteristic (ROC) curve [129] for all of the classes used in this experiment using a 3-layer En-HMAX model (30 training images). The area under the curve (AUC) was used to characterize the classification confidence in a specific binary classification task (e.g., camera versus not-camera) with a unity value reflecting a 100% accuracy. In 4 out of 7 classes, using the En-HMAX model led to the highest AUC. The performance of the En-HMAX model was only marginally lower than the LASSO-HMAX model in 2 classes and the Ridge-HMAX model in 1 class. Table 3.5 reports the F1-scores [130], and the corresponding precisions and recalls, achieved with different models for a 3-layer

Table 3.5: F1-scores for 3-layer Arrangement with 30 training images

HMAX model configuration	F1-Score	Precision	Recall
LASSO-HMAX [4]	0.37	0.25	0.66
En-HMAX	0.63	0.51	0.83
Ridge-HMAX	0.49	0.39	0.66

Table 3.6: Classification results for the scene category database

Feature types / recognition model	Classification performance
The En-HMAX model [135]	**76.4 ± 0.5**
BSC [131]	72.5 ± 0.3
Rasiwasia [132]	72.2 ± 0.2
Liu [133]	63.32
Bosch [134]	72.7

En-HMAX model (30 training images). Results reflect the higher performance of the En-HMAX model when compared to the LASSO-HMAX and Ridge-HMAX models. However, it performs less than deeper models of object recognition at the cost of the efficiency and computational complexity.

3.8.2 Scene Classification Scores

Table 3.6 shows the complete results of the classification performance using 100 images per class for training and the rest for testing, the same set-up used in other scene recognition methods [131–134]. Average classification results across 20 independent runs and the standard deviations are reported. The classification rate is 76.4%, which is higher than the best results of 72.5 %, achieved in [131]. Although recognising scene images is considered dramatically different than recognising object images, i.e., locations of features, and characteristics of features, the En-HMAX achieved high scores for scene image classifications. This indicates that the En-HMAX model successful in recognising scene images alongside object images.

3.9 Lateral Connections

Neuroscience studies have shown that there are two types of lateral connection in the primates visual cortex: excitatory and inhibitory, where long-range horizontal connectivity is intrinsic in the primary visual cortex (V1) [52, 136]. Inspired by these studies, lateral connections of the En-HMAX model were investigated. The connections showed in Figure 3.10 comprises all three C layers response. Extracting the features from all layers through the visual hierarchy has allowed investigating the features effectiveness of each layer separately. The importance of utilising different levels of features through the hierarchy was studied. The impact of concatenating low-and high-level features of the En-HMAX model was quantified. A structure similar to that of the En-HMAX model was used, however, using an elastic net regulariser for dictionary learning in all layers. The rationale was to quantify the effectiveness of features in different layers with similar techniques in all layers. In order to smoothly fine tune the parameters, a small number of bases in the S_2 layer and S_3 layer was selected. This helps to rapidly tune the parameters for an enhanced performance. Models formed by the lateral connection comprised 50 S_1 bases of dimensions 10×10, 40 S_2 bases of dimensions $12 \times 12 \times 50$ and 36 S_3 bases of dimensions $13 \times 13 \times 40$.

The bases were learned by elastic net regularisation with 50,000 patches arbitrarily extracted from images or C patches. Regularisation parameter λ_1 ,λ_2 were 0.15. All models were implemented in MATLAB; a softmax classifier [137] was used to perform classification.

Several models were emerged using the lateral connections of the En-HMAX model as the followings:

- model 1 which comprises the features of the first two layers, i.e., C_1 and C_2 responses;
- model 2 and model 3 comprising C_2 and C_3 and C_1 and C_3 features, respectively;
- model 4 and model 5 were formed classically by using features from single layers, C_2 and C_3, respectively;
- model 6 utilises low level, mid-level and high-level features of the hierarchy.

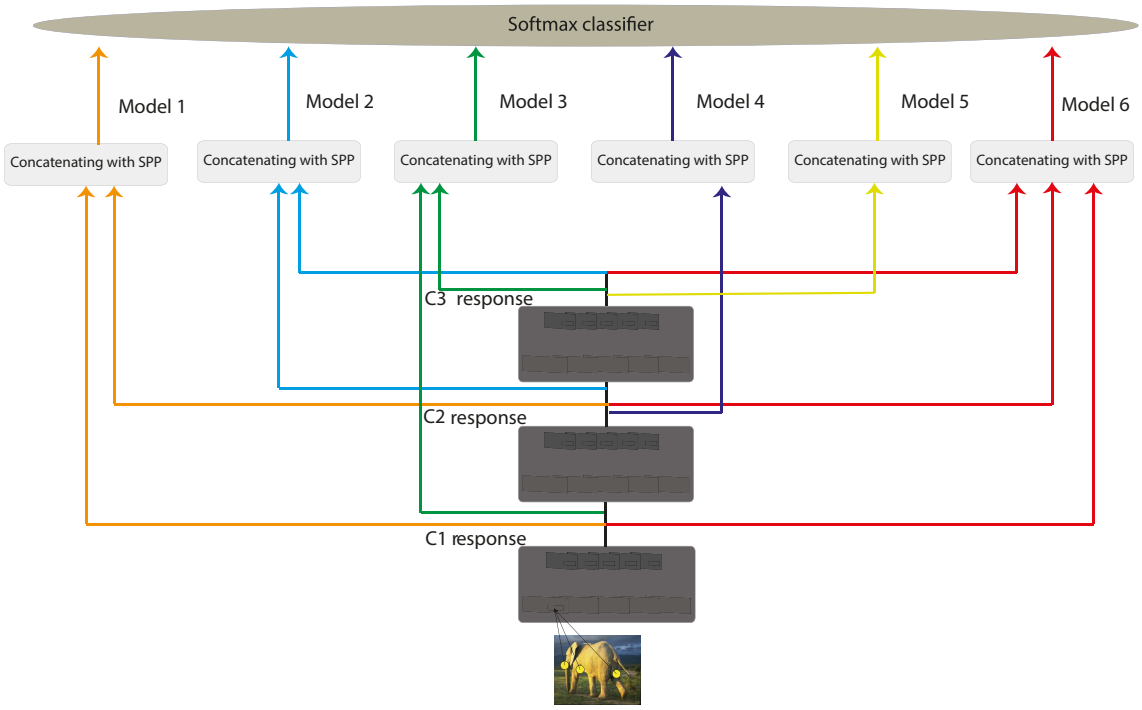


Figure 3.10: The lateral connections combining different layers of the proposed En-HMAX. Different types of features are extracted using all possible combinations of layers output. For instance, model 1 is formed using C_1 layer response and C_2 layer response.

3.9.1 Cross Validation

In the lateral connections study, 12.2% of the image data was used for training, and the remaining image data (87.8%) was used for testing. This training ratio is considered small enough to quantify the model generalisation to the unseen image data. Cross-validation was used to establish a more accurate platform to compare the performance across different models constructed using the lateral connections. Testing the accuracy of different independent images across many iterations produces a valuable approximation of performances. The testing images were regarded as new data on condition that data are independent and identically distributed (i.i.d.), where all the image dataset have similar number of image samples across all the categories. The classification accuracy is measured and averaged across 5 random splits of train and test sampled images.

3.9.2 Chance Level Performance

To evaluate models of the lateral connections, an image dataset was formed using images extracted from Caltech-101 dataset [110] was used. The classes of Caltech

Figure 3.11: The dataset used for the lateral connections study [110].

101 image dataset contain an inconsistent number of images. Therefore, only four of the classes with the higher number of images were selected. Accordingly, the En-HMAX model generalisation capabilities were tested, to new images. These categories (as shown in Figure 3.11 are bonsai, faces, airplanes and car-sides. For the faces class of images, only eight different individuals were selected. The training and testing of the model were done on those same eight individuals within the faces category.

The performance was interpreted as how much the classification outcomes diverge from the rate accomplished by a random classification, for instance, in a two-class and a four-class classification scenarios, the chance levels are 50% and 25%, respectively. On such conditions, the training image data is expected to be equally distributed among all classes. For this particular study, the classification performance was conducted over four object image classes. Each class consists of 15 images and 108 images for training and testing, respectively. As a result, the chance level of this study is 25%

3.9.3 Scores of The Lateral Connection Experiment

Table 3.7 summarises the results of the models extracted from the full lateral connection system. Interestingly, the features extracted using model 1 achieved the highest performances. Model 6 that was formed using a combination of all the available feature layers also achieved high performances, however, slightly less than model 1. This indicates that using elastic-net regulariser in three consecutive lay-

Table 3.7: Mean classification accuracy in percentage ∓ standard deviation (SD).

Model Architecture	**Training Size**	
	15	**30**
Model 1	82.6387 ∓3.7183	82.8496 ∓4.0386
Model 2	67.1295 ∓5.2277	74.1398 ∓10.0892
Model 3	76.7360 ∓2.6495	79.8386 ∓5.3964
Model 4	69.8372 ∓9.1673	74.0322 ∓9.9380
Model 5	62.2040 ∓3.8459	70.4838 ∓8.0494
Model 6	79.3402 ∓1.9766	82.3656 ∓4.0437

ers may sparsify some of the important features of the S_3 layer. Therefore, in the standard En-HMAX model, an ICA method was used in the first layer of the model followed by two layers of the elastic net regulariser for dictionary learning.

Model 1 and model 4 share the same architecture of the original HMAX model. However, model 1 shows better performances in both of the used training sizes. The advantage of model 1 over the other models can be explained by the ample number of bases selected in the lateral connection experiment. The standard En-HMAX model comprises 1024 bases in the S_3 layer, while the lateral connections system comprises only 36 bases in the S_3 layer.

Figure 3.12 displays the results using 15 and 30 images for training samples per category, respectively. The classification accuracy of each model is represented by a box plot. The wide spread of in the box plots of model 2 and model 4 shows that these models were not as stable as the other models.

Figure 3.13 shows the classification accuracy of the individual categories when using a training size of 15 and 30 images. Car-side category scores the highest

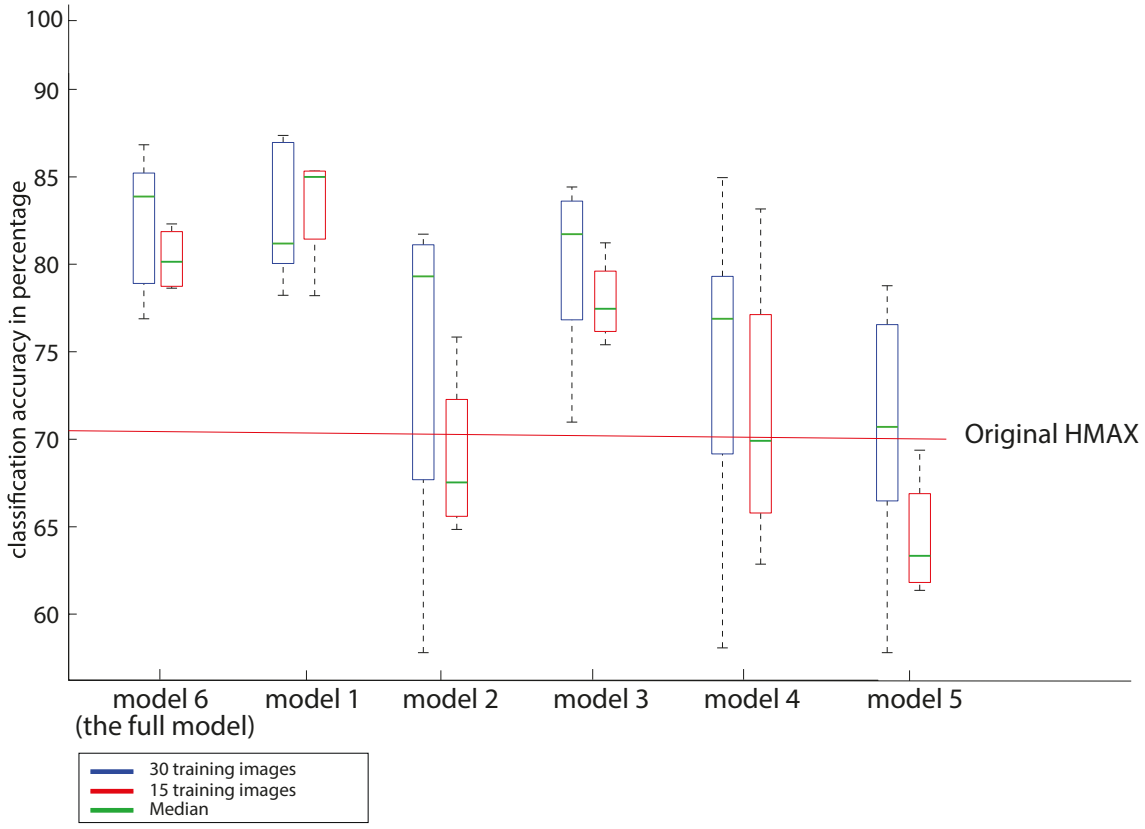


Figure 3.12: The performance of the proposed models using both 15 and 30 training images. From the plot, the following numeric values can be observed: the median (in green), 25% quantile (lower edge of the blue box), 75 % quantile (upper edge of the blue box), minimum value (lower black terminal) and maximum value (upper black terminal). The average classification accuracy of the original HMAX model is represented by the horizontal red line.

accuracies while planes categories attain the lowest even when the chance level used in all experiments is even. This indicates that the success of recognising a class of images is related to the class identity. Many factors within each class of images determine the class recognition difficulty, for instance, the variation in object pose, size, background and location in the image. These factors dramatically contribute to the discrepancies between the different dataset in solving object recognition tasks.

3.10 Visualization of Higher-level Features

The En-HMAX model, among other object recognition models, extracts discriminant features of the objects in the images. Figure 3.14(a and b) shows an example of visualising the strongest features of an object. Figure 3.14(c) shows examples of patterns that represent strong activations in the used dataset. The bases were

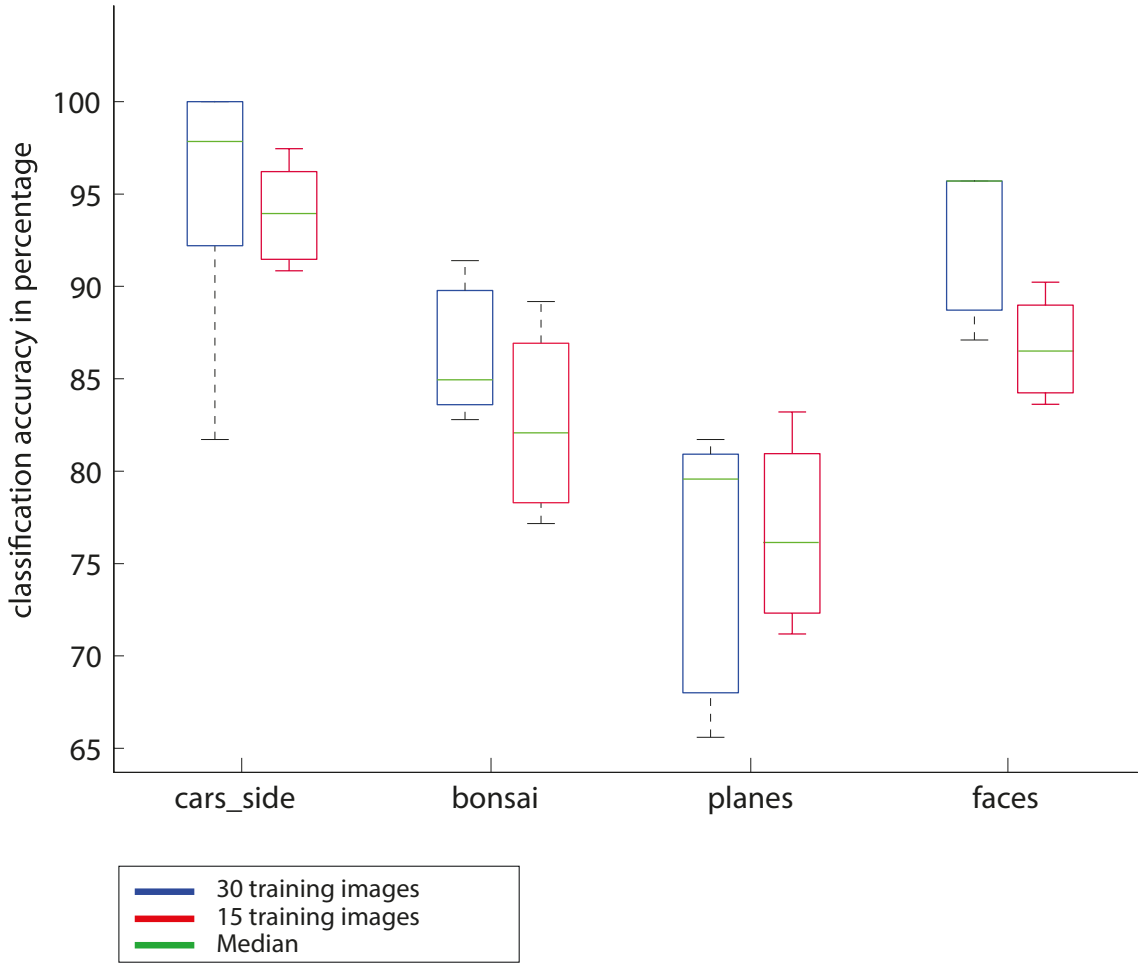
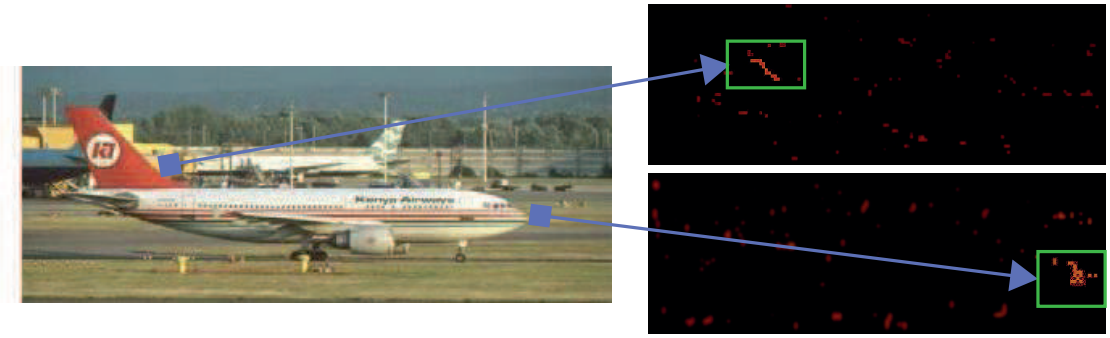


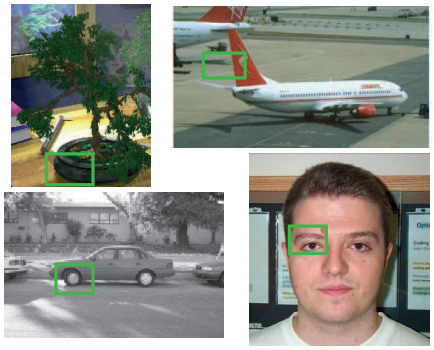
Figure 3.13: The classification accuracy of the individual categories of model 6 using a training size fo 15 images and 30 images. From the plot, the following numeric values can be observed: the median (in green), 25% quantile (lower edge of the blue box), 75 % quantile (upper edge of the blue box), minimum value (lower black terminal) and maximum value (upper black terminal).

activated by a particular syntactic subject. For example, the 22nd feature map, as shown in Figure 3.14(b) bottom, is highly activated by a ">" shape; the 32nd feature map, as shown in Figure 3.14(b) top, is highly activated by a "\" shape.

Similarly to [4], in order to visualise the bases of higher layers of HMAX model, the bases were projected on the dataset. Bases from higher layers were combined with that of the lower layers. Due to the shrinkage resulted from MAX pooling, the patches of the higher layers were resized to counterbalance the sub-sampling. As a result, the bases of higher layers were projected into the input domain. Figure 3.15 shows the visualised receptive fields of S_1, S_2 and S_3. The given examples demonstrate the bases of the following categories: car-side, faces-easy, bonsai and airplanes. S_1 bases resemble simple lines and edges, similar to that of Gabor filter. The S_2 bases, however, resemble basic parts of the objects, such as complex corners.



(a) image (b) feature maps



(c) stongest activations

Figure 3.14: visualization of feature maps. (a) Input image from Caltech 101 data set. (b) Some of the feature maps of the input image. The arrows specify the highest responses and their equivalent locations in the images. (c) Some of the Caltech 101 images that have the strongest responses. The green squares mark the receptive fields of the highest response.

However, some S_3 bases resemble more complex parts of objects, such as noses of faces and leafs of the bonsai.

3.11 Comparison With The Original HMAX Model

The En-HMAX was designed to improve the performance of the original HMAX model. It was developed to preserve the same mechanism and structure of the original model, however, with advanced computational methods that achieve similar objectives. The designed En-HMAX differs from the previously developed HMAX model [3] in different aspects. It can be summarised as the followings:

1. The filters of the En-HMAX model were learned from natural images. In contrast to the original HMAX, in which a hand-crafted Gabor function with different scales and orientations were applied to the input images.

2. Similarly to the original HMAX, the feature maps of the En-HMAX were

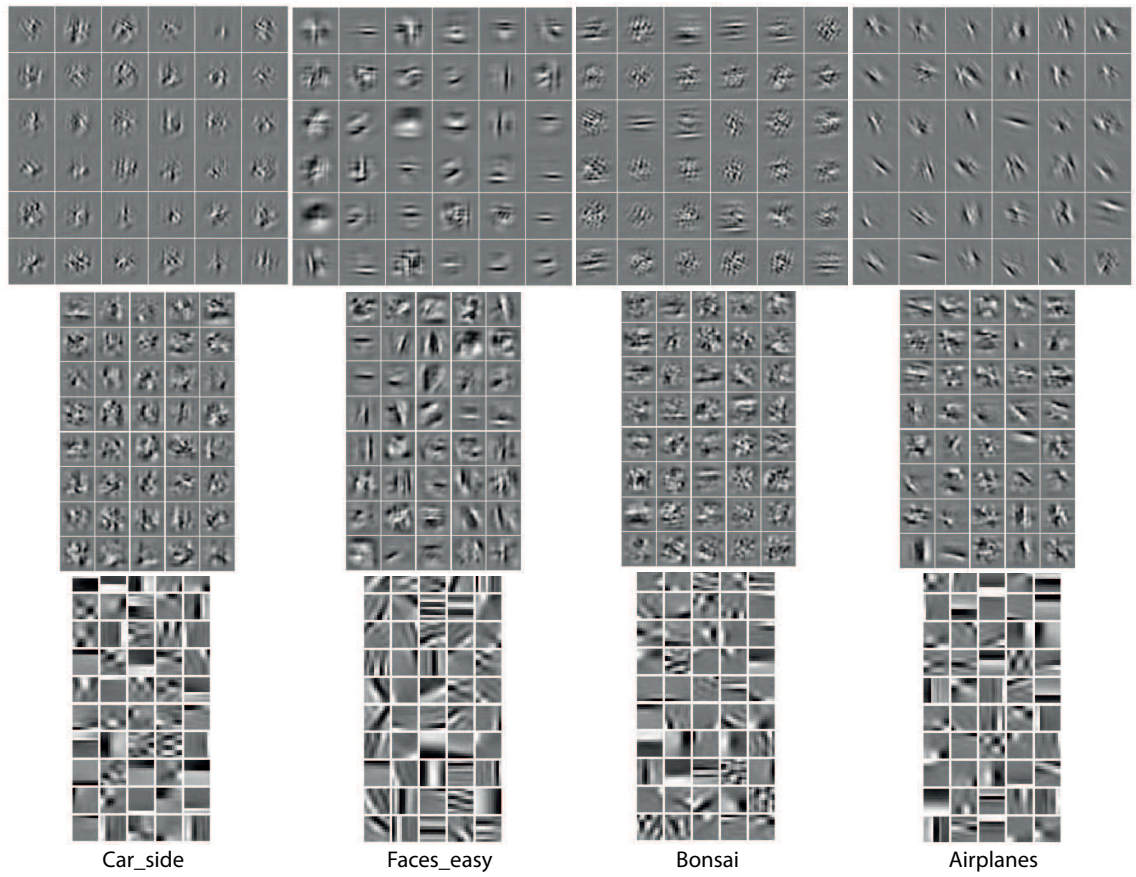


Figure 3.15: visualization of S_1 bases (bottom), S_2 bases (middle) and S_3 bases (top) learned from Caltech 101 dataset. The categories learned are: car-side, faces-easy, bonsai and air-planes from left to right.

down-sampled using a pooling operation. However, in the En-HMAX, the ℓ_1-norm pooling was adapted as the main pooling method in the C layers. The ℓ_1-norm pooling provides more invariance to transformations and occlusion, as it represents all data points in an image patch.

3. In the En-HMAX, the bases of S_2 and S_3 layers were learned by sparse coding using an elastic net regulariser. In the original HMAX, however, a template matching with a radial basis function was used to produce the S_2 features.

4. The En-HMAX model can deal with input images of different sizes and orientations, i.e., portrait and landscape, using a developed SPP method [138]. Only the distinctive data point from each view point is passed to the proceeding layers using both the pooling layer and the SPP layer. The SPP layer encourages the model invariance to position and scale, especially when it is used in the higher layers of the model. Additionally, the SPP layer offers more flexibility and scalability. The HMAX model, however, can only deal

with images of fixed sizes. Images of different sizes were cropped [139], or warped [80]. The cropping and warping approaches have few limitations, for instance, the cropped area does not necessarily cover the whole object. Additionally, the warped part of the images could result in undesirable geometric distortion. Content loss or distortion could compromise the recognition accuracy. Moreover, a pre-defined scale may not be appropriate when object scale alternates.

3.12 Testing The En-HMAX Model with Occlusions

A growing body of evidence supports the proposition that biological systems are able to recognise an object under partial occlusion [29, 140]. As a result, the En-HMAX was tested for occlusions. This section focuses on the capabilities of the En-HMAX model to recognise object and scene images under partial occlusion, for instance, recognising a coffee mug using only the partially available visual content, as shown in Figure 3.16. The En-HMAX model was inspired by the human visual cortex, a powerful platform to decode occlusion. Therefore, in this section, to tackle occlusions, the En-HMAX was used to perform the task with no extra knowledge, for instance, providing the En-HMAX model with three-dimensional images that may show the inconsistency of the occludee, or training the En-HMAX model with patterns of occlusion to better understand occlusions.

To conduct the experiments in this section, the following types of occlusion were used:

- class-A occlusions: images of object and scene dataset artificially occluded with different percentages of standard image distortions.
- class-B occlusions: images of geometrical shapes disordered with complex patterns of occlusion.

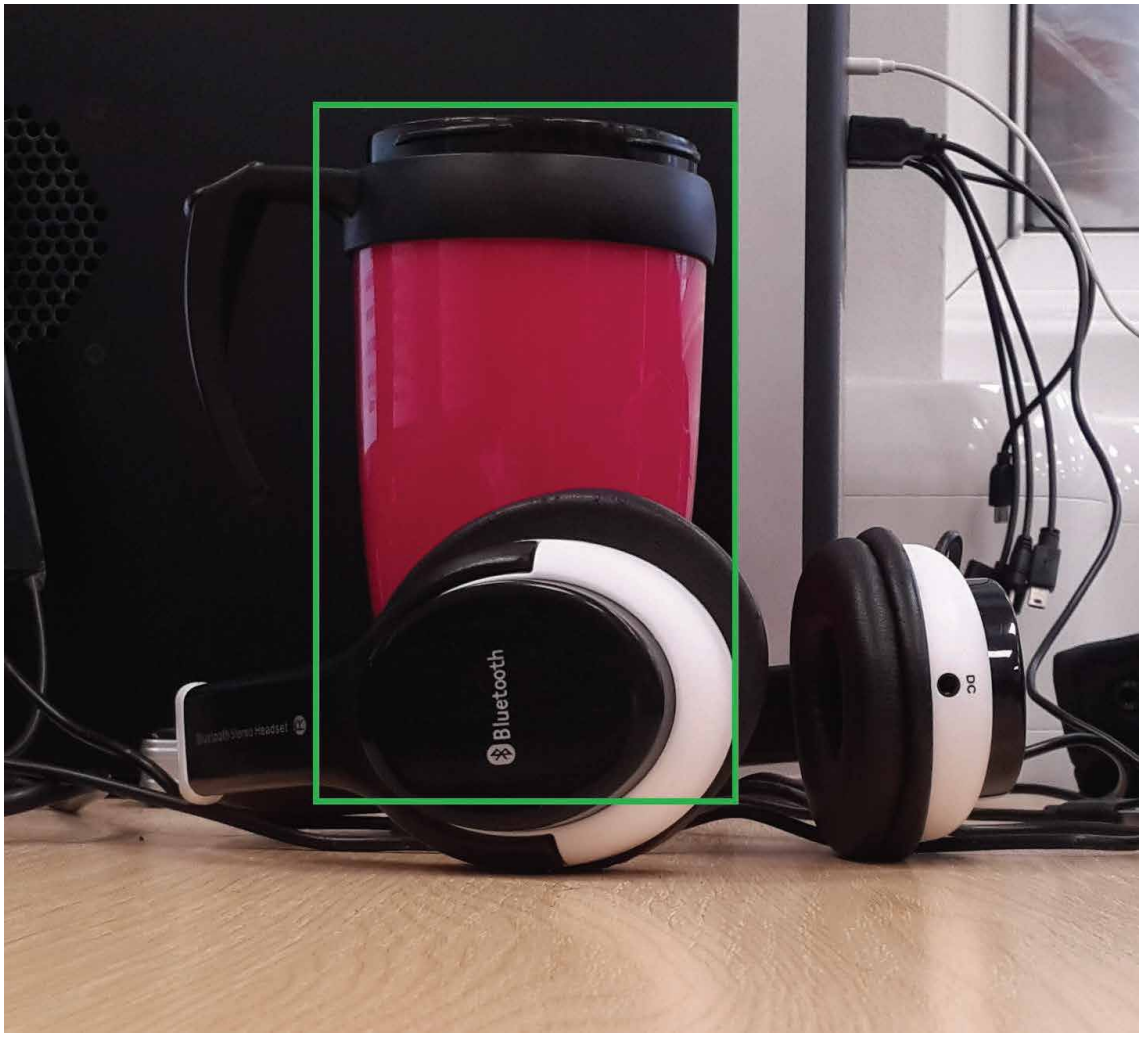


Figure 3.16: Example detection and recognition of a cup under partial occlusion.

3.12.1 Dataset

3.12.1.1 Object and Scene Dataset

To investigate the performance of the En-HMAX model under partial occlusion, image categories from objects and scenes were used. Due to the stark differences in the nature of each of the above datasets, class-A occlusions were applied to the above datasets and the results for each dataset were reported individually.

For the object image dataset, the classes were collated from Caltech 101 dataset [110] and Caltech 256 dataset [6]. Occlusions were then applied to both types of the image dataset. The occlusions were applied to the images with different sizes and shapes, as shown in Figure 3.17.

For the scene image dataset, the classes were collected from Fifteen scene categories [111] dataset. The rationale for using this dataset was to investigate the En-HMAX model robustness against occlusions on scenes. The dataset contains a

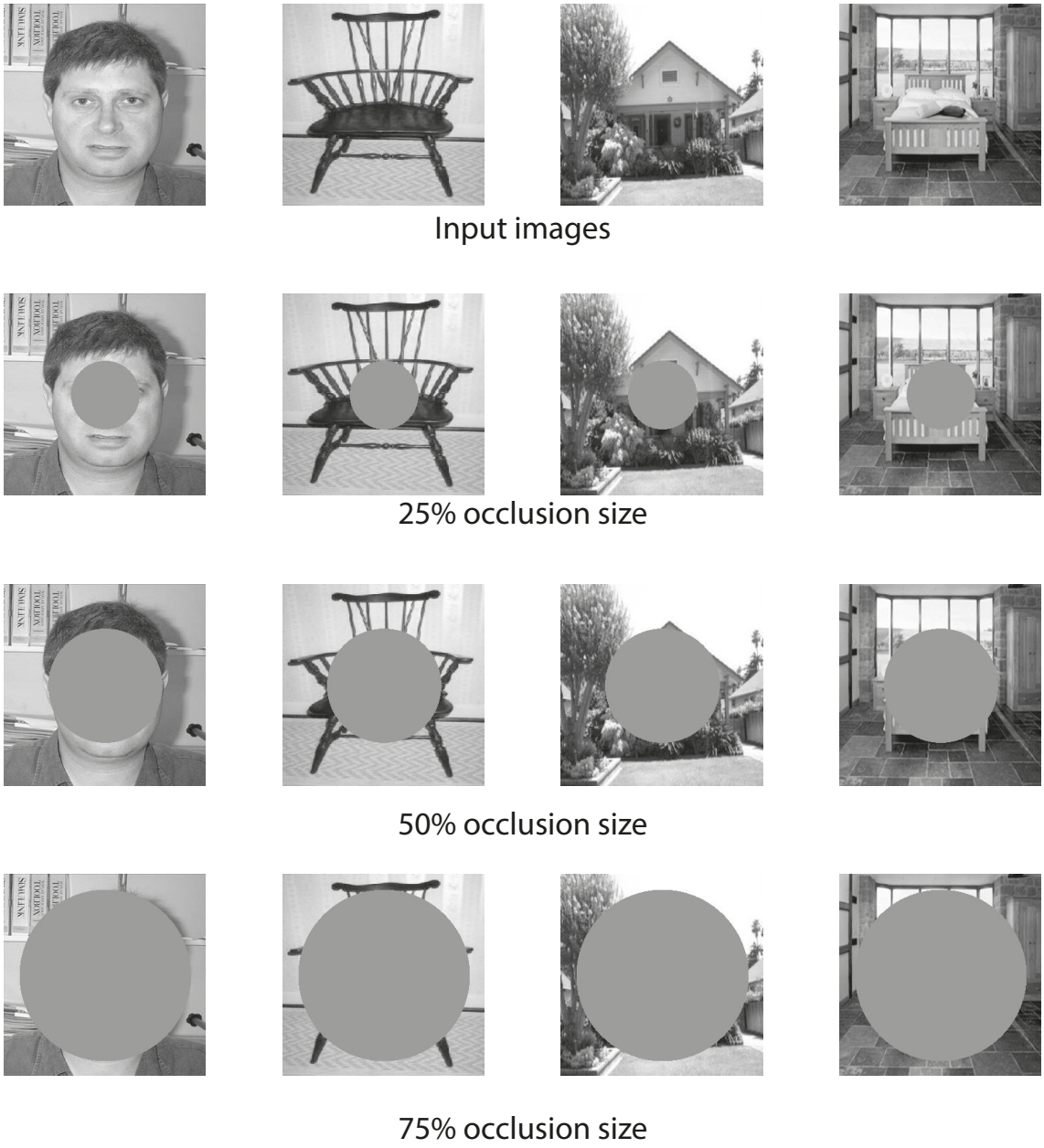


Figure 3.17: Samples of class-A occlusions applied to the images of the object and the scene datasets

plethora of scene images that belong to 15 categories. Each category consists of 200 to 400 images, with an average image size of 300 × 250 pixels. Similarly, Class-A occlusion type was applied, to evaluate the model robustness.

3.12.1.2 klab Dataset

A set of various partially occluded stimuli was created by the klab [141]. The dataset comprises patterns rather than objects or scenes. Therefore, it is considered challenging. Sensitive parts of each pattern were occluded. The labels of this dataset are based on human subjects decision. The recognition process is based on under-

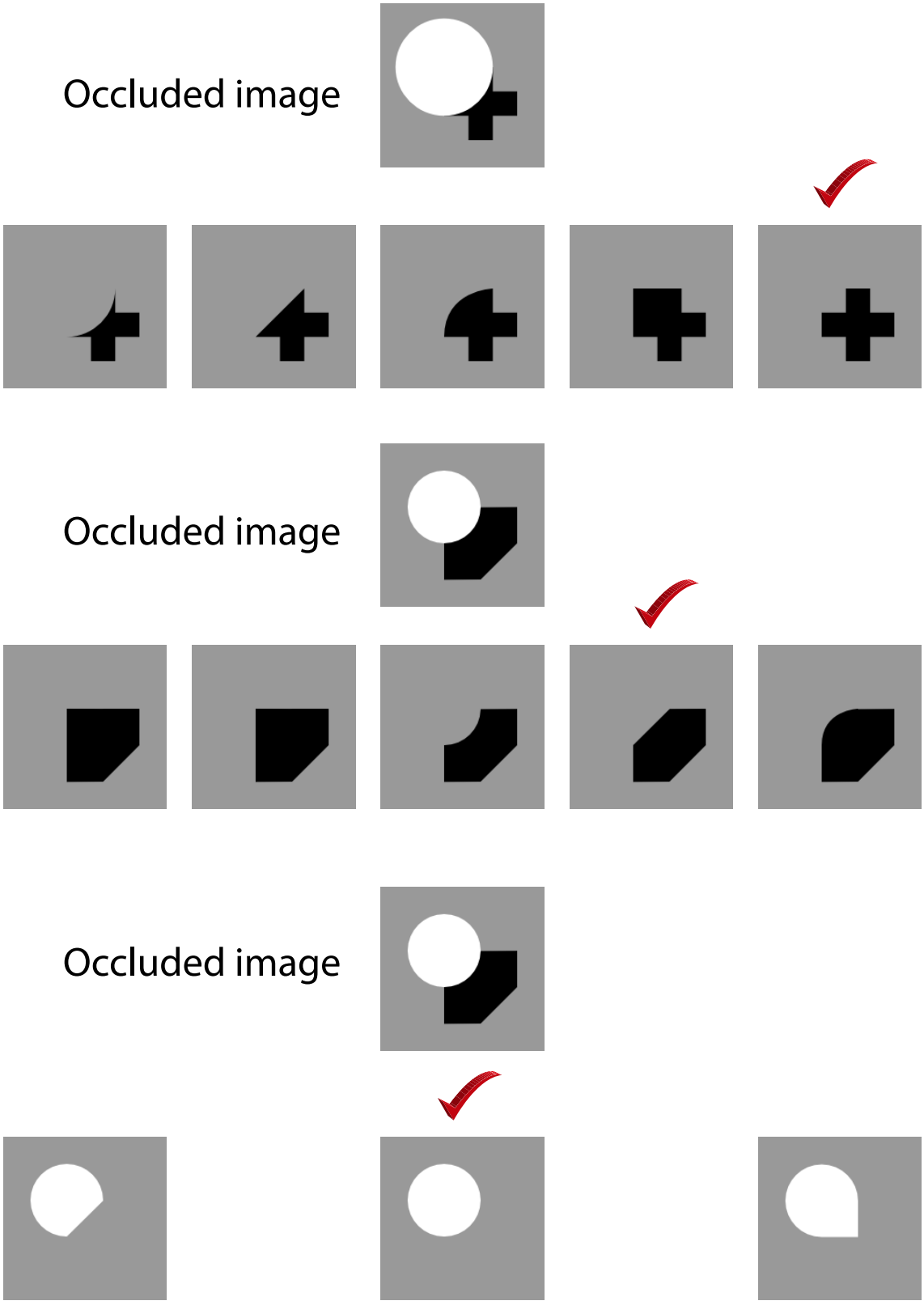


Figure 3.18: Samples of class-B occlusions [141]. The patterns above are disordered by other patterns. Images below are the potentials of the original image.

standing the original pattern to decode the occlusions, see Figure 3.18.

3.12.2 Occlusions

Class-A Occlusions

To test the robustness of the En-HMAX model, artificial occlusions were applied to the objects image dataset and the scene image dataset. Class-A occlusion is a classical type of occlusions, in which the occluded part of the images is filled with pixel values of 128. Different areas of the images were occluded using a variation of occlusion size. Therefore, for each experiment, important parts of the objects and scenes are blocked.

Class-B Occlusions

Class-B occlusions require sophisticated mechanisms to retrieve the original patterns. Examples of Class-B occlusions are shown in Figure 3.18. Human subjects were used to set the ground truth for this dataset. Experiments show that subjects tend to choose the shapes with the red marker in the provided examples corresponding to the original images. Therefore, to solve class-B occlusions, it was speculated that mechanisms of attention and top-down processing for visual associations, similar to that of the human visual cortex may be necessary.

3.12.3 Experimental Testing

Experiment 1 - Processing class-A occlusions

In this experiment, the En-HMAX model was tested with class-A occlusions of different sizes. The object image dataset and the scene image dataset were used in this experiment. As discussed earlier, the En-HMAX model was only used to process the data and extract features, with no prior knowledge of the nature of the task. The classical HMAX model was also used in this experiment using similar experimental settings. A comparison of the performances of the two models was made. The training samples were selected randomly from the original images. The testing samples were images with class-A occlusion. Cross-validation was applied over 20 independent runs. This type of validation is generally used in Caltech 101 dataset. The mean accuracy and the standard deviation were reported for each experiment for both the En-HMAX model and the HMAX model. Refer to section(3.5), for more information about the cross-validation used in this experiment.

Experiment 2 - Processing class-B occlusions

In this experiment, the En-HMAX model was tested with class-B occlusions. The feature maps for each image were extracted separately. To quantify performances, the Euclidean distance was used to measure the similarity between the feature maps, see Figure 3.19. For each feature map $\mathbf{z}$, the distance r from the original image feature map $\mathbf{p}$ is calculated as shown below:

$$r = \|\mathbf{z} - \mathbf{p}\| . \tag{3.5}$$

The smaller distance from the occluded pattern is considered as the En-HMAX model decision. The correct decisions were then aggregated over the whole image dataset. The classification accuracy was calculated based on the dataset ground truth.

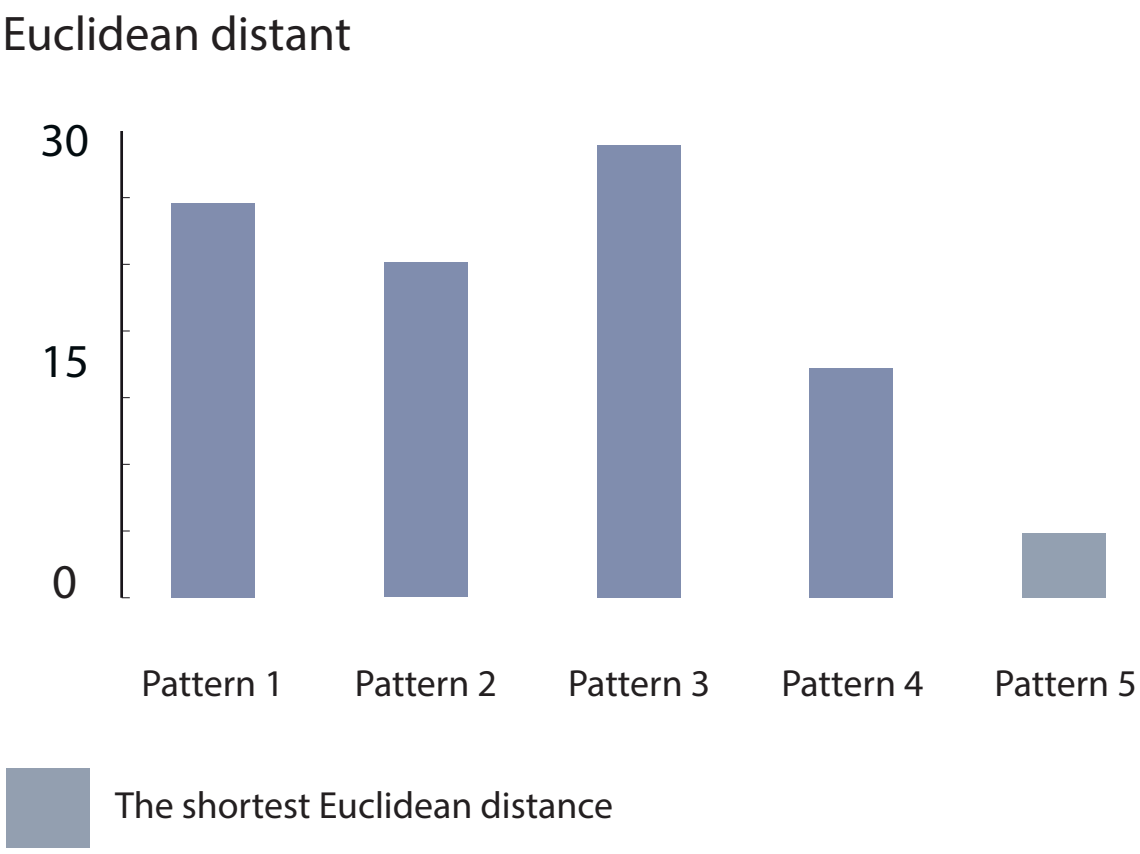


Figure 3.19: An example that shows the method of quantifying the classification accuracy of Experiment 2.

3.12.4 Results

3.12.4.1 Experiments 1

The classification results of Experiment 1 are shown in Table (3.8, 3.9). Various percentages of the occlusion size were used to test both versions of the HMAX model. On some occasions, class-A occlusions prevent the models from encountering implicit features in the images. The results show that the En-HMAX model is robust to class-A occlusions. The results also show that the En-HMAX model outperforms the original HMAX model by a large margin in all classification scenarios. It can be noticed that the HMAX model exhibits robustness against occlusions applied to both the object image dataset and the scene image dataset. In particular, when an occlusion size of 25% is applied to the images. Table 3.10 shows the confusion matrix for 50% occlusion percentage on the scene image dataset. It can be noticed that when applying occlusions, the classifier tend to become more biased toward the coast category. Yet, most classes included correct classifications. The ROC curve was used to show the classifier performance as shown in Figure 3.20. Most classes of the scene dataset have scored an AUC of 1. However, classes with the lowest AUC were bedroom, industrial and store, suggesting that they are more sensitive to

Table 3.8: Classification accuracy in a percentage of different sizes of class-A occlusions applied to the object dataset.

Objects

Occlusion size	HMAX [3]	En-HMAX [135]
∼25%	54.090 ± 0.17	99.818 ± 0.003
∼50%	43.272 ± 0.07	70.636 ± 0.05
∼75%	28.500 ± 0.04	29.000 ± 0.03

Table 3.9: Classification accuracy in a percentage of different sizes of class-A occlusions applied to the scene dataset.

Scenes

Occlusion size	HMAX [3]	En-HMAX [135]
∼25%	25.100 ± 0.13	99.166 ± 0.005
∼50%	17.466 ± 0.07	69.766 ± 0.13
∼75%	14.666 ± 0.03	20.833 ± 0.06

Table 3.10: Confusion matrix for the scene image dataset within an occlusion size of 50%.

A 50% Central Occlusion

	suburb	coast	forest	highway	insidecity	mountain	country	street	building	office	bedroom	industrial	kitchen	livingroom	store
suburb	**100.00**														
coast		**100.00**													
forest		33.33	**66.67**												
highway		100.00		**0.00**											
insidecity		40.00			**60.00**										
mountain		13.33				**86.67**									
country		40.00				6.67	**53.33**								
street		66.67						**20.00**			13.33				
building		6.67							**93.33**						
office										**100.00**					
bedroom		6.67									**93.33**				
industrial		6.67								6.67		**86.67**			
kitchen		13.33								26.67	40.00		**20.00**		
livingroom		13.33												**86.67**	
store		60.00													**40.00**

occlusions.

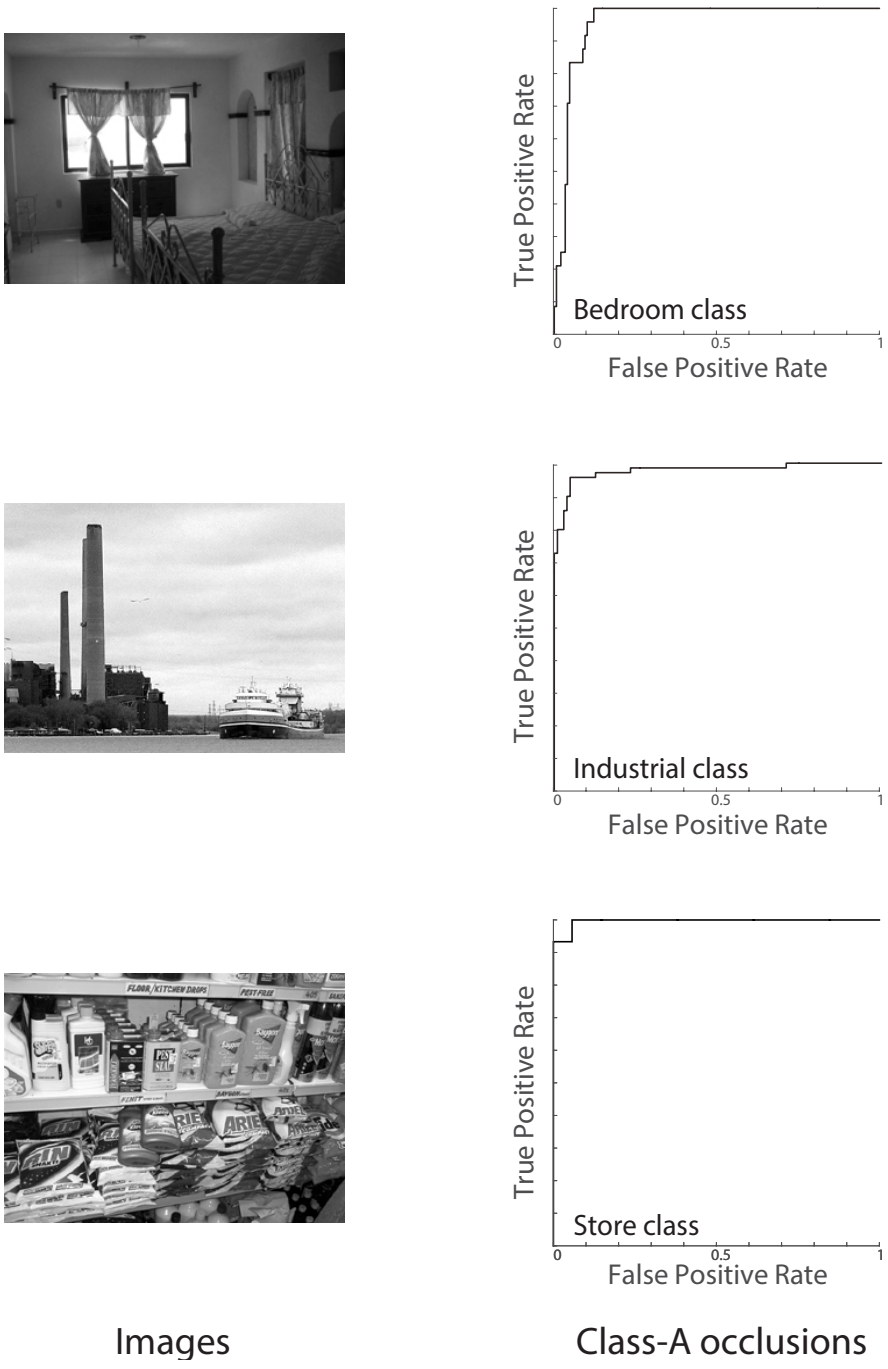


Figure 3.20: The ROC curve that shows the performance of the classifier in recognising the scene occluded images (size of 25%). All fifteen classes are included in this analysis. Only classes with the lowest AUCs are denoted in the figure. The vertical and horizontal axes denote the true positive and false positive rates, respectively.

3.12.4.2 Statistical Regularities

Applying class-A occlusions to the dataset has produced an overlapping between the input images, see Figure 3.21. However, when using the En-HMAX model to process the images, the overlapping has vanished. In particular, C_3 layer feature maps. The activations of C_3 layer are normally distributed among the different values of the spectrum.

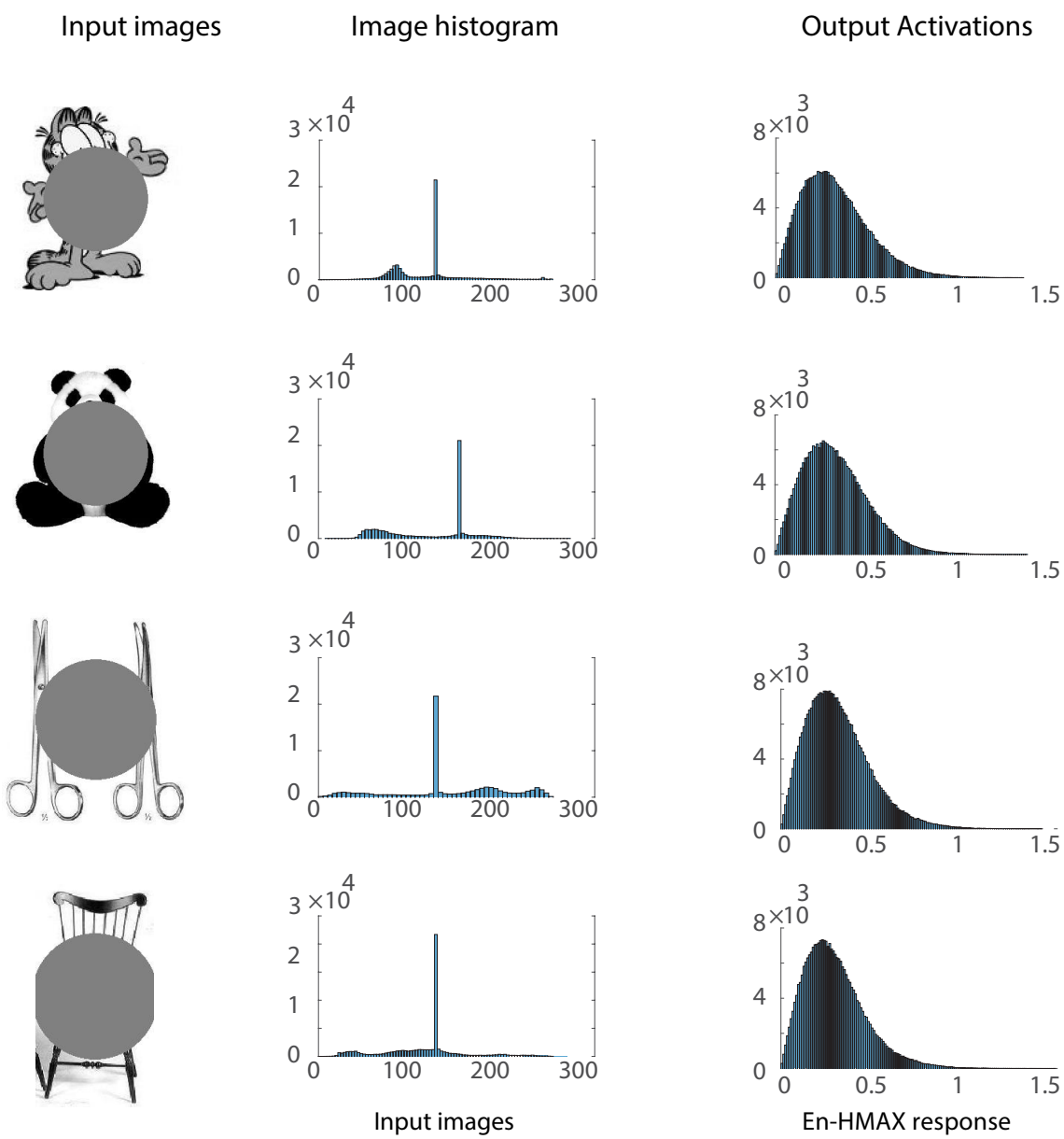


Figure 3.21: A histogram representation of class-A occlusion image dataset. First column: a histogram representation of some object images with 50% class-A occlusion. Second column: a histogram representation of the non-zero coefficients of the En-HMAX model activations.

The norm pooling and dictionary learning using the elastic net regulariser have enabled the En-HMAX model to overcome the overlapping caused by class-A occlusions.

The learning in the S layers of the model is unsupervised and sequential, and it develops through the hierarchy of the model. Therefore, the input images are filtered through the hierarchy of the En-HMAX model. As a result, the redundant data are removed from the images, and only the formative features are considered for classification.

Table 3.11: The classification accuracy in percentages for recognising klab dataset [141].

Model Architecture	Total performance in percentage
Our model	33.333 %
HMO [140]	30 %
GaborJet [142]	30 %
HMAX [3]	30 %

3.12.4.3 Experiments 2

In this experiment, the En-HMAX model was tested with class-B occlusions. The results are displayed in Table 3.11. The En-HMAX model outperformed other models of object recognition with a performance of 33.333%. The performances of other well-known models in the field were also reported in this experiment. In particular, the HMO model [140] , Gabor Jet [142] and the HMAX model [3]. However, the highest accuracy achieved was 30% as shown in Table 3.11.

With scenarios of low variation, the En-HMAX model showed acceptable performance. However, in the high-variation stimuli, it has failed to match the human performance by a large margin. It is not surprising that "ventral stream pathway" inspired models are not nearly as effective as human performance. The results shown for this dataset were excepted due to the difficulty of the task. Although the En-HMAX model has slightly outperformed other well-known models of object recognition to solve class-B occlusion, feed-forward models are still incapable of efficiently decoding this type of occlusions. In order to solve class-B occlusions, methods that mimic the mammal's visual cortex capabilities of attention and top-down processing are needed.

3.13 Chapter Summary

In this chapter, the En-HMAX model of the visual cortex was presented. It was compared with the original HMAX model, in addition to, other available states of art enhanced versions of the HMAX model. The model performance and sparsity were quantified. Details explanations of the structure of the En-HMAX model was given. The En-HMAX model provides two essential elements for image classification: selectivity and invariance. The main reason for using an elastic-net regulariser for the HMAX model was to encourage the grouping effect when the atoms in the dictionary are highly correlated. Results show that the En-HMAX model outperforms the original HMAX model (by $\sim$40%) as well as the two special cases of the En-HMAX model, i.e., the LASSO- and Ridge-HMAX models, by $\sim$19% and $\sim$9%, respectively.

Furthermore, in this chapter, the lateral connections experiment was presented. Features with different degree of complexity were investigated for recognition. The performances of different combinations of features were investigated and reported.

The En-HMAX model was then tested with two types of occlusions: class-A occlusions and class-B occlusions. Class-A occlusions are classical occlusions, in which different parts of the images were occluded artificially. Class-B occlusions, however, are more complex occlusions that require complex methods for associating the new patterns with the original images. Different sizes of class-A occlusions were used to test the performance of the En-HMAX model. The original HMAX model was tested under similar conditions. A comparison was made between the two models. The analysis performed in this chapter showed the robustness of the En-HMAX model to tackle different types of occlusions.

After developing and testing the En-HMAX model, a more intensive study of the effective regions of vision that contribute to object and scene recognition will be performed in the next chapter.

Chapter 4

Objects and Scenes Classification with Selective Use of Central and Peripheral Image Content

4.1 Introduction

In the previous two chapters, a novel object and scene recognition model inspired by the visual cortex was proposed. The En-HMAX model was tested with images of objects and scenes with a variety of conditions. The robustness of the model was then tested with occlusion. In this chapter, a careful examination of the contribution of various regions of vision in cortex-inspired models is explained. It is believed that investing such mechanisms in artificial models could extremely enhance the recognition speed of high-resolution images, due to the dramatic reduction of the processed pixels of the images.

The developed En-HMAX model was used to investigate the classification scheme of object and scene images. The importance of the peripheral and central image content was investigated individually for each image dataset. In various conditions, images were occluded by windows and scotomas of varying sizes. Furthermore, inspired by the eccentricity in the human eye, the images were processed to match a similar degree of foveation. Foveation has reduced the size of the images by a factor of $\frac{1}{2}$.

In this chapter, the developed En-HMAX model was tested for the following specific features of the human brain:

1. flexible utilization of peripheral versus central vision to enhance scene and object recognition performance;

2. central foveation to reduce the size of the processed visual data without compromising the scene recognition performance.

The En-HMAX model, alongside other well-known CNNs were used to perform this analysis. By introducing a varying number of visual angles of scotoma and window occlusions, the scene recognition process was investigated. Additionally, a second experiment was conducted to focus on the contribution of parafoveal versus peripheral areas of the images. The experiment included a larger number of visual angles on both datasets.

This chapter will first briefly discuss the structure of hierarchical models and CNNs. In addition, the image dataset and the applied window and scotoma conditions will be introduced. The process of foveation will be discussed. Then, the experimental testing settings will be explained in detail. This included introducing the classification wiring, cross-validation and statistical analysis. The results will later be discussed. Finally, a summary of the major contributions of this chapter is provided.

4.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have been developed from neural networks. They have proven successful in machine vision and objects recognition. Similarly to the HMAX model, CNNs have a similar structure that extracts the important features of the input images. It takes advantages of stacking together layers of learned filters and MAX pooling to achieve invariance. These layers are usually followed by a fully connected layer where each neuron is connected to every other neuron in the adjacent layer. Neurons in each layer operate independently from other neurons and do not share any connections. In this chapter, well-known CNN models are compared with the En-HMAX model. In particular, AlexNet [47], VGG19 [46] and GoogLeNet [48].

4.3 Scenes and Objects Image Datasets

Scene and object image datasets were utilised to test the En-HMAX model for image classification. The scene image dataset included man-made as well as natural scene images. Scene images were extracted from a scene categories dataset that was collated by Li and Perona [112] and augmented by Lazebnik et al. [111]. This scene image dataset is considered as one of the most complete scene category datasets [111].

The images of the scene dataset have different dimensions but on average are of 300 × 250 pixels. The classes in the scene image dataset are: bedroom (216 images), suburb (241 images), industrial (311 images), kitchen (210 images), living room (289 images), coast (360 images), forest (328 images), highway (260 images), inside city (308 images), mountain (374 images), open country (410 images), street (292 images), tall building (356 images), office (215 images) and store (315 images).

The object dataset includes 11 classes extracted from the Caltech 101 database [110]. The object dataset comprises the following classes: car sides (123 images), dollar bills (52 images), faces easy (435 images), garfield (34 images), inline skates (31 images), motorbikes (798 images), pagodas (47 images), pandas (38 images), scissors (39 images), trilobites (86 images) and windsor chairs (56 images).

4.4 Images With Scotoma and Window

Conventionally, the window and scotoma paradigms are used to study the basic visual processes in reading [143]. In addition, they have been utilised to study the perception within the first eye fixation on a scene, for recognising the gist [144–148]. To simulate the effectiveness of peripheral versus central information in an image, the classic paradigms of scotoma and window was applied to the image dataset [149]. The radii of scotoma and window were varied to test the performance of the En-HMAX model in the classification of original and foveated images in the scene image dataset as well as the object image dataset. Examples for applying window and scotoma on an image in its original and foveated forms are shown in Figure 4.1(B, C). The scotoma and window were utilised in the experiments of Larson and Loschky [149]. They were used on human subjects to demonstrate recognition accuracy to investigate the contribution of peripheral versus central vision. The term “window” is originated by the analogy of looking at a scene through a window,

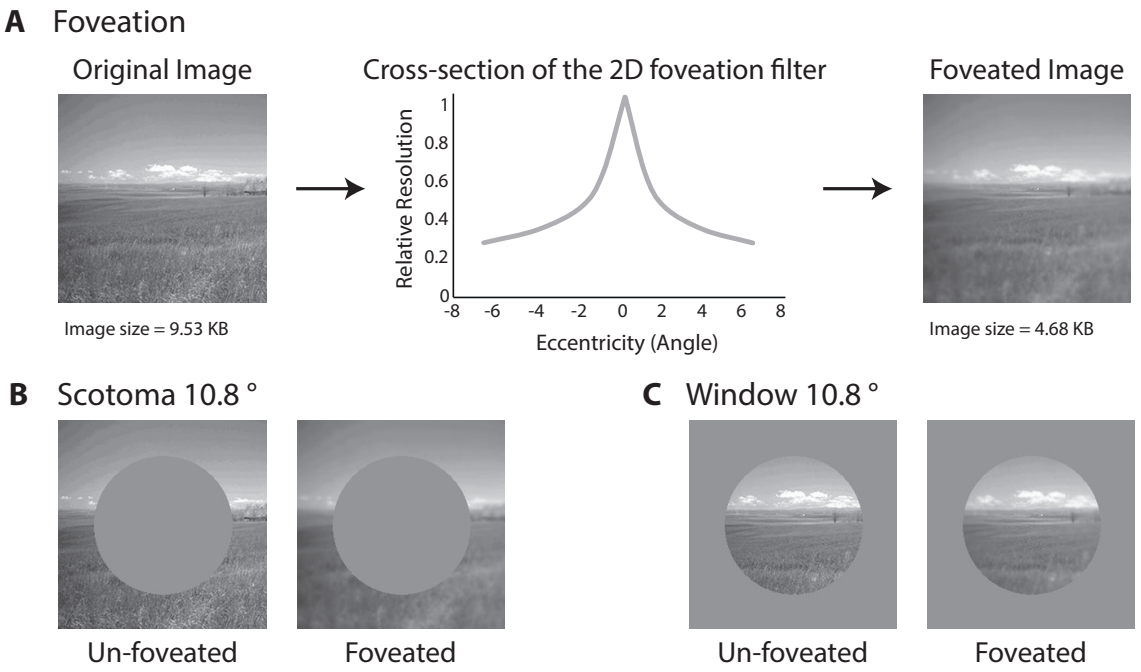


Figure 4.1: An example of pre-processing an image with Foveation, scotoma and window conditions. Window condition is when a circular region blocks the peripheral. In the scotoma condition the central area is blocked and only the periphery is shown. (A) Foveating an image using a 2D filter. (B) Examples of the scotoma condition. (C) Examples of the window condition. The image shown in the figure is extracted from a scene category dataset [111, 112].

for example, a porthole. In the window paradigm, the visual information outside the window is absent. The term "scotoma" is derived from an analogous medical condition in which a certain part of the visual field is blocked. In the scotoma paradigm, the information outside the scotoma is unaltered and the centre-based information is blocked. In a similar fashion, scotoma and window paradigms were used to investigate the role of peripheral and central vision in the En-HMAX model.

The rationale for using the window and scotoma paradigms was to introduce disorder to the images on a selected region of vision to quantify the decline in En-HMAX performance in recognising image categories. In line with the earlier mentioned experiments [146, 149], scotoma and window have been generated using the filters shown below:

$$\begin{aligned} h_g(n_1, n_2) &= \exp\left(\frac{-(n_1^2 + n_2^2)}{2\sigma^2}\right) \\ h(n_1, n_2) &= \left(\frac{h_g(n_1, n_2)}{\sum_{n_1}\sum_{n_2} h_g}\right) \end{aligned} \tag{4.1}$$

where σ denotes the standard deviation, $h_g(n_1, n_2)$ corresponds to the distribu-

tion function, $h(n_1, n_2)$ are the generated normalised multivariate Gaussian filters and (n_1, n_2) represent the filter's dimension. The mask was then discretised by setting all pixel values inside/ outside the mask to 128 to form the scotoma and window, respectively.

In the first experiment of this study, the distance between the model and the images was assumed to be 26.6 inches. This assumption is, in line with the settings in the experiment by Larson and Luaschy [149], where the number of pixels per degree is calculated based on a trigonometric notation [150]. The value of σ in equation (5) was adjusted accordingly. However, in the second experiment, in order to allow the En-HMAX model to be more generalizable to various viewing settings, a wider range of scotomas and windows was applied.

4.5 Foveation

The human visual system segments slower higher resolution acquisition and faster lower resolution acquisition into the central and peripheral regions of the retina, respectively. The retinal information decreases in resolution towards the periphery without compromising performance in scene recognition [149]. Therefore, foveation was introduced in the below experiments. The effect of introducing foveation to the images was investigated with regard to the En-HMAX model.

The amount of information of the visual scene varies depending on the location of the fixation point. The fixation point corresponds to the fovea, i.e., the centre of the eye's retina, and demonstrates the highest resolution in the scene. An example of the fixation point is observing on the computer screen the computer mouse pointer.

The foveation in the human eye can be explained by the non-uniform distribution of the ganglion cells and the photoreceptors in the retina. The eye quality of decoding the visual scene depends immensely on the ganglion cells. The density of these cells has a high value at the fovea and drops dramatically toward the periphery. When a human observes the scene, images with different resolutions are transmitted to the front visual channel, and accordingly to the early layers of the visual cortex of the brain. Regions of vision around the fovea are perceived with the highest resolution and sampled with the highest density.

To simulate human foveation, a pyramid of low pass filters was used [151]. Each input image, e.g. Figure 4.2A, was passed through six repeated layers of filters

cascaded with a down-sampling stage. Starting from the centre of each image, the filtering and down-sampling parameters were set such that at each pyramid layer the image resolution was halved [151]. A cross-section of the symmetrical resolution maps used to generate eccentricities is shown in Figure 4.2A. The maximum relative resolution was assigned to the centre. Relative resolution declined smoothly toward the periphery. Foveation was applied to the input images such that the contrast c is calculated with

$$c(f, e) = c_0 \exp\left(\alpha f \frac{e + e_2}{e_2}\right) \tag{4.2}$$

where f is the spatial frequency, c_0 is the minimum contrast threshold, e is the retinal eccentricity, e_2 is the half-resolution eccentricity and α is the decay constant.

In computer vision, foveation can be considered as an image compression method [152]. It reduces the size of the images, and the computational resources required to process them, i.e., the speed of encoding and decoding. Figure 4.1A shows an example of a foveated input image with a size reduction of $\sim 52\%$.

4.6 Experimental Testing

Two experiments were conducted to investigate the relative value of peripheral versus central data in rapid categorisation of scene and object images.

Experiment 1: The contributions of peripheral and central image content to classification

The rationale of this experiment was to quantify the performance of the En-HMAX model in the classification of various scene and object images data under window and scotoma occlusion conditions. This experiment comprised two parts. In part one, the original images of the scene and object datasets were classified. In part two, all images were first foveated before repeating the analysis exactly as in part 1.

In both parts of Experiment 1, the En-HMAX model and the classifier were trained with the full, original images. The En-HMAX model and the classifier were then tested with the same images but overlaid with windows or scotomas of four different visual angles, namely, 1°, 5°, 10.8° and 13.6°. Figure 4.2A depicts fully the arrangement of training and testing data in Experiment 1.

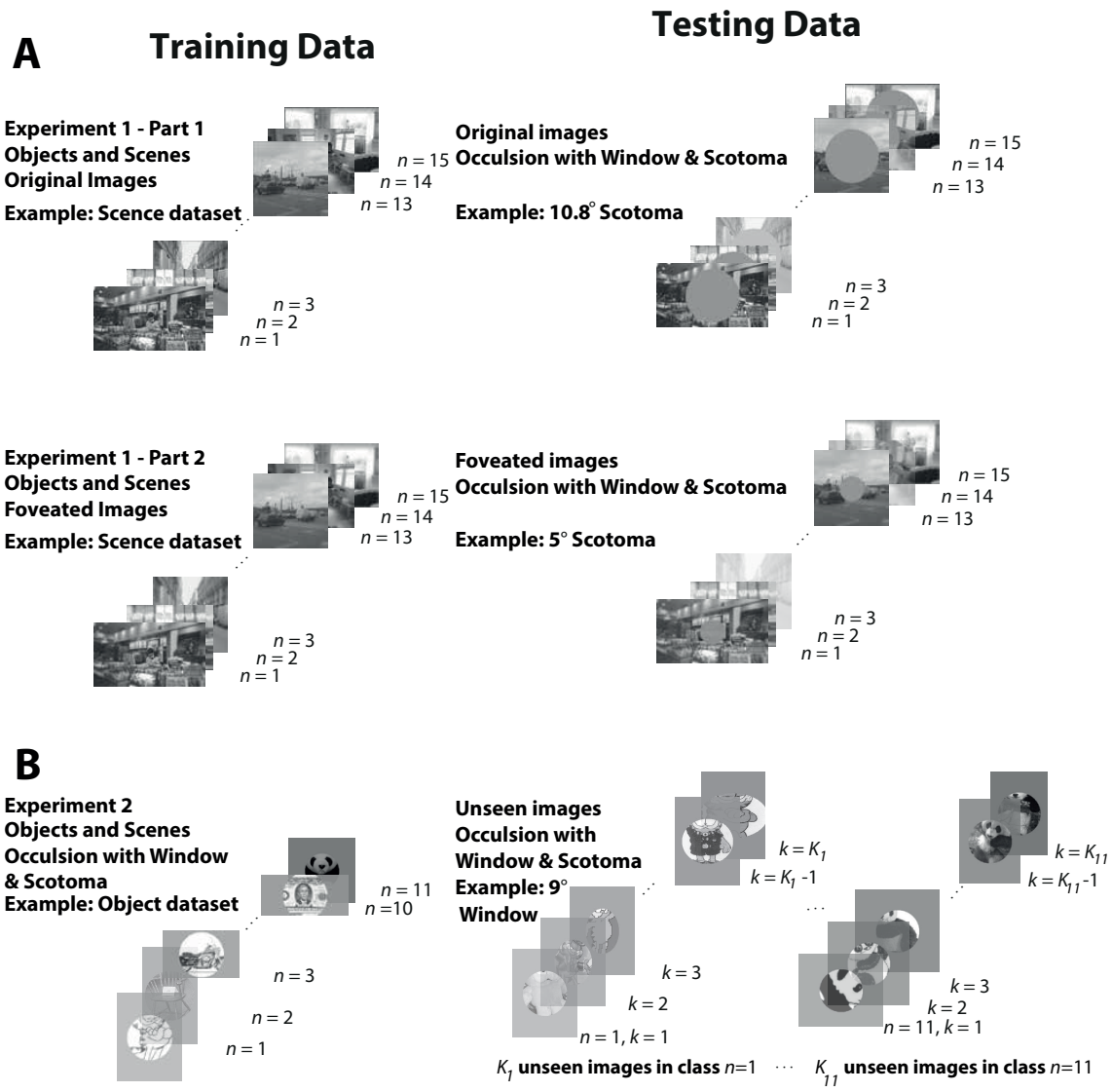


Figure 4.2: The configuration of the experiments. Similar settings have been used for Experiment 1 and Experiment 2. In Experiment 2, the number of testing images varies, depending on the size of each class. The letters n and k represent the class number and the image number in each class, respectively. Images shown in the figure are extracted from Caltech 101 dataset [110] and scene category dataset [111, 112].

These visual angles for the window and scotoma were selected following Larson and Loschky [149]:

- the presence and absence of foveal information (1°);
- the presence and absence of parafoveal and foveal vision against the peripheral vision (5°);
- the representation of approximately equal viewable area inside the window and outside the scotoma (10.8°). This is calculated on a per-pixel basis, and averaged across the whole dataset; and finally

- the presence and absence of peripheral information (13.6°).

Experiment 2: Generalisation to Unseen Images

The motivation behind this experiment was to study the ability of the En-HMAX model in generalisation to unseen images when trained with occluded images of scenes and objects. The motivation was to explicitly measure the efficiency of each region of vision in both datasets. A larger number of visual angles with a smaller step size was used, for instance, blocking the central vision of the entire scene dataset and observe how the model behaves with these changes.

The En-HMAX model and the classifier were trained with images of varying visual occlusions, ranging from 1°–19° with a fixed step size of 2°. The En-HMAX model and the classifier were then tested with unseen images for every visual occlusion angle. An example of this classification design for the objects dataset with a 9° window is shown in Fig.4.2B. This experiment can produce a more precise measurement of the effective region of vision depending on the dataset.

4.7 Classification

A multi-class linear support vector machine [70, 72] was used to classify the images. In particular, the LIBLINEAR library [72] was utilised. To solve the multi-class problem, the one-vs-the rest method was used, as implemented in LIBLINEAR. The image data were divided into training and testing sets. In the two experiments, there was a different number of testing images in each class. Therefore, to avoid bias, the aggregated classification scores were normalised across categories. In addition, the receiver operating characteristic curve (ROC) [129] was calculated for all of the classes used in Experiment 1. Various thresholds were used for each class to perform the binary classification. The classes which respond poorly against other classes have an ROC curve close to the diagonal, while classes that respond selectively against other classes have a curve far from the diagonal. Thus, the ROC curve was considered as a quantitative analysis of the classifier selectivity and specificity.

4.8 Cross-validation

In both parts of Experiment 1, the following cross-validation approach was adopted. A fixed number of images per category was used to report the overall error rates, with the same number of test samples. The total number of images per category in Experiment 1 was 15 images. Ten images per category were selected randomly to train the En-HMAX model. The same number of images was used for testing.

In Experiment 2, a subset of 30 images per class was selected randomly to train the En-HMAX model. The remaining images in each class were used for testing. The number of test images varies depending on the overall number of images in each class (31 images to 798 images). In both experiments, the classification for 20 independent runs was repeated. The average classification scores and the standard deviations were reported for each classification scenario.

4.9 Statistical Analysis

To test the statistical significance of the main findings in the conducted experiments, a paired t-test and a sign test were performed. For every sample of data, a statistical test for normality was performed using the two numerical measures of shape: the skewness and excess kurtosis. For normally distributed data, the paired t-test was used, and for non-normal data, the non-parametric sign test was used. Following the main analysis, posthoc comparisons were performed.

4.10 Results

4.10.1 Experiment 1

Figure 4.3 shows the results of Experiment 1 (both parts); in which the En-HMAX model was trained with complete images and was tested on images with window or scotoma occlusions. Results are reported for original and foveated images in Fig. 4.3A and B, respectively.

In the scene classification, in the cases of 1° and 5° visual angle scotoma (Figure 4.3A), accuracies of approximately 89 ± 1% were achieved. A non-parametric *sign test*, with a risk $\alpha = 0.05$, shows that there was no significant difference in the scores

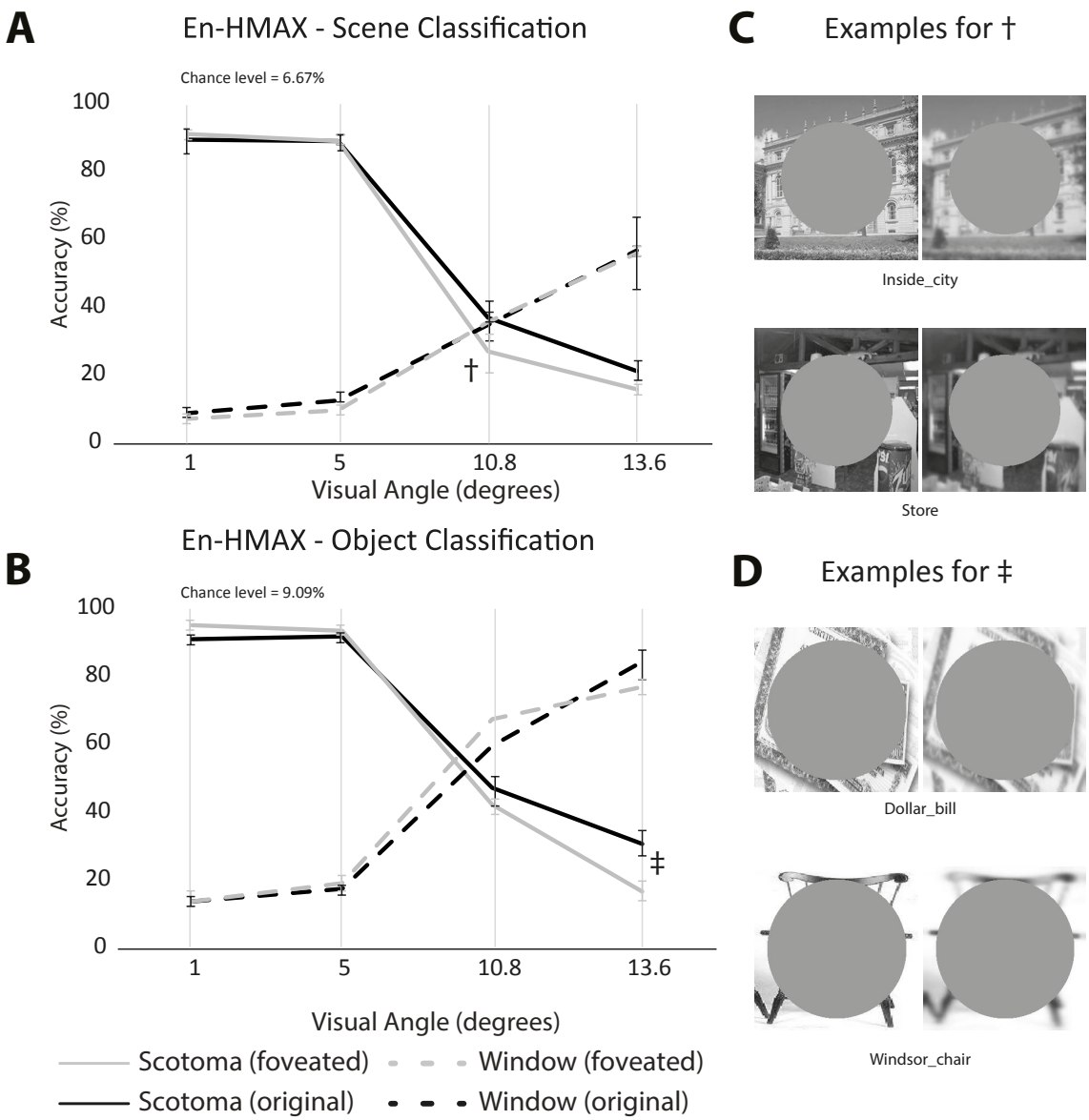


Figure 4.3: Classification accuracy with the En-HMAX model as a function of visual angle and viewing condition (scotoma and window) for scene (A) and object (B) images with and without foveation. (C) Examples for the 10.8° scotoma condition for both the original and the foveated data. (D) Examples for 13.6° scotoma condition for both the original and the foveated data. Images shown in the figure are extracted from Caltech 101 dataset [110] and scene category dataset [111, 112].

for 1° scototma (M=89.2, SD = 5.6) and 5° scotoma (M= 88.9, SD = 4.2); z(19) = 0.8, p = 0.3.

This indicates that the En-HMAX model can achieve the maximum performance even in the absence of parafoveal vision. Further increase in the size of scotoma led to a considerable degradation of the classification scores, see Figure 4.3(C, D), for instance, the accuracy at 13.6° scotoma reduced to 23 ± 6%. This level of accuracy was however above the chance level accuracy (6.67%; 15 classes). Predictably, the scene classification performance was poor at 1° and 5° visual angle window con-

ditions. This score increased as the window became larger such that at the 13.6° window condition, the classification accuracy reached 57%.

In the 10.8° and 13.6° scotoma conditions, the classification scores for original images were significantly higher than that for foveated images. A *paired samples t-test*, with a risk $\alpha = 0.05$, was used to quantify the significance of these results. For scotoma 10.8°, there was a significant difference in the scores for original images (M = 37.1, SD = 12) and foveated images (M = 27.4, SD = 11.1); t(19) = 3.0, p = 6.5×10^{-3}. Similarly, for scotoma 13.6°, there was a significant difference in the scores for original images (M = 21.6, SD = 7.9) and foveated images (M = 15.9, SD = 4.8); t(19) = 2.6, p = 0.016.

In addition, in the window condition, irrespective of the window size, no significant difference was observed between the classification scores for the original and the foveated images of scenes.

Figure 4.3B shows object classification results. Classification accuracies exhibited the same trends as in Figure4.3A. However several interesting observations were made:

- For object classification, the cross-over of window and scotoma conditions occur at visual angles of 9.7° (original) and 9° (foveated). However, for scenes, it shifted to right to 10.8° (original) and 10.8° (foveated). This indicates that the En-HMAX model relies more on the central image content for recognising objects.

- At window 13.6°, a non-parametric *sign test*, with a risk $\alpha = 0.05$, shows that the classification performance of objects (M = 84.7, SD = 7.6) was significantly higher than that of scenes (M = 57.1, SD = 25.5); z(19) = 2.6, p = 7.2×10^{-3}. This indicates that the peripheral region of the scene images is more effective for the recognition process.

- At the 13.6° scotoma condition, a *paired samples t-test* shows that the objects classification score achieved for the foveated images (M = 17, SD = 6.2) was significantly lower than that observed for the original images (M = 31.1, SD = 6); t(19) = 6.9, p = 1.32×10^{-6}.

Figure 4.4 shows the accuracies of individual classes in the scene dataset. The scene images were categorised according to whether they were natural (green), man-

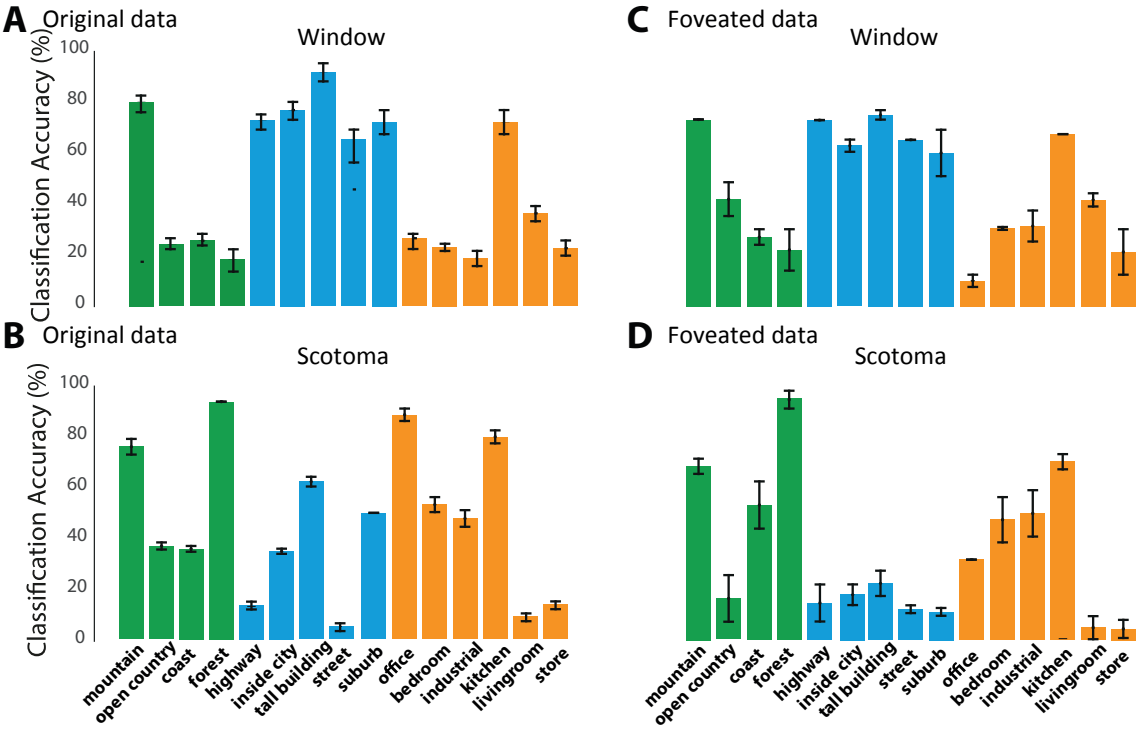


Figure 4.4: Individual class accuracies for the scene dataset at an angle of 10.8° in the window and scotoma conditions. The classes are categorised according to whether they are natural (green), man-made and out-door (blue) or man-made and in-door (amber) scenes.

made and out-door (blue) or man-made and in-door (amber) scenes. Only the following viewing conditions were used to both the original and the foveated version of the images: window 10.8° and scotoma 10.8°. The rationale for selecting only these two conditions was to observe how scene classification was affected when central or peripheral image content was blocked. Interestingly, for this dataset, the performance drop was not category-dependent, for instance, some of the classes, such as the mountain and kitchen, retained good classification accuracy in all scenarios whilst other classes did not. Another interesting observation was that out-door scene images were least affected by the 10.8° window occlusion. However, it was observed that classification accuracy dropped more when scene images were masked with a 10.8° scotoma than when they were masked with a 10.8° window. This highlights the difference between outdoor images and natural images in terms of the location of the features.

For completeness, the En-HMAX model was compared with the original HMAX model in terms of the individual class accuracies. Figure 4.5 shows that the En-HMAX model outperforms the HMAX model in recognising the datasets individual accuracies. Markers below the diagonal indicate that the En-HMAX model outper-

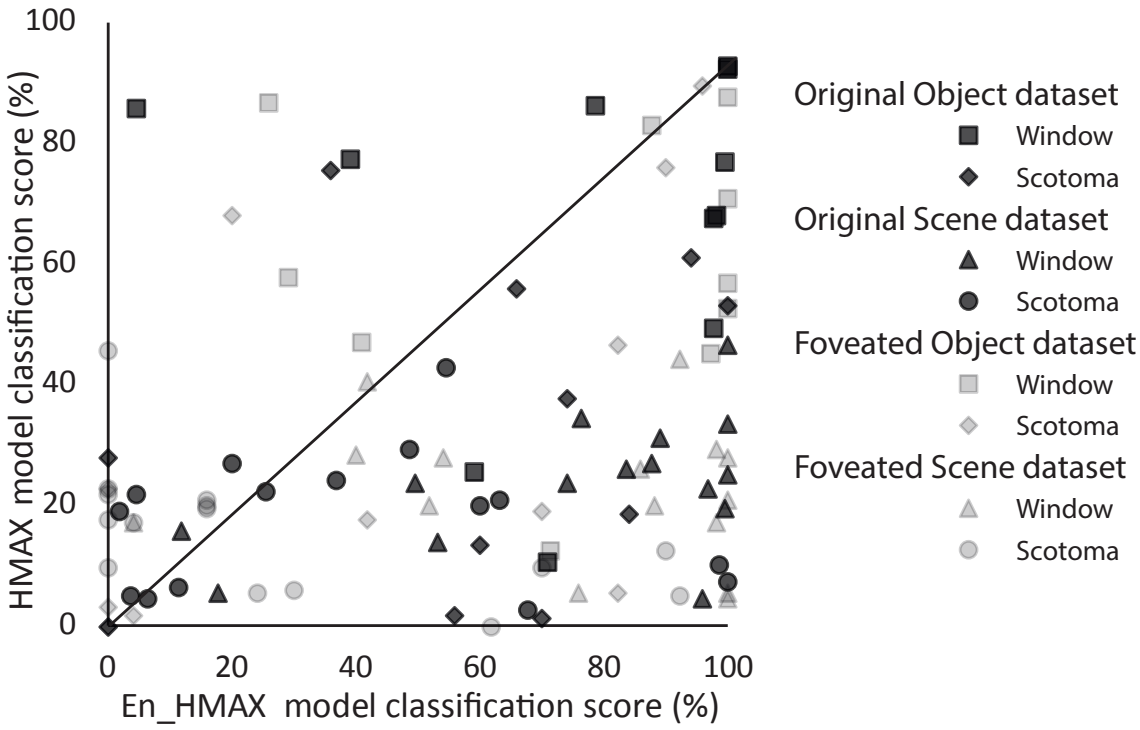


Figure 4.5: A comparison between the accuracy of the En-HMAX and the HMAX models. Markers of the scattered diagram represent the classes accuracies of both models. Classes below the diagonal indicate that the En-HMAX model outperforms the HMAX model. The figure shows accuracies of the 10.8° scotoma and the 10.8° window conditions. Both the original and foveated dataset were used in this analysis.

forms the HMAX model, in recognising a certain class of the dataset. A visual angle of 10.8° was used in both, scotoma and window conditions as a representative example. There are 104 markers in Fig.4.5, of which 76 markers lie below the diagonal line. Two markers located exactly at the diagonal line, and only 26 markers appear above the diagonal line.

From the receiver operating characteristic (ROC) curve shown in Fig.4.6A, it was observed that the En-HMAX model responded selectively for the majority of the classes. A larger area under the curve (AUC) has been reported when the selected visual angle is 5° scotoma. In Figure 4.6(B), the confusion matrix was used to visualise the performance of the individual classes. Each class in the vertical axis describes the instances in an actual class while each column describes the instances in a predicted class. With a 5° scotoma, all the classes of the dataset were identified successfully. The classification scores were high even when the foveation was introduced to the images, as shown in Figure4.6(B). This is more visible in the scene dataset. This indicates that the recognition process can be completed successfully without relying on the parafoveal vision.

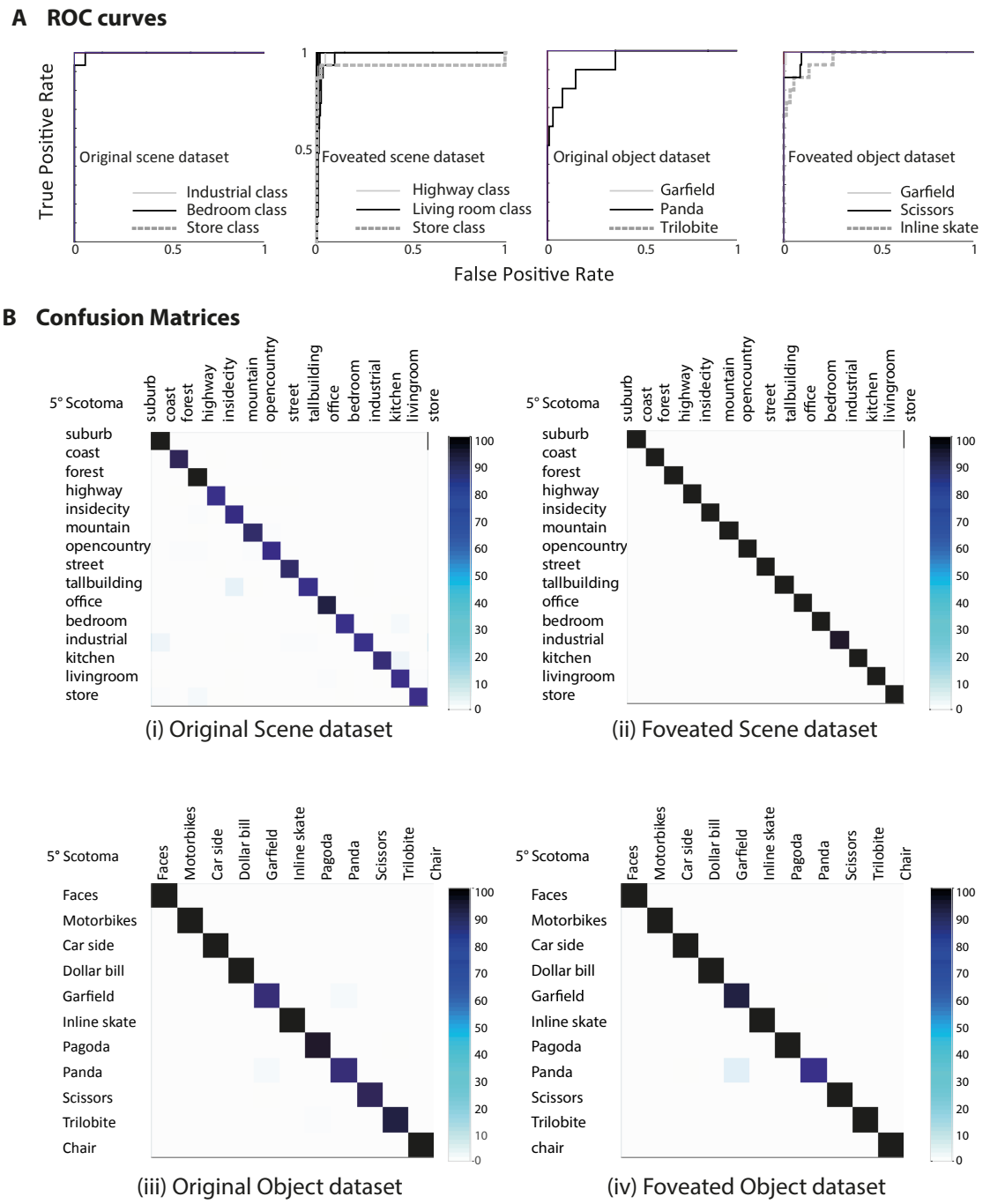


Figure 4.6: Classification analysis of Experiment 1. (A) The ROC curves of our used datasets within a visual angle of 5° scotoma. All classes have been included in the analysis. Only classes with the lowest area under the curve (AUC) are visible. The vertical and horizontal axes denote the true positive and false positive rates, respectively. (B) Confusion matrices are for the 5° scotoma condition. The vertical axis represents the actual classes, and the horizontal axis represents the predicted classes. The scores have been averaged over 20 independent runs.

Measuring the efficiency of each region of vision, the peripheral vision has proved to be more efficient to achieve maximum recognition performance. Classification accuracies of up to ∼90% for scenes and objects were possible. Window and scotoma

analysis suggested that object and scene recognition were sensitive to the availability of data in the centre and the periphery of the images, respectively. Similar to the observations made in human studies, the experiments showed that the En-HMAX model has utilised a relative order of importance depending on image category. The obtained modelling results have matched the hypothesis that centre-based vision is more important than the peripheral vision for recognising objects. Also, the results showed that introducing foveation does not compromise the recognition performance even when the parafoveal vision is blocked.

4.10.2 Convolutional Neural Networks

For completeness, three well-known CNN models were selected (AlexNet [47], VGG19 [46] and GoogLeNet [48]) to reproduce the results of Experiment 1. CNNs showed a similar pattern to the En-HMAX model in object recognition tasks (Figure 4.7), for instance, cross-over points locations. Similarly to the En-HMAX model, the cross-over points in the object dataset are located to the left of that in the scene dataset. This suggests that CNNs rely more on the central image content for recognising objects. For the scene image dataset, similar prioritisation for the peripheral data is observed as the followings:

1. The cross-over points of the peripheral and central vision for scene classification are located to the right from that of object classification on the spectrum of the visual angles;

2. the poor classification performance when the peripheral vision is blocked at window 13.6° for scene image classification.

At the same time, the recognition pattern of CNNs showed differences from that of the En-HMAX model recognition pattern, for instance, the drop in performance has dramatically increased in the absence of the parafoveal vision at scotoma 5°. This shows that the En-HMAX model relies more on the peripheral image content for recognising scenes, due to its abstract architecture. The similarity in the behaviour between the CNN models and the En-HMAX model suggests that they both prioritise similar features in the images. The similarity might suggest that both structures utilise sophisticated visual eccentricity biases, as the primate visual system does.

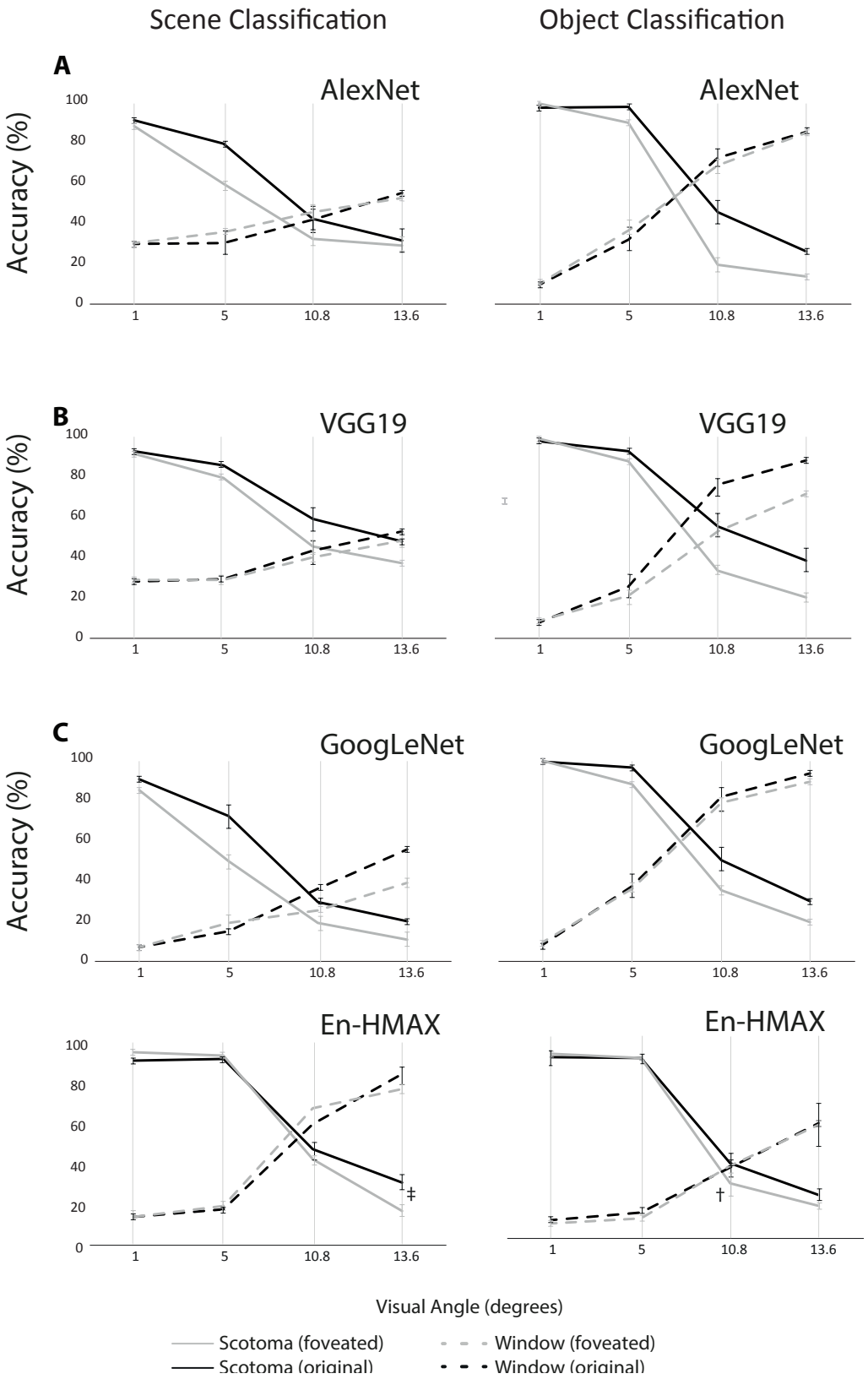


Figure 4.7: Replicating Experiment 1 using three well-known models of CNN, namely, AlexNet [47], VGG19 [46] and GoogLeNet [48].

4.10.3 Experiment 2

In Experiment 2, the behaviour of the En-HMAX model was computationally tested in classifying occluded scene and object images with windows and scotomas of vary-

ing radii. Figure 4.8(A) shows that the recognition accuracy for recognising unseen images of the scene dataset was stable to the point that more than 50% (a visual angle of 10.8°) of the image data was blocked by the scotoma. However, when the scene dataset is peripherally blocked by the window conditions, the performance starts dropping earlier from a visual angle of 13° and downward. In Fig.4.8(B), the performance of object classification under the window condition is almost symmetrical. However, across the whole spectrum of visual angles, the performance under the scotoma condition was lower than that of the window in a large margin. This observation reaffirms that object recognition is more dependent on the central image content.

The performance of object classification in the window condition declined dramatically from $\sim 80\%$ to $\sim 58\%$ in the range of 7° to 3°. In the scene classification and in the presence of scotoma, a similar decline in performance took place between 13° to 17°. However, the reduction in correct classification from $\sim 75\%$ to $\sim 60\%$ was less when compared to the reduction observed for object classification. When normalised to the maximum score achieved in each condition, these reductions were $\sim 23\%$ to $\sim 10\%$ in the window versus scotoma, respectively. This dramatic decline in the object classification trend occurred when the visual data around the parafoveal vision was blocked. As a result, the En-HMAX model may behave differently at this particular range of the visual angles.

The results have shown that higher performances in recognising unseen images of objects and scenes can be achieved using only the more relevant image content, i.e., peripheral vision for scene recognition and central vision for object recognition. Results show that $\sim$50% of the visual field would be enough to achieve $\sim$96% of the maximum accuracy in classification of unseen images. It was envisioned that this method could be invested for a large scale categorisation by reducing the amount of processed data.

4.11 Discussion and Concluding Remarks

Models of scene recognition [127, 153] show that rapid categorisation can be performed at the early perceptual stages of the visual cortex hierarchy [154, 155]. Experimental results on human subjects have shown that with a stimulus of an exposure time of 100ms, humans can categorise scenes at both: the super-ordinate level (e.g.

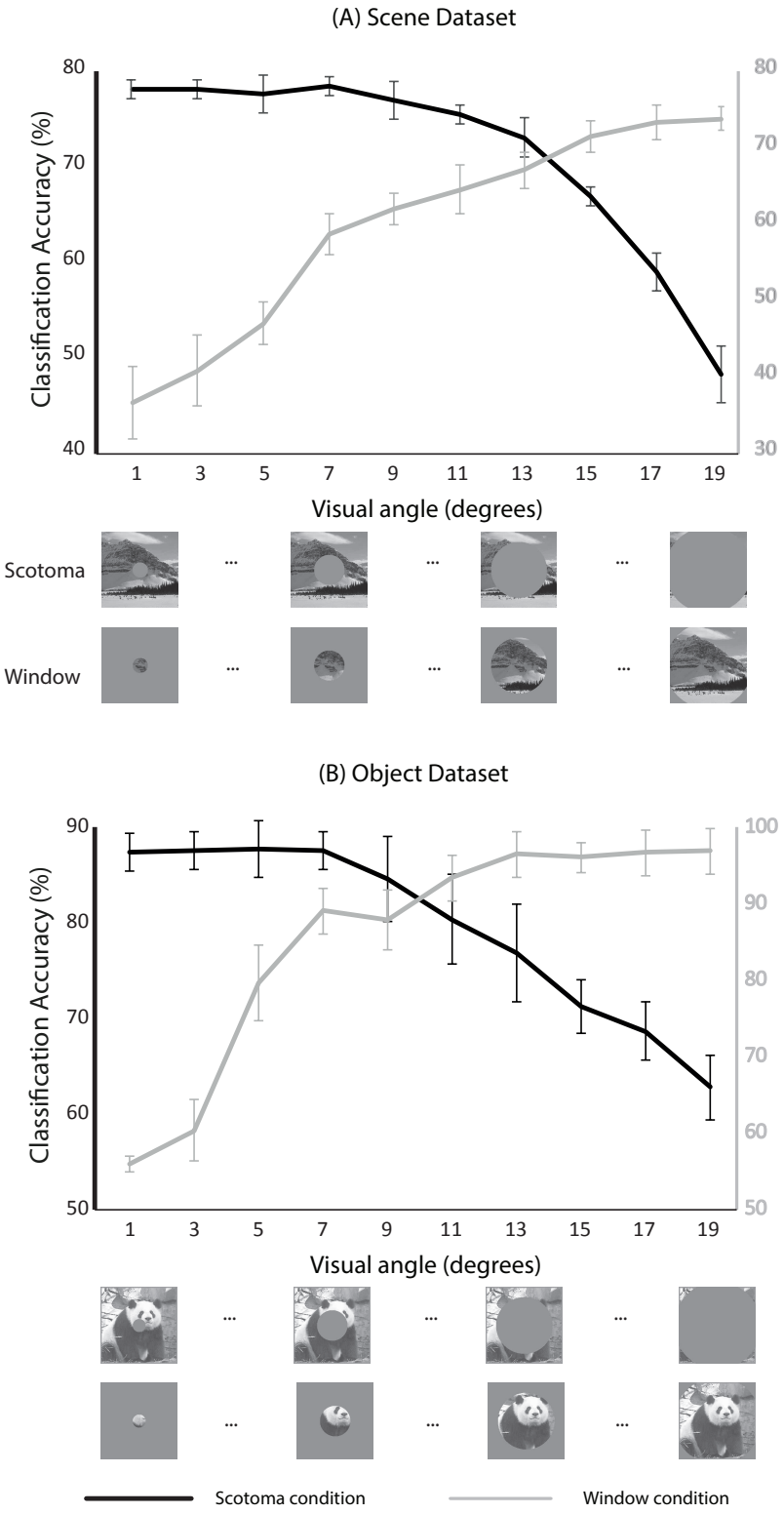


Figure 4.8: The classification accuracy trend over percent of each of the shown visual angles. The above scores have been calculated with respect to unseen images of both the scene dataset and the object dataset. Images shown in the figure are extracted from Caltech 101 dataset [110] and scene category dataset [111, 112].

man-made versus natural) and the basic level (e.g. highway versus forest) [156].

This study showed that peripheral image content, that is beyond 5° eccentricity, is more efficient in recognising the gist of a scene than central image content.

Varying the scotoma size below that range did not reduce the model performance. In Experiment 2, two types of datasets were tested, namely, object and scene image dataset. Applying the same conditions to both datasets, a similar conclusion was reached, that is, the peripheral vision for scene images are more important than central vision and the reverse is also true for objects. Additionally, this study showed that foveation has no significant impact on gist recognition in the absence of parafoveal vision, that is 5° scotoma (Figure 4.3). This finding indicates that a maximum performance of scene recognition can be achieved using only the foveated peripheral image content.

The advantage of central vision in object recognition is mainly explained by the fact that objects are generally located in the centre of the images. This indicates that the model is performing recognition based on the objects within the images and not their backgrounds. Also, when normalising performances in Experiment 2, the decline in object recognition was 13% faster than the decline in scene recognition, especially when occlusions block parafoveal vision in the range of 7° to 3°. This observation corroborated the importance of parafoveal vision for object recognition [93].

Interestingly, outdoor man-made scene classes were less dependant on the peripheral image content. With a 10.8° window, these type of scenes scored relatively higher performance. This is due to the alteration of the global properties of scene categorisation schemes [157]. Scene recognition depends on local features settings within each type of scenes [33]. Examples of local features are the presence of cars, pedestrians, and bicyclists in a street in outdoor-man made scenes [14]. Therefore, the En-HMAX model can extract local features across man-made scene images without particularly relying on the peripheral vision.

In Experiment 2, this finding was further investigated using a larger scene and object image datasets. the relative importance was inferred for each region of vision for both datasets: peripheral image content for scene dataset and central image content for object dataset. Blocking the less relevant image content produced the same performance pattern in both scenarios. This finding can help reduce both the computational and time requirements to perform classification.

4.12 Chapter Summary

In this chapter, an investigation to the contributions of peripheral versus central vision and its effect on the En-HMAX recognition process was performed. The proposed approach can dramatically decrease the size of the required visual data for scene and object recognition. State of art models of object recognition may be too computationally expensive to run on a computer with modest specifications. Three possibilities to overcome this problem: local processing, cloud processing and a mixture of the two. Cloud processing remains an important tool especially for devices with low processing capability. Most systems use a mixture of local processing and cloud processing, given the increasing power of mobile graphics units. However, transferring all image data to a remote cloud may be an unrealistic solution, due to the band-width related issues [158]. This means that the foveation could be performed locally to reduce information and the rest done in the cloud. As such, significant data reduction can be particularly attractive. An important finding of this chapter was that the maximum classification performance, equal to when the whole image is available, can be achieved with only half of the input image content. This observation sets Cloud computing as a viable option for this task. It can be an important factor to solve the band-width dilemmas in real-time Cloud-based object recognition applications [158].

After investigating the importance of the peripheral vision for scene recognition, novel hierarchical topologies of object recognition that depends on the context of the object are introduced in the next

Chapter 5

Hierarchical Topologies for Context-Based Object Recognition

5.1 Introduction

In the previous chapters, a novel method for enhancing the performance of object recognition systems was developed. Then, to study the performance contributing factors of object and scene recognition, different regions of vision were investigated. In this chapter, topologies for context-based object recognition are introduced. They perform the object recognition process based on the context in which the object is located. The environment was detected before (or during) the process of recognising the objects. It is shown that the environment of the object can give a great deal of knowledge about the identity of the object, for instance, it is more likely to see a camel in a desert and a computer monitor in an office. In this chapter, a combination of deep and shallow models for object and scene recognition is used. Three novel topologies that provide a trade-off between classification accuracy and decision sensitivity are developed. This chapter proposes the following novel contributions to enhance the performance of the existing methods of object recognition:

- novel three topologies that provide a trade-off between the recognition accuracy and decision sensitivity;
- an enhanced object recognition performance outperforming a single GoogLeNet by 13%;
- a high level of confidence in the decision making, where the final decision not

only depends on the object feature for classification but also on the surrounding peripheral environments. This is essential in highly sensitive applications of object recognition such as driverless cars;

- a novel topology that recognises the object environment (whether indoor or outdoor scenes) without a specific classifier and only by inference;
- novel decision-making mechanisms that provide a no-decision state for the low confidence scenarios.

This chapter will first discuss the importance of understanding the environment in which the objects are located. It will briefly discuss the architecture of the used shallow and deep models for object recognition. It will then discuss the details for utilising the posterior probability of the classifiers. Furthermore, a brief discussion of the used image datasets will be provided. Then, the proposed topologies for object recognition will be explained in detail. The classification scenarios will also be provided. Moreover, a summary of all the active results with comparisons with other models is provided in the results section. Finally, a chapter summary is provided that discusses the main findings in this chapter of the thesis.

5.2 Shallow Models

In this chapter, the models that consist of five convolutional layers or less are considered shallow models, see Figure 5.1. Below is a brief description of the architecture of the shallow models used in the experiments of this chapter. In particular, the HMAX model, the En-HMAX model and AlexNet.

5.2.1 HMAX

The HMAX model [2–4] consists of four layers that comprise convolutional and pooling layers. The alternation of convolution and pooling has proven efficient to extract a high-level representation of objects. The HMAX model has attracted the attention of many researchers in the field of machine vision because of its good performance and abstract architecture.

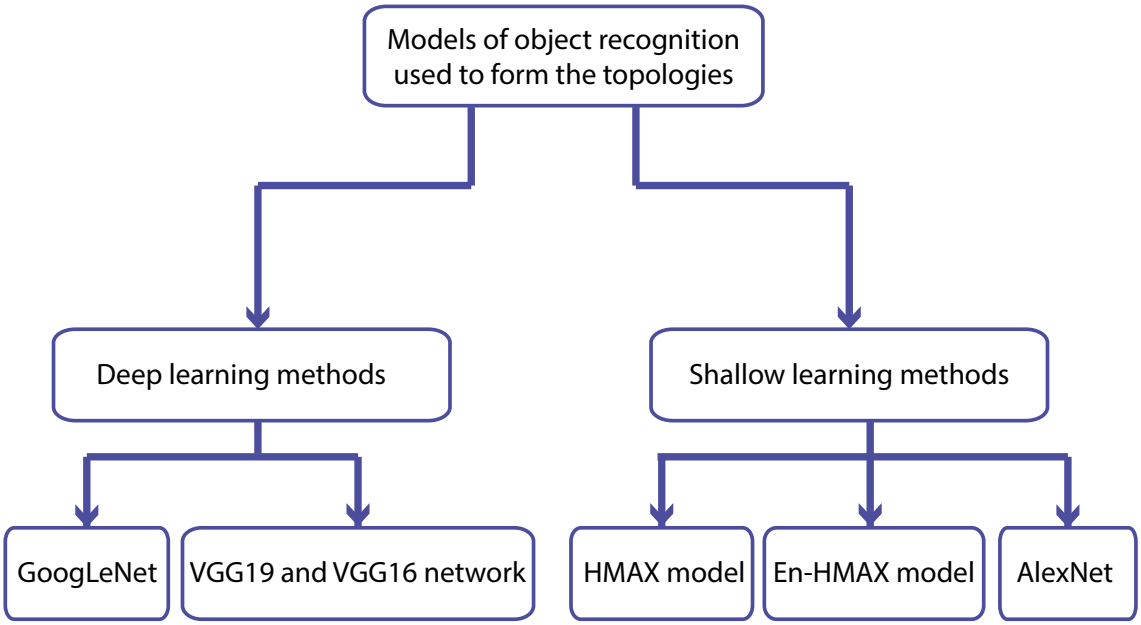


Figure 5.1: The taxonomy of object recognition models used to form the hierarchical topologies.

5.2.2 En-HMAX

The En-HMAX model [135, 159] has an increased number of layers. It comprised three convolutional layers and three pooling layers. Similar to the HMAX model, the En-HMAX model preserves the same architecture, however, using different techniques. It has outperformed the original HMAX model in a large margin on many datasets.

5.2.3 AlexNet

The AlexNet [47] is a convolutional neural network that consists of five convolutional layers, three pooling layers and two fully connected layers. It comprises 60 million parameters to be fine-tuned. It transforms objects in the input images into distinctive features. The AlexNet model operates in a similar fashion to the HMAX model. They share similar hierarchal structure and the same classic alternation of convolutional and pooling layers. Across shallow models, it achieved the highest performances on many datasets [160]. The success of AlexNet has attracted the attention of researchers of computer vision towards CNNs. Due to its simplicity and good performance, in this chapter, the AlexNet is considered as the default model for indoor versus outdoor categorisation task [47].

5.3 Deep Models

Object recognition models that consist of more than five convolutional layers are considered deep models. Below is a brief description of the architecture of well-known deep models used in the experiments of this chapter. In particular, VGG16 network, VGG 19 network and GoogLeNet.

5.3.1 VGG16 and VGG19

The VGGNet architecture introduced in [46] is designed to significantly increase the depth of the existing CNN architectures with 16 or 19 convolutional layers. The last three layers of both versions, i.e., VGG16 and VGG19, are the following layers:

- Fully connected layer: in this layer, the input data is multiplied by the weight matrix and then adds a bias vector. Neurons in a fully connected layer are connected to all activations in the previous layer;
- softmax layer: in this layer, a softmax function is used for classification purposes. It is considered as the multi-class generalisation of the logistic sigmoid function, also known as the normalised exponential layer;
- classification layer: in this layer, the output predicted label is generated. It is formed by cross-entropy loss function that defines the preexisted trained classes.

5.3.2 GoogLeNet

The GoogLeNet model [48], also known as the inception model, is significantly deeper than the previously explained CNN models. It comprises 57 convolution layers with 5 million parameters to fine-tune. A key feature in the design of GoogLeNet is applying the network in network architecture introduced in [161], in the form of inception modules. Inception module uses a set of parallel convolution layers with a MAX pooling stage along each module. A concatenating layer is used to concatenate the responses of each individual module. In this work, the used version of GoogLeNet comprises a total of 9 inception modules. A more detailed overview of GoogLeNet architecture can be found in [48].

5.4 Transfer Learning

Transfer learning is increasingly becoming a powerful tool in the field of machine learning [162]. It involves utilising the stored knowledge of a model acquired for solving a particular task and applying it to solve a different problem. For instance, the knowledge acquired while learning to distinguish between different types of trucks could be utilised to recognise different types of cars.

Fine-tuning a network with randomly initialized weights is extremely complicated and time-consuming task. Therefore, in this chapter, fine-tuned pre-trained networks on 1000 class images of ImageNet dataset [160], were used as a starting point to learn the new tasks of the experiments. The previous knowledge of the selected well-known models of CNN was utilised to extract features. These models were trained and fine-tuned to do a different task, which is recognising daily-life objects.

5.5 Posterior Probability

The posterior probability is the conditional probability that is computed after an occurrence of a relevant event. In the field of pattern recognition, the posterior probability indicates the uncertainty of assessing a particular class of images. The posterior probability is produced when a generative model makes a decision [163]. Higher posterior probabilities indicate higher confidence of the classifier's decision. Figure 5.2 shows an example of how an indoor classifier distributes posterior probabilities for a given input image. Usually, the maximum posterior probability is used to determine the class label. In this chapter, the maximum posterior probability was utilised to indicate the confidence in the classifier. A threshold for each classifier was set and accordingly, the classifiers made decisions based on their confidence. The threshold was set based on the average posterior probability of all the testing dataset.

5.6 Datasets

The image classes were collected from ImageNet dataset [160], Caltech 101 dataset [110] and Caltech 256 dataset [6]. These classes were categorised into two uncorre-

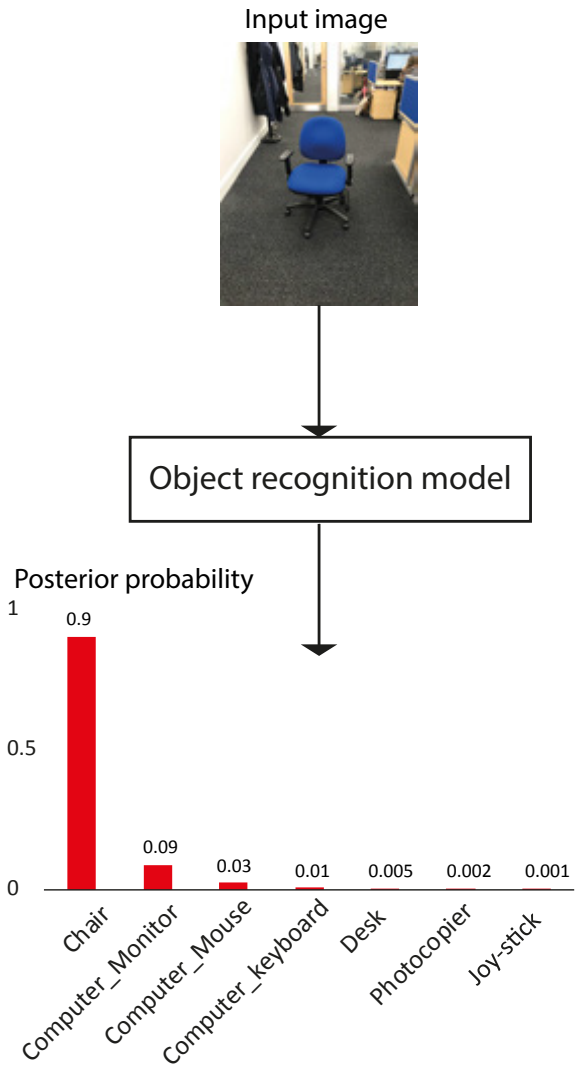


Figure 5.2: The distribution of posterior probabilities of an input image. It can be seen that in this example, the classifier is 90% confident that the object in this image is a chair.

lated set of images: outdoor and indoor. The outdoor image subset does not contain classes of the indoor image subset and the reverse is also true.

Figure 5.3 shows six examples of the dataset, reflecting the richness of the dataset in terms of the variety of objects and their backgrounds.

5.7 Classification

In this chapter, the classification settings are briefly explained. In this section, for all classification scenarios, the extracted features were classified using a linear support vector machine (SVM) [70]. In each of the experiments, 50% of the dataset was allocated for testing the classifier. In addition, to ensure that the classification scores were not biased by the random choice of training samples, the classification

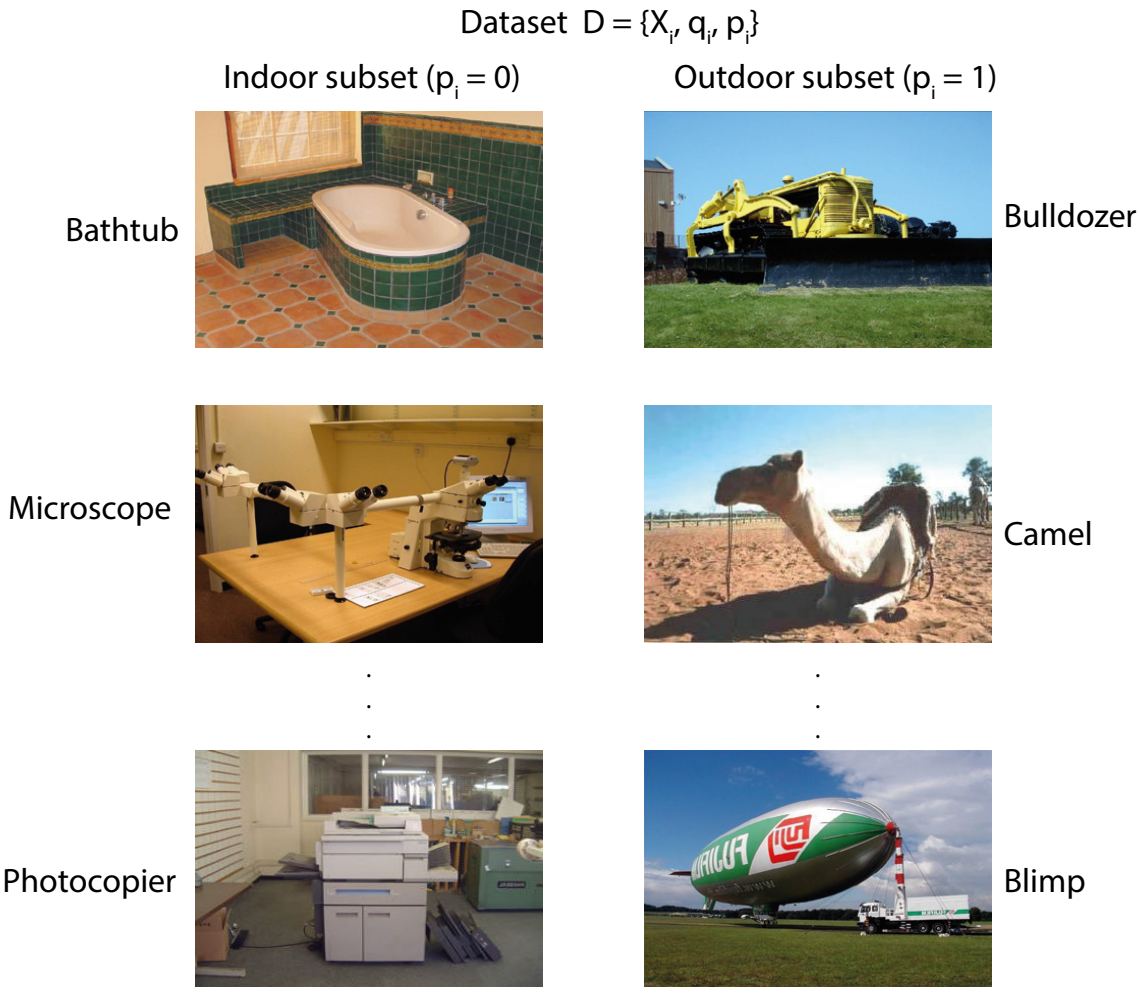


Figure 5.3: Selected indoor and outdoor images from our dataset.

was repeated for 20 runs where the random selection in each round is independent from the other. The average classification score and the standard deviation are reported.

5.8 Proposed Topologies

The hierarchical topologies developed in this section are designed to achieve an improved classification performance over the existing methods of object recognition. Additionally, providing higher confidence level and decision sensitivity. In this section, a detailed description of the proposed topologies is provided. The method and the architecture of each topology are explained. The designed topologies obtain the environment in which the object is found as an essential component of the recognition process. Furthermore, the designed topologies comprise a decision-making stage that can be tuned to increase the confidence or the decision sensitivity for the process of object recognition.

Topology-A and topology-B consist of three different models for object recognition. They comprise one shallow model for recognising the environment and two deep models for object recognition. Topology-C, however, consists of only two models for object recognition. The environment type, whether indoor or outdoor, in topology-C, is categorised by inference.

The architecture of topology-A was inspired by the human visual system, where scenes are rapidly categorised in a small time of 50ms which give a clear information about the identity of the objects within [164]. However, topology-B and topology-C are purely computational with less relevance to biology. Topology-B was designed to minimise the error chance in the first stage of topology-A, the scene recognition stage. The scene recognition stage was designed in-parallel to other stages of object recognition with a different mechanism in the decision-making stage. Topology-C was designed to minimise the number of models in topology-A and topology-B. Only two models for object recognition are used in topology-C for understanding the environment and for identifying objects. Finally, each of the below topologies have several advantages and disadvantages. The below subsections will discuss these in more details.

5.8.1 Topology-A

Figure 5.4 shows the basic structure of topology-A. In the used dataset $\mathbb{D} = \{\mathbf{X}_i, q_i, p_i\}_{i=1}^{N}$, each image $\mathbf{X}_i$ has class label q_i (for example: chair) and category label p_i (for example: indoor). The indoor category is denoted by using $p_i = 0$ and the outdoor category by using $p_i = 1$. For a given image, q_i* denotes the predicted class label and p_i* denotes the predicted category label. The confusion matrix of the indoor versus outdoor classifier $C_M = \{c_{ij}\}_{i,j=1}^{2}$ was used to calculate the ratio of the correctly classified images (see Fig.5.5). Using the total probability theorem, the overall accuracy in topology-A can be claculated as shown below:

$$\begin{aligned} Accuracy(\%) = \frac{100}{\sum_{i,j} \mathrm{c_{ij}}} \Big[& \mathrm{c}_{11}\, \mathbb{P}(q* = q \mid p* = p = 0) + \\ & + \mathrm{c}_{22}\, \mathbb{P}(q* = q \mid p* = p = 1) + \\ & + \mathrm{c}_{12}\, \mathbb{P}(q* = q \mid p* = 1, p = 0) + \\ & + \mathrm{c}_{21}\, \mathbb{P}(q* = q \mid p* = 0, p = 1) \Big] \end{aligned} \tag{5.1}$$

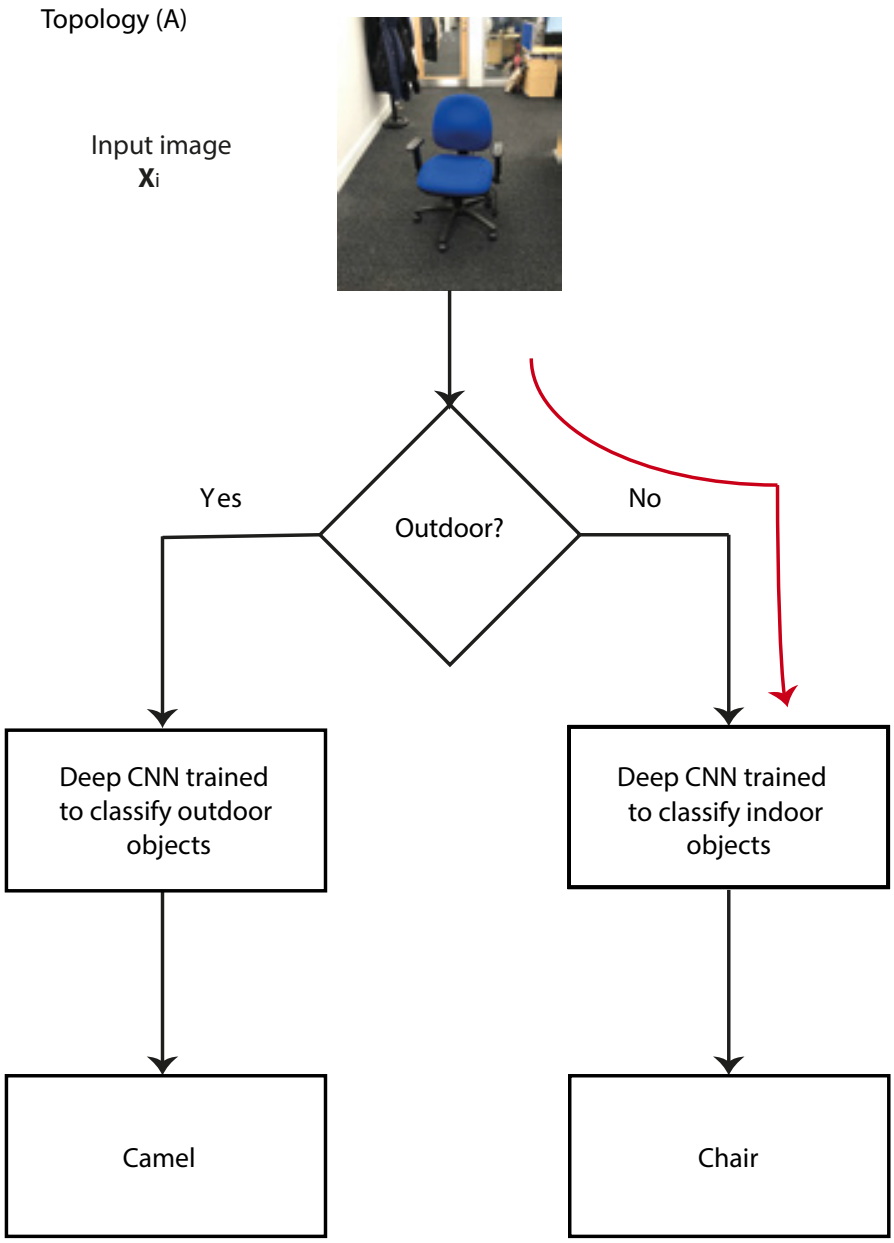


Figure 5.4: The structure of topology-A. The input image is first categorised (i.e., indoor and outdoor) then classified (i.e., chair, microscope).

5.8.2 Topology-B

In topology-B, shown in Figure 5.6, the three classifiers operate in parallel to identify an object in an input image. The object identity depends on the decision of all three classifiers. The three classifiers have an equal influence in making the final decision. Making an incorrect decision in any of the stages does not guarantee an incorrect class label in the final stage. The posterior probability is used to quantify the reliability of the classifiers. Classifiers with higher confidence level have more influence on making the final class label decision.

In the experiments performed in this chapter, the mean of the posterior proba-

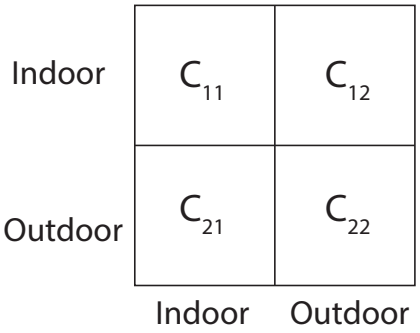


Figure 5.5: The confusion matrix of the indoor versus outdoor classifier. c_{11} and c_{22} represent images that were classified correctly. c_{12} and c_{21} represent images that were misclassified.

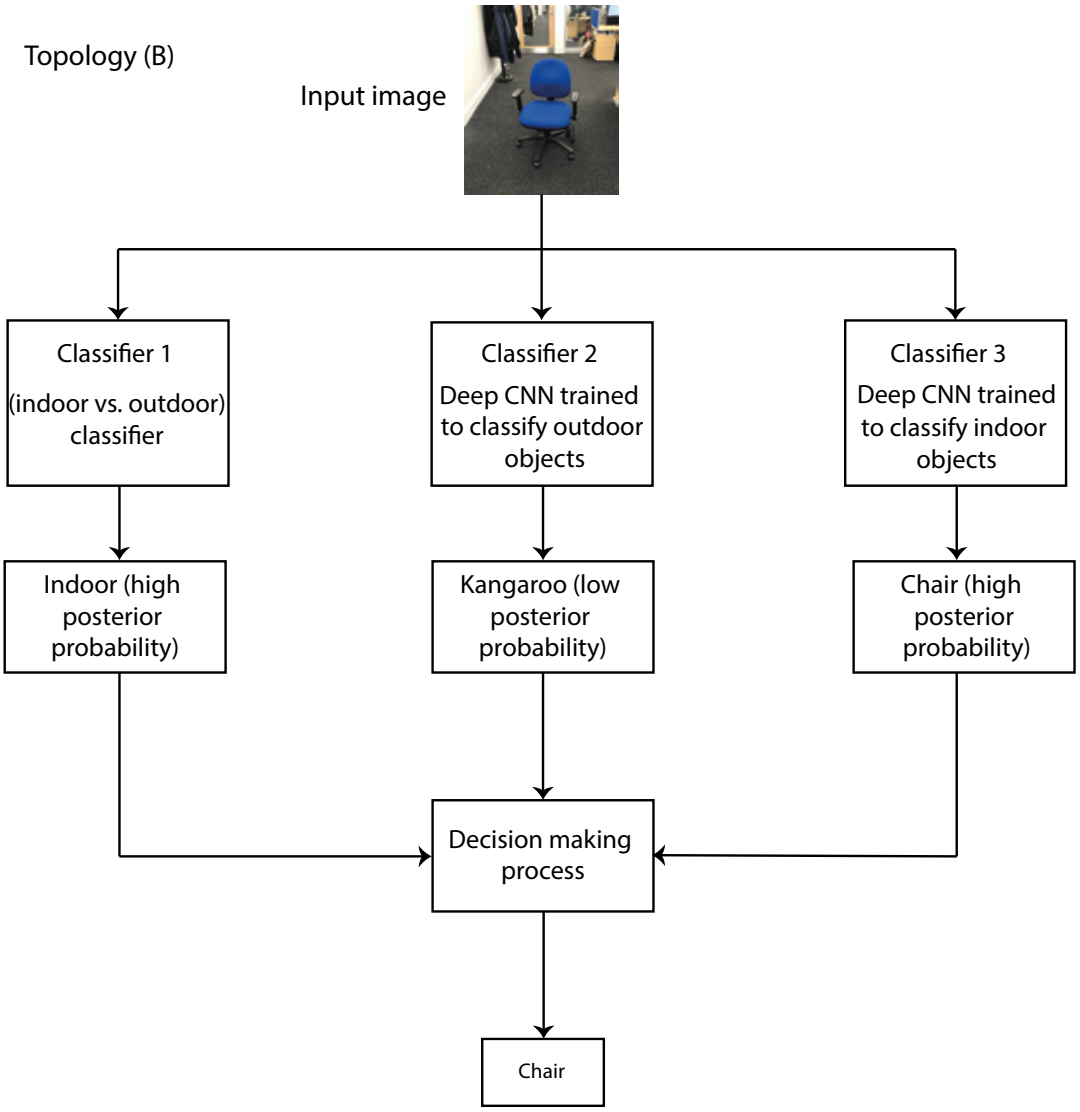


Figure 5.6: The structure of topology-B. In topology-B, the classifier that categorises indoor versus outdoor images operates in parallel with other classifiers.

bilities of the whole testing data $\mathbb{D}$ was set as a confidence threshold. However, an optimal confidence threshold can be tuned differently depending on the classification context. The final decision is based on the posterior probabilities of all three

Table 5.1: The decision-making process of topology-B. The table shows only 2 possible scenarios of the 16th possible combinations. In all other scenarios, a no-decision state will be produced. The ✓marker denotes higher confidence, X marker denotes lower confidence and d denotes the "do not care status".

		Confidence	
Indoor classifier (1)		✓	X
Outdoor classifier (2)		X	✓
Indoor versus	Indoor decision	✓	d
outdoor classifier	Outdoor decision	d	✓
Classifier selection		1	2

classifiers as shown in Table 5.1.

5.8.3 Topology-C

In this topology, shown in Figure 5.7, only two classifiers were used to predict the class label and the category label. Table 5.2 shows the scenarios in which this topology make the final decision.

In this chapter, the collected image dataset has two separate image subsets. The image classes of the indoor subset do not correlate with the image classes of the outdoor subset. This suggests that when an indoor classifier is used, classes from the outdoor subset tend to give lower posterior probabilities than classes from the indoor subset. Figure 5.8 shows an analysis of the average posterior probability for both the indoor classifier and the outdoor classifier. In this analysis, GoogLeNet was used to produce the figures. As expected, in both scenarios, i.e., indoor classifier and

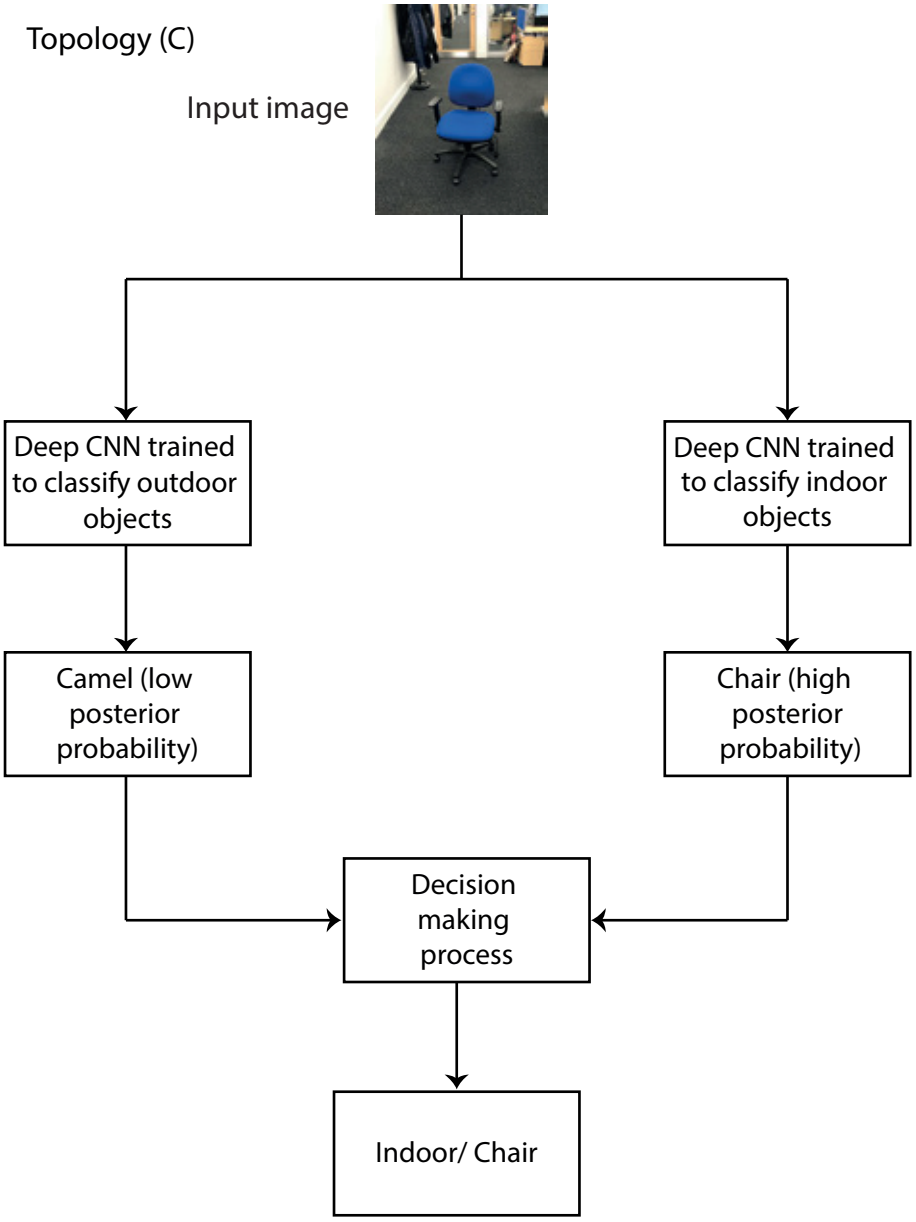


Figure 5.7: The structure of topology-C. In topology-C, no classifier is used to categorise the environment (indoor and outdoor), however, it is able to categorise the environment by inference.

outdoor classifier, testing a classifier with unseen images within the same training categories produced a significantly higher posterior probability than testing it with different image categories. For the indoor classifier, the Mann-Whitney U test, with a risk $\alpha = 0.05$, shows that the posterior probabilities for indoor test images (M= 87.6, SD = 18.9) were significantly higher than that of outdoor test images (M = 41.7, SD = 21.5); Z-score = 22.3, p-value < 0.05. Similarly, for the outdoor classifier, the above test shows that the posterior probabilities of the outdoor test images (M = 74.0, SD = 26.4) were significantly higher than that of indoor test images (M= 31.2, SD = 18.0); Z-score = 20.9, p-value < 0.05. The data above comprises unpaired

Table 5.2: The decision-making process of topology-c. The ✓marker denotes higher confidence and X marker denotes lower confidence

	Confidence			
Indoor classifier (1)	✓	X	✓	X
Outdoor classifier (2)	X	✓	✓	X
Classifier selection	1	2	No-decision	

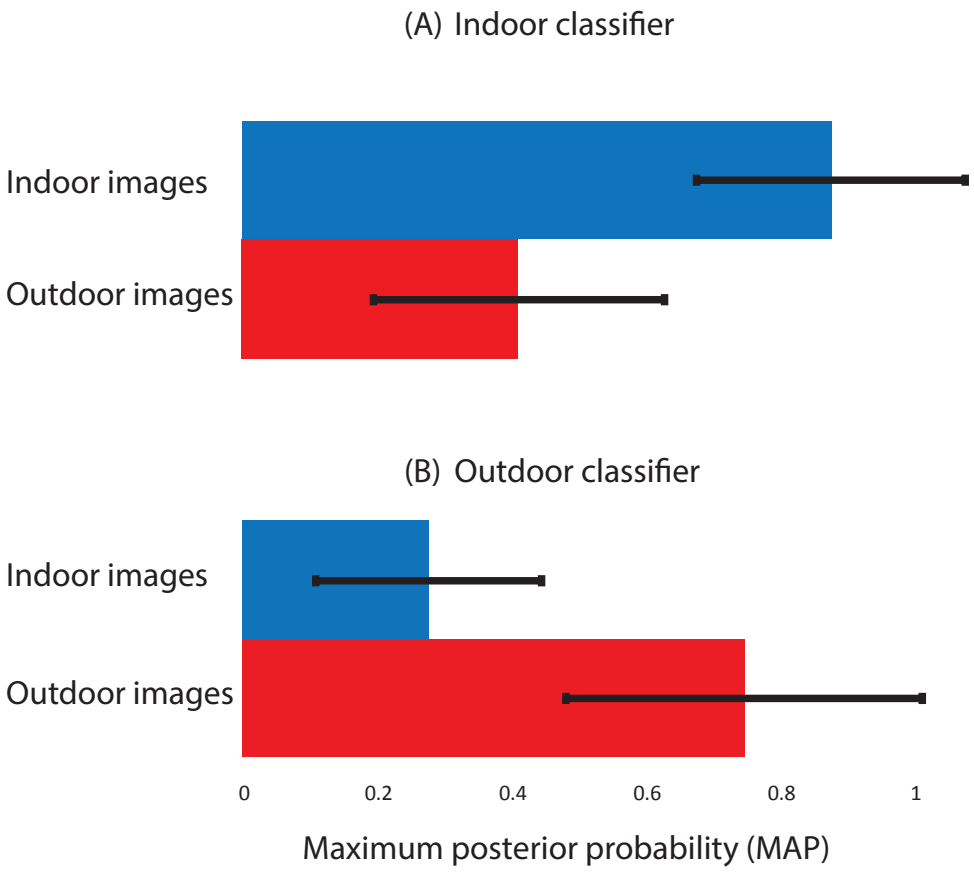


Figure 5.8: An example of the average posterior probability of the indoor and the outdoor classifiers using GoogLeNet. (A) Indoor classifier. (B) Outdoor classifier. This chart illustrates the decorrelation in the average posterior probability between the indoor classifier and the outdoor classifier of topology-C.

non-parametric samples. Therefore, we used Mann-Whitney U method to test for significance. Therefore, we hypothesised that the posterior probability can give a

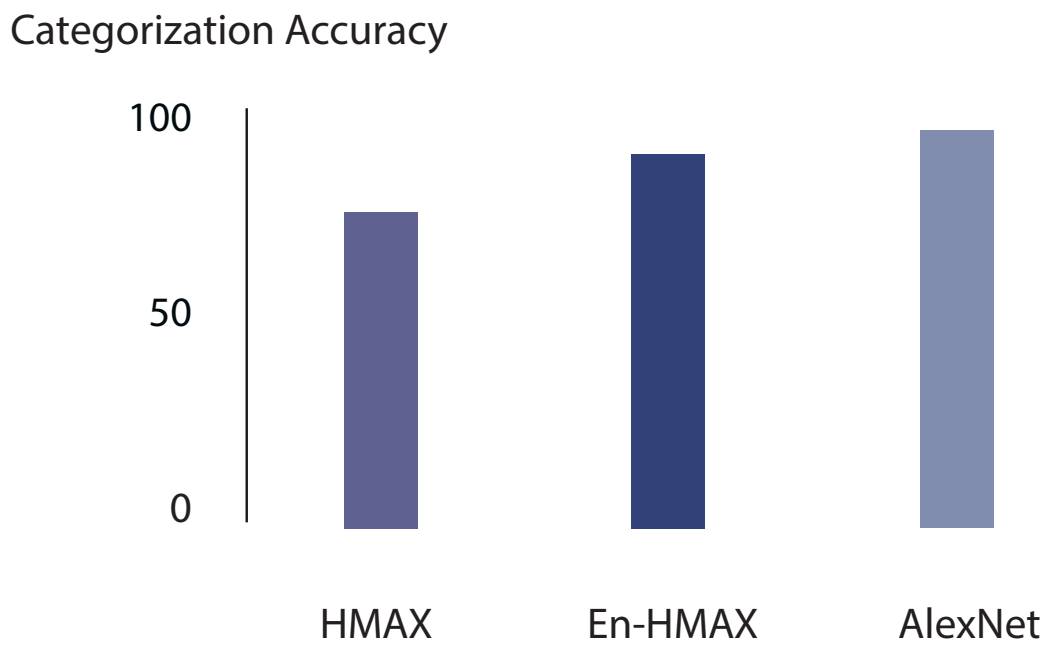


Figure 5.9: Results of categorising indoor and outdoor images.

notion of the image category, i.e., indoor versus outdoor.

5.9 Results

The below subsections display the results for the discussed topologies in the previous sections.

5.9.1 Indoor Versus Outdoor

Models of object recognition tend to produce higher performances in a binary classification scheme. The chance level in binary classification scenarios is 50%. In this chapter, shallow models were utilised for categorising indoor and outdoor scenes. Figure 5.9 shows a comparison in classification performance between these models. It can be noticed that AlexNet outperforms other shallow models for the categorisation task, with a high accuracy of 99.46%. The En-HMAX model achieves higher scores of 87.96%, however, it is still far less than the performance of AlexNet. This is due to the large size of the image data, in which the En-HMAX model cannot handle efficiently due to its abstract architecture. The same applies to the HMAX model, where 75.03% of classification accuracy is achieved. Therefore, AlexNet was elected as a default model with regard to all indoor versus outdoor categorisation schemes, i.e., topology-A and topology-B.

In topology-A, AlexNet spread the images to either the indoor classifier or the outdoor classifier. Although AlexNet has a very high classification performance, the few incorrect decisions it makes lead to failure in the output stage. This is due to the uncorrelated image data used in both classifiers. In another word, the indoor classifier knows nothing about the outdoor environment and the reverse is also true. Therefore, when an outdoor image passes the indoor classifier, an incorrect class label will be guaranteed.

In topology-B, the decision of AlexNet has less impact on the final class label due to the structure of the topology. An incorrect decision at any stage does not guarantee an incorrect class label. In topology-C, however, no shallow network is used to categorise the scene type. The scene type is inferred from the indoor and the outdoor classifier.

5.9.2 Classification Scores Using Topology-A

In Figure 5.10, AlexNet, VGG16, VGG19 and GoogLeNet were utilised as the main platforms to quantify the performance of topology-A. To compute the classification accuracy of the whole classification task, the above models were used individually. In particular, all the image dataset was used without segregating it into an indoor subset and an outdoor subset. This process was repeated for each of the above models separately. As a result, the classification accuracy of each of the above models was quantified for the comparison with topology-A. A similar process was performed for topology-B a topology-C.

Finally, topology-A scores were compared with the above scores. For completeness, the comparisons are only performed between a certain classification model and the topology that is formed within the same model, for instance, the VGG19 network results are compared with topology-A that is formed by only the VGG19 models.

For all used models, topology-A outperformed the original models. For example, in AlexNet, an increased classification performance of 7% is achieved. The difference is constantly decreased for deeper models.

This is particularly interesting because deeper models are capable of understanding large data. Therefore, using a bigger abject dataset is believed to increase the above differences dramatically.

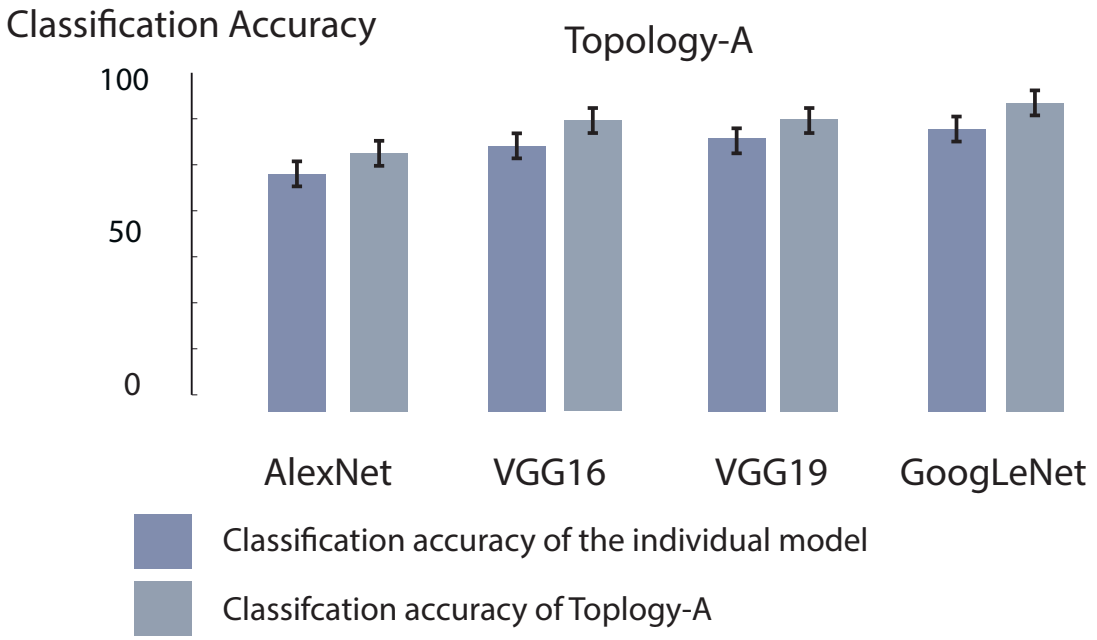


Figure 5.10: Results of topology-A. AlexNet is used as a default model for categorising indoor and outdoor images. The classification accuracies in the second-row represent the performance of below models to individually classify the whole dataset.

Topology-A has the following advantages:

1. Advanced performance over using a single network.

2. Only two models can operate to recognise each input image.

The disadvantages of topology-A can be summarised as the followings:

1. It involves three different classifiers that require more memory in terms of implementation.

2. An incorrect decision in the first stage guarantees an incorrect class label. The first stage (indoor versus outdoor classifier) has more power in making the final decision.

5.9.3 Classification Scores Using Topology-B

Figure 5.11 shows the classification scores of using topology-B. It also shows the percentages of the no-decision state. In line with topology-A, similar models were used in this experiment to form this topology. AlexNet was used to categorise the indoor and outdoor images in all scenarios. In the above calculations, the no-decision state is considered as a correct classification. It can be noticed that deeper models such as GoogLeNet and VGG19 do not outperform other models when using this

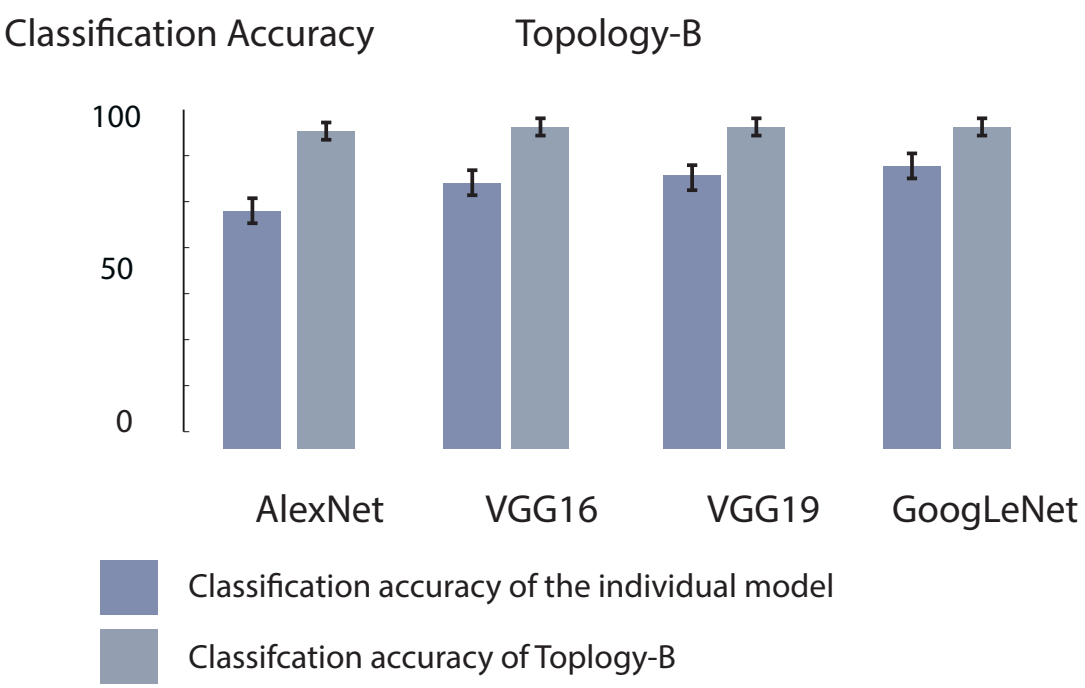


Figure 5.11: Results of topology-B. AlexNet is used as a default model for categorising indoor and outdoor images for all the below calculations.

topology. The performances are more balanced. However, the topology formed by VGG19 tends to make more decisions than other models. The decision-making conditions can be tuned using an optimised threshold. In this experiment, the mean posterior probability of all the testing images was used as a threshold of confidence.

Topology-B has the following advantages:

1. The decision-making process depends equally on all three classifiers.
2. It achieves the highest performance among the other topologies.
3. It is designed to make no decisions when a lower confidence level is obtained. The confidence threshold can be tuned depending on the allocated task. Applications with higher risks, for instance, autonomous cars, need higher confidence threshold. The "no-decision" state is an important measure in such applications.

The disadvantages of topology-B can be summarised as the followings:

1. It requires more memory in terms of implementation because of the three classifiers in its architecture.
2. It is more computationally expensive than the other topologies because it needs all three classifiers to operate simultaneously.

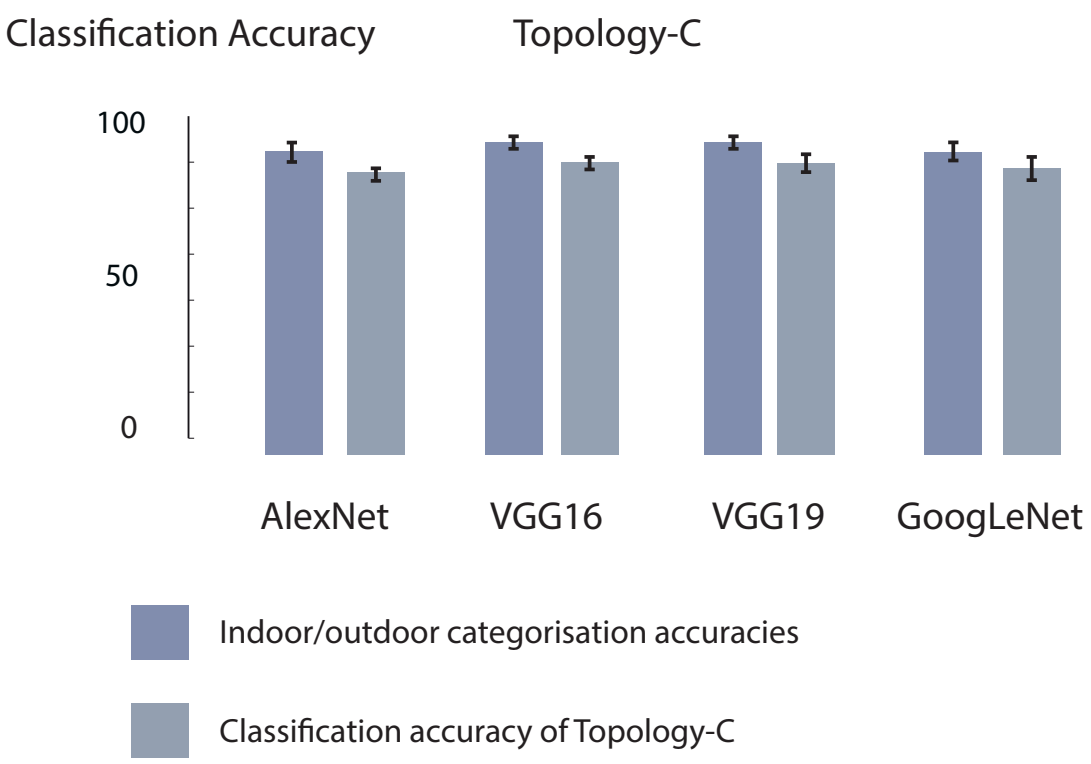


Figure 5.12: The results of topology-C

5.9.4 Classification Scores Using Topology-C

In topology-C, the objects are classified using only two classifiers as shown in Figure 5.6. Similar to topology-A and topology-B, the same previously explained models were used to form topology-C. Furthermore, the classification scores were reported in a similar fashion. Unlike topology-B, there was no allocated classifier for categorising the indoor and the outdoor environments. Instead, the category label was inferred throughout the process of recognising an object. Figure 5.12 shows the categorisation and classification scores of topology-C. A high categorisation accuracy of 95% was achieved using VGG19. This is particularly interesting because this score is achieved without using a specific classifier for the task. In this topology, the percentages of the no-decision state are less than that of topology-B. However, the classification accuracies are slightly decreased. Interestingly, VGG19 performs slightly better than other models using this topology.

Topology-C has the following advantages:

1. It involves only two classifiers for the recognition process.

2. It infers the category label without using a specific classifier, i.e., indoor versus outdoor classifier.

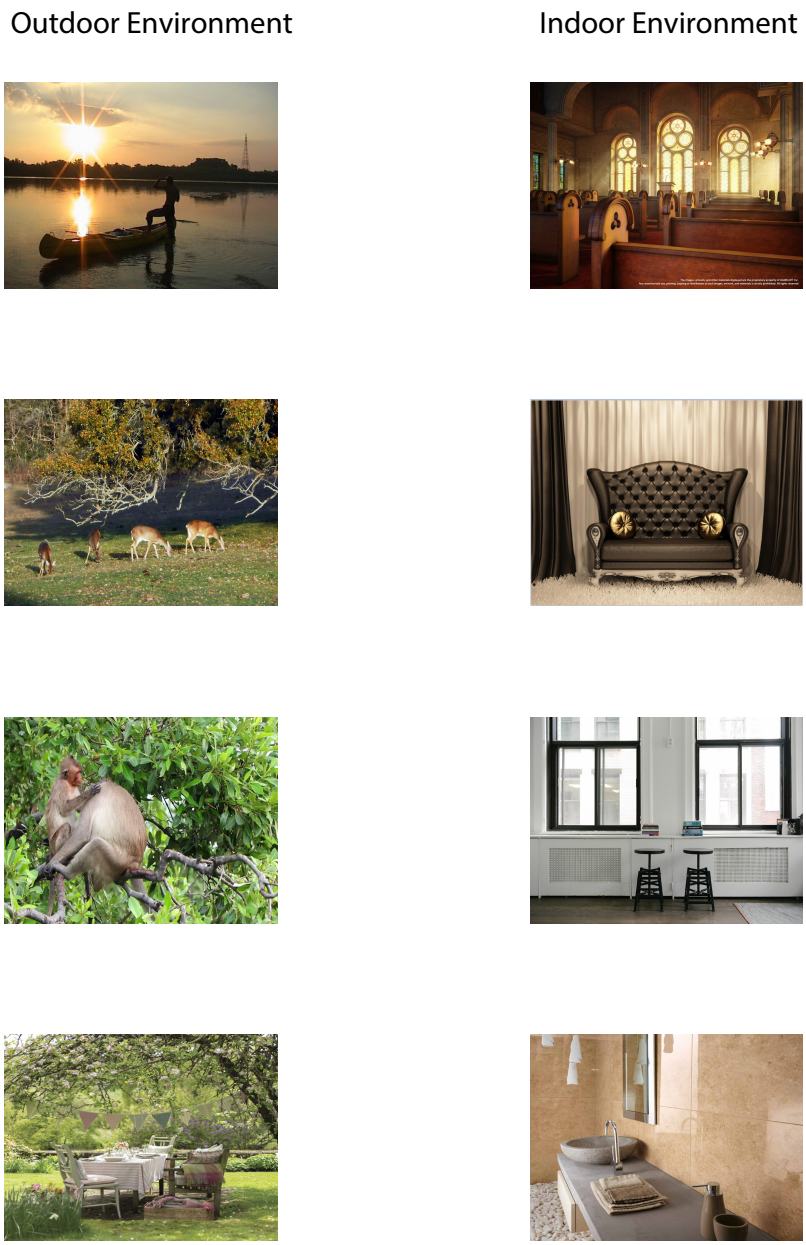


Figure 5.13: Examples of the image dataset used for the indoor and outdoor environment for real-time implementation.

3. It makes no decision when a lower confidence level is obtained.

The disadvantages of topology-C can be summarised as the followings:

1. It provides reduced performance comparing to the other topologies due to the decreased number of the classifiers in its architecture.

2. It shows lower decision frequency than other topologies, due to the limited number of input parameters in the decision-making stage.

Figure 5.14: Examples of the real-time implementation of the indoor and outdoor classifier using AlexNet. This experiment has taken place in the research lab at Newcastle University. The outdoor scene is the view from the window of the office.

5.10 Real-time Implementation

Models of object recognition comprise many layers of convolutions. Each layer consists of many filters. Using an advanced number of layers can help to extract high-level features that provide models with invariances. Accurate models of object recognition require huge computational resources. Training and fine-tuning these models consume a tremendous amount of time. Although training object recognition models are computationally expensive, their implementation has been significantly reduced due to the introduction of transfer learning. This is done using the following two steps:

1. Using an optimised fine-tuned pre-trained network as a feature extractor.

2. Training a new classifier to learn the newly produced features.

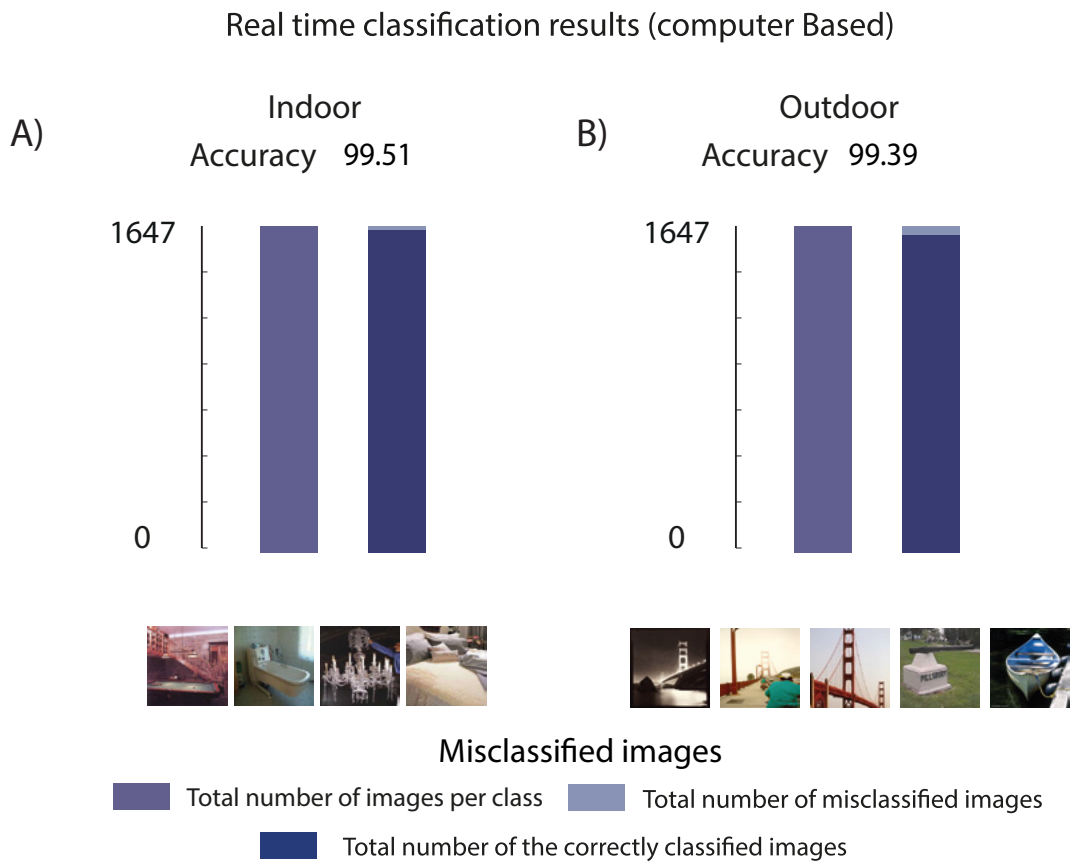


Figure 5.15: Computer-based results of the real-time experiment of the indoor and outdoor classifier.

Using the above procedure of transfer learning, the real-time implementation of the indoor and outdoor scheme was performed. More details about the implementation is provided below.

For the real-time implementation experiment of the indoor and outdoor classifier, an image dataset was collected to train a pre-trained AlexNet model to learn both schemes. The indoor image dataset consists of 2202 images, while the outdoor image dataset consists of 2167 images. The images of the indoor dataset contain different outdoor backgrounds. Furthermore, it comprises many different objects that can be usually seen in an outdoor environment, for instance, wild animals such as bears, airplanes and boats. Similarly, the images of the indoor dataset contain indoor backgrounds and objects that are likely to be seen in an indoor context, for instance, a computer keyboard, a computer mouse and microscope. Figure 5.13 shows examples of the indoor and the outdoor images used in this experiment.

A pre-trained (on 1000 objects from the ImageNet dataset) AlexNet network was trained using the above dataset. The experiments were implemented in Matlab on a dual-core i5 processor (3.4 GHz) PC with 32G RAM without GPU acceleration. The

experiments were done using an average quality computer web-cam in the research office. The results have shown a classification accuracy of 99.46% to classify new environments. Figure 5.14 shows examples of the real-time implementation for this experiment.

Figure 5.15 shows the individual accuracies for each class of images. It can be noticed that the performances are balanced for each subset. Only four images were misclassified from the indoor subset. Additionally, five images were misclassified from the outdoor subset. The accuracies were 99.51% and 99.39% for the indoor classifier and the outdoor classifier, respectively.

5.11 Chapter Summary

This chapter has developed three topologies for object recognition. The introduced topologies provide a trade-off between three essential elements for image classification: classification accuracy, decision sensitivity and computational complexity. This is important in applications with high risks such as the driver-less cars. A no-decision state is an important measure for the least confident scenarios. Furthermore, the decision sensitivity can be tuned depending on the used type of the application. In topology-A, two models can operate to recognise an object for each input image. The categorisation stage filters the input images to either the indoor classifier or the outdoor classifier. This topology is less complex than other topologies. However, an incorrect decision at the first stage can cause an incorrect image class label. In topology-B, the problems of topology-A were tackled by electing the decision via all classifiers simultaneously. All three classifiers operate at the same time and a voting procedure decides the final decision. This topology is computationally complex, as it needs three classifiers to operate simultaneously for each input image. However, it provides higher classification accuracies, in addition to, providing the capability of tuning its decision sensitivity. Topology-C provides the advantages of topology-A and topology-B. The voting includes only two classifiers to infer the image category and class. This topology also offers to control the sensitivity of the decision making. Results show that with the proposed topologies, the performance of GoogLeNet can be improved by 13%.

The evaluation process was performed using Caltech 101 dataset, Caltech 256 dataset and ImageNet dataset. This chapter extends the knowledge regarding the techniques that could shape the object recognition process in the real world. In particular, application specific scenarios and will serve as a base for future studies in the field. The next chapter will draw the conclusion of the entire work on object recognition and propose future works which will push the development of object recognition system design even further.

Chapter 6

Conclusions and Future Work

6.1 Summary and Conclusion

In this chapter, a brief summary of this thesis and a review of the main contributions are provided. This thesis has provided novel methods for developing object recognition technologies. In particular, a comprehensive survey on the improvement of object recognition has been presented in Chapter 2. The research work was initiated with the intention of examining a model to recognise three-dimensional objects using an ample number of two-dimensional images. More specifically, it was designed from a feature-based system that extracts invariant features from two-dimensional images that represents the real three-dimensional world. The recent hierarchical object recognition methods have been listed and analysed. In addition, an overview of recent publicly available datasets of object recognition has been listed with their merits and characteristics.

Firstly, the objective was to select the feature-based approach to conduct the research. Out of the available approaches, in this thesis, two main approaches were used for object recognition: hierarchical feed-forward approaches and deep learning approaches. Therefore, an examination was made to assess well-known models of object recognition, such as the HMAX model, sparse HMAX model, AlexNet, VGG net and GoogLeNet. As a result, a decision was made to develop on the above models and use them in different stages over the course of this research.

In Chapter 3, in order to further enhance the recognition accuracy for both objects and scenes, the En-HMAX model was proposed. The En-HMAX model provides sparsity-grouping trade-off, such that informative features of objects and

scenes are preserved for classification. The En-HMAX model was compared with the original HMAX model and other hierarchical models for object recognition. The model sparsity was quantified. The En-HMAX model provides two essential elements for image classification: selectivity and invariance. The main reason for using an elastic-net regulariser for the HMAX model was to encourage the grouping effect when the atoms in the dictionary are highly correlated. Results show that the En-HMAX model outperforms the original HMAX model (by ∼40%) as well as the two special cases of the En-HMAX model, i.e., the LASSO- and Ridge-HMAX models, by ∼19% and ∼9%, respectively. Furthermore, in Chapter 3, the lateral connections experiment was presented. Features with different degree of complexity were investigated for recognition. The performances of different combinations of features were quantified and reported.

In Chapter 4, the developed En-HMAX model (in Chapter 3) was tested against occlusions. All the occlusions generated in this chapter have a single location, shape and pixel value. As a result, the dataset comprised occluded dataset that is highly overlapped. Using an elastic-net dictionary learning in HMAX model scheme has encouraged the grouping effect when atoms in the dictionary are highly correlated. As a result, the En-HMAX model showed outstanding performance when encountering such a highly correlated data, such as that of class-A occlusions. The experimental results show that hierarchical structures such as the En-HMAX model allow for substantial robustness in recognizing objects under partial occlusion. The En-HMAX model provides two elements essential for image classification: selectivity and invariance.

In Chapter 5, The recognition behaviour of the En-HMAX model that mimics the basic structure of the ventral visual stream was further investigated. As a result, a study that highlights the contribution of the peripheral versus central vision to scene and object images was conducted. The En-HMAX model was tested with object and scene image datasets with varying occlusion conditions to reaffirm that peripheral image content, that is beyond 5° eccentricity, is more efficient in recognising the gist of a scene than central image content. In addition, this study showed that introducing foveation increases the object classification performance of the En-HMAX model at 1° scotoma. However, it had no impact on recognising the gist of the scene in the absence of parafoveal vision.

The advantage of central vision in object recognition is mainly explained by the

fact that objects are generally located in the centre of the images. This indicates that the Eh-HMAX model recognizes the objects within the images and not their backgrounds. Also, when normalizing performances, the decline in object recognition was 13% faster than the decline in scene recognition, especially when occlusions block parafoveal $[3^\circ - 7^\circ]$ section of the image. This observation corroborated the importance of parafoveal vision for object recognition [93].

The prevailing advantage of the peripheral vision in scene recognition can be explained by the characteristics of scenes. The formative information of the scene is spread and compressed at the periphery of the images. Therefore, the En-HMAX model intrinsically captures the usefulness of the peripheral image content when recognising scenes. Interestingly, results suggested that outdoor man-made scene classes were less dependant on the peripheral image content. With a 10.8° window, these scene sub-types scored relatively higher performance. It was speculated that the reason for this observation is that scene recognition depend on local features within each type of scene [33]. Examples of local features are the presence of cars, pedestrians, and bicyclists in a street in outdoor-man made scenes [14]. Therefore, the En-HMAX model can extract local features across man-made scene images without particularly relying on the peripheral vision. Further data and research are required to test this hypothesis.

A further investigation was made to the relative importance of each region of vision for both datasets, that is, peripheral image content for scene dataset and central image content for object dataset. Blocking the less relevant image content produced the same performance pattern in both scenarios. A key outcome of this experiment may be this finding that by selectively blocking image regions, the computational requirement of image classification can be reduced which is of significant importance in real-time robotic vision applications.

The state of art models for object recognition may be too computationally expensive to run on a computer with modest specifications. Three possibilities to overcome this problem are: 1) local processing, 2) Cloud processing and 3) a combination of the two. Cloud processing remains an important tool especially for devices with low processing capability. Most future systems may use a combination of local and Cloud processing, given the increasing power of mobile graphics units and mobile connectivity. However, transferring all image data to a remote Cloud may be unrealistic, due to the band-width related issues [158]. This limitation may necessitate

data is either reduced or compressed locally before transmission ideally without any performance degradation. The results in Chapter 5 showed that foveation can be an appropriate candidate for local data compression. Another important finding in the study made in Chapter 5 was that the maximum classification performance, equal to when the whole image is available, can be achieved with only half of the input image content. This observation offers significant bandwidth saving and data reduction and can be an important factor to solve the band-width dilemmas in real-time Cloud-based object recognition applications [158].

In Chapter 6, three topologies for object recognition were developed to further optimise the previously discussed platforms for object recognition. The recognition process in the developed topologies depends heavily on the environment in which the object is found. The topologies presented in Chapter 6 provides three essential elements for image classification: classification accuracy, decision sensitivity and computational complexity. In topology-A, two models can operate to recognise objects for each input image. The categorisation stage filters the input images to either the indoor classifier or the outdoor classifier. This topology is less complex than other topologies. However, an incorrect decision at the first stage may guarantee an incorrect image class label. In topology-B, the problems of topology-A were tackled by electing the decision via all classifiers. All three classifiers operate simultaneously and a voting stage decides the final decision. This topology is computationally complex, as it needs three classifiers to operate simultaneously for each input image. However, it provides higher classification accuracies, in addition to, providing the capability of tuning its decision sensitivity. Topology-C provides the advantages of both topology-A and topology-B. The voting includes only two classifiers to infer the image category and class label. This topology also offers to control the sensitivity of the decision making. Results show that with the proposed topologies, the performance of GoogLeNet can be improved by 13%.

6.2 Future Work

The aspirations of this thesis involve presenting new research horizons in the future of object recognition. However, there are some limitations which need to be considered as future work in order to improve the performance of the biologically inspired models of the visual cortex. Key issues are

1. Most of the utilised datasets in recent object recognition systems are intended to enable the developed object recognition models to generalise to new settings, for instance, different object backgrounds, illumination, object orientation, pose, position and scale. However, there are limited resources to image dataset that target the top down processing and attention that the human brain can solve. In this thesis, the klab dataset [141] was utilised for this purpose. However, this dataset is limited in terms of variety of patterns and number of images.

2. Although the developed techniques in this thesis for object recognition were shown to perform efficiently, the processing time should reduce, especially when the image resolution of modern cameras is dramatically increasing.

3. Although the developed models in this thesis are biologically inspired, they do not provide techniques for attention and top-down processing that exist in the human visual system. Introducing attention in object recognition models can provide the following advantages:

   (a) it can provide the capabilities of learning new sets of objects within a single image and identifying the learned objects in different environments;

   (b) it equips models with the capability of recognition in a highly cluttered environment.

4. Due to the time requirement of learning a dictionary to generate optimised filters, in the developed models for object recognition in this thesis, the dictionary learning process takes place off-line. However, in order to build a general model for object recognition for a new environment, the dictionary learning process must take place online.

5. Employ recent mobile technologies to perform object recognition off-Cloud. This requires the developed models to have a more abstract architecture. Deep learning models can only be trained using Cloud computing. Providing a more abstract model for object recognition can help to achieve that aim. At the same time, implementing a highly optimised deep learning method on a parallel chip such as field programmable gate arrays (FPGAs) can help to solve that problem.

6. Investing in different regions of vision for different recognition tasks, for instance, high focus on the peripheral image content for tasks that involve scene recognition. Similarly, focus the model processing on the central image content for applications that involve object-based recognition.

References

[1] S. K. Vashist, O. Mudanyali, E. M. Schneider, R. Zengerle, and A. Ozcan, "Cellphone-based devices for bioanalytical sciences," *Analytical and Bioanalytical Chemistry*, vol. 406, no. 14, pp. 3263–3277, 2014.

[2] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, 2007.

[3] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[4] X. Hu, J. Zhang, J. Li, and B. Zhang, "Sparsity-regularized hmax for visual recognition," *PloS One*, vol. 9, no. 1, pp. 3263–3277, 2014.

[5] D. K. Prasad, "Survey of the problem of object detection in real images," *International Journal of Image Processing (IJIP)*, vol. 6, no. 6, p. 441, 2012.

[6] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[7] G. Ghazaei, A. Alameer, P. Degenaar, G. Morgan, and K. Nazarpour, "Deep learning-based artificial vision for grasp classification in myoelectric hands," *Journal of Neural Engineering*, vol. 14, no. 3, pp. 3263–3277, 2017.

[8] "Number of mobile phone users worldwide from 2013 to 2019," https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide, Accessed: 2018-02-21.

[9] "Number of mobile phone users in india from 2013 to 2019," https://www.statista.com/statistics/274658/forecast-of-mobile-phone-users-in-india/, Accessed: 2018-02-21.

[10] "Google goggles," https://play.google.com/store/apps/details?id=com.google.android.apps.unveil&hl=en_GB, Accessed: 2018-02-21.

[11] "Camfind," https://camfindapp.com/, Accessed: 2018-02-21.

[12] "Toyota suspends us driverless car tests after fatal uber accident, howpublished = http://www.bbc.co.uk/news/business-43478158, note = Accessed: 2018-05-16."

[13] T. Serre, M. Riesenhuber, J. Louie, and T. Poggio, "On the role of object-specific features for real world object recognition in biological vision," in *Biologically Motivated Computer Vision*, vol. 2525, 2002, pp. 387–397.

[14] A. Teichman and S. Thrun, "Practical object recognition in autonomous driving and beyond," in *Advanced Robotics and its Social Impacts (ARSO), 2011 IEEE Workshop on*, 2011, pp. 35–38.

[15] M. Hu, Z. Wei, M. Shao, and G. Zhang, "3-D object recognition via aspect graph aware 3-D object representation," *IEEE Sig. Process. Lett.*, vol. 22, no. 12, pp. 2359–2363, 2015.

[16] F. Fang and S. He, "Viewer-centered object representation in the human visual system revealed by viewpoint aftereffects," *Neuron*, vol. 45, no. 5, pp. 793–800, 2005.

[17] P. A. McMullen and M. J. Farah, "Viewer-centered and object-centered representations in the recognition of naturalistic line drawings," *Psychological Science*, vol. 2, no. 4, pp. 275–278, 1991.

[18] J. Driver and P. W. Halligan, "Can visual neglect operate in object-centred co-ordinates? an affirmative single-case study," *Cognitive Neuropsychology*, vol. 8, no. 6, pp. 475–496, 1991.

[19] V. Zografos, "Pose-invariant, model-based object recognition, using linear combination of views and bayesian statistics," Ph.D. dissertation, UCL (University College London), 2009.

[20] A. Guéziec and R. Hummel, "Exploiting triangulated surface extraction using tetrahedral decomposition," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 4, pp. 328–342, 1995.

[21] D. Meagher, "Geometric modeling using octree encoding," *Computer Graphics and Image Processing*, vol. 19, no. 2, pp. 129–147, 1982.

[22] D. Jang, K. Kim, and J. Jung, "Voxel-based virtual multi-axis machining," *The International Journal of Advanced Manufacturing Technology*, vol. 16, no. 10, pp. 709–713, 2000.

[23] H. H. Bülthoff and S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," *Proceedings of the National Academy of Sciences*, vol. 89, no. 1, pp. 60–64, 1992.

[24] M. J. Tarr, P. Williams, W. G. Hayward, and I. Gauthier, "Three-dimensional object recognition is viewpoint dependent," *Nature Neuroscience*, vol. 1, no. 4, pp. 229–289, 1998.

[25] M. J. Tarr and H. H. Bülthoff, "Image-based object recognition in man, monkey and machine," *Cognition*, vol. 67, no. 1-2, pp. 1–20, 1998.

[26] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition.* IEEE, 2000, pp. 300–305.

[27] A. Srivastava, X. Liu, and U. Grenander, "Universal analytical forms for modeling image probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1200–1214, 2002.

[28] J. Sullivan, A. Blake, M. Isard, and J. MacCormick, "Bayesian object localisation in images," *International Journal of Computer Vision*, vol. 44, no. 2, pp. 111–135, 2001.

[29] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, no. 7045, 1999, pp. 1150–1157.

[30] S. Arslan, A. Saçan, I. H. Toroslu, E. Acar *et al.*, "Comparison of feature-based and image registration-based retrieval of image data using multidimensional data access methods," *Data & Knowledge Engineering*, vol. 86, pp. 124–145, 2013.

[31] K. Ohba and K. Ikeuchi, "Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 9, pp. 1043–1047, 1997.

[32] C.-Y. Huang, O. I. Camps, and T. Kanungo, "Object recognition using appearance-based parts and relations," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 5, no. 7045. IEEE, 1997, pp. 877–883.

[33] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[34] D. P. Huttenlocher and S. Ullman, "Recognizing solid objects by alignment with an image," *International Journal of Computer Vision*, vol. 5, no. 2, pp. 195–212, 1990.

[35] W. Grimson and T. Lozano-Perez, "Model-based recognition and localization from tactile data," in *Proceedings. 1984 IEEE International Conference on Robotics and Automation.*, vol. 1. IEEE, 1984, pp. 248–255.

[36] O. D. Faugeras and M. Hebert, "A 3-d recognition and positioning algorithm using geometrical matching between primitive surfaces," in *Proceedings of the Eighth international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc., 1983, pp. 996–1002.

[37] A. Califano and R. Mohan, "Multidimensional indexing for recognizing visual shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 4, pp. 373–392, 1994.

[38] Y. Lamdan, J. T. Schwartz, and H. J. Wolfson, "Object recognition by affine invariant matching," in *Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1988, pp. 335–344.

[39] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, 2002.

[40] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comp. Vis. Imag. Underst.*, vol. 110, no. 3, pp. 346–359, 2008.

[41] C. Harris and M. Stephens, "A combined corner and edge detector," in *in Proc. 4th Alvey Vision Conference*, vol. 15, no. 7045, 1988, pp. 147–152.

[42] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, no. 7045. Ieee, 2005, pp. 994–1000.

[43] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, 2001.

[44] E. N. Mortensen, H. Deng, and L. Shapiro, "A sift descriptor with global context," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 184–190.

[45] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105.

[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[49] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of The 22nd ACM International Conference on Multimedia*, no. 7045, 2014, pp. 675–678.

[50] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *J. Neurophysiol.*, vol. 28, no. 2, pp. 229–289, 1965.

[51] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," *Pattern Recognition*, vol. 15, no. 6, pp. 455–469, 1982.

[52] S. Dura-Bernal, T. Wennekers, and S. L. Denham, "Top-down feedback in an HMAX-like cortical model of object perception based on hierarchical Bayesian networks and belief propagation," *PLoS ONE*, vol. 7, no. 11, pp. 229–289, 2012.

[53] L. Zhang and W. Lin, *Selective visual attention: computational models and applications*. John Wiley & Sons, 2013.

[54] C. Theriault, N. Thome, and M. Cord, "Extended coding and pooling in the hmax model," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 764–777, 2013.

[55] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[56] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[57] J. W. Bacus, "A whitening transformation for two-color blood cell images," *Pattern Recognition*, vol. 8, no. 1, pp. 53–60, 1976.

[58] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 22, no. 2, pp. 1–6, 2018.

[59] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[60] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[61] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[62] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2691–2698.

[63] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 689–696.

[64] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 1033–1040.

[65] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 543–550.

[66] B. Shen, B.-D. Liu, and Q. Wang, "Elastic net regularized dictionary learning for image classification," *Multimedia Tools App.*, pp. 1–14, 2014.

[67] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Stat. Soc. Series B*, vol. 67, no. 1, pp. 301–320, 2005.

[68] E. Raninen and E. Ollila, "Scaled sparse linear regression with the elastic net," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4336–4340.

[69] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal Stat. Soc. Series B*, vol. 58, no. 1, pp. 267–288, 1994.

[70] V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[71] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fmri data," *NeuroImage*, vol. 26, no. 2, pp. 317–329, 2005.

[72] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[73] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *IEEE computer society conference on Computer vision and pattern recognition*. IEEE, 1997, pp. 130–136.

[74] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[75] M. Pontil and A. Verri, "Support vector machines for 3d object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 637–646, 1998.

[76] J. R. Parker, *Algorithms for Image Processing and Computer Vision*. John Wiley & Sons, 2010.

[77] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comp. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[78] L. Li, S. Jiang, and Q. Huang, "Learning hierarchical semantic description via mixed-norm regularization for image understanding," *IEEE Transactions on Multimedia*, vol. 14, no. 5, pp. 1401–1413, 2012.

[79] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[80] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 580–587.

[81] D. Zoccolan, N. Oertelt, J. J. DiCarlo, and D. D. Cox, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Proc. Nat. Aca. Sci.*, vol. 108, no. 21, pp. 8748–8753, 2009.

[82] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain," *Nature*, vol. 435, no. 7045, pp. 1102–1107, 2005.

[83] A. Maashri, M. DeBole, M. Cotter, N. Chandramoorthy, Y. Xiao, V. Narayanan, and C. Chakrabarti, "Accelerating neuromorphic vision algorithms for recognition," in *Proc. 49th IEEE Design Autom. Conf.*, no. 7045, 2012, pp. 579–584.

[84] G. Orchard, J. G. Martin, R. J. Vogelstein, and R. Etienne-Cummings, "Fast neuromimetic object recognition using FPGA outperforms GPU implementations," *IEEE Trans. Neural Net. Learn. Sys.*, vol. 24, no. 8, pp. 1239–1252, 2013.

[85] E. T. Carlson, R. J. Rasquinha, K. Zhang, and C. E. Connor, "A sparse object coding scheme in area V4," *Current Biol.*, vol. 21, no. 4, pp. 288–293, 2011.

[86] Z. Ying and D. Castanon, "Statistical model for occluded object recognition," in *Information Intelligence and Systems, 1999. Proceedings. 1999 International Conference on.* IEEE, 1999, pp. 324–327.

[87] B. Pepikj, M. Stark, P. Gehler, and B. Schiele, "Occlusion patterns for object class detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.* IEEE, 2013, pp. 3286–3293.

[88] R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester, "Object detection with grammar models," in *Advances in Neural Information Processing Systems*, 2011, pp. 442–450.

[89] D. Meger, C. Wojek, J. J. Little, and B. Schiele, "Explicit occlusion reasoning for 3d object detection," in *British Machine Vision Conference.* Citeseer, 2011, pp. 1–11.

[90] C. F. Olson and D. P. Huttenlocher, "Automatic target recognition by matching oriented edge pixels," *Image Processing, IEEE Transactions on*, vol. 6, no. 1, pp. 103–113, 1997.

[91] S. Z. Der and R. Chellappa, "Probe-based automatic target recognition in infrared imagery," *IEEE Transactions on Image Processing*, vol. 6, no. 1, pp. 92–102, 1997.

[92] K. Grill-Spector and R. Malach, "The human visual cortex," *Annual Review of Neuroscience*, vol. 27, pp. 649–677, 2004.

[93] R. Malach, I. Levy, and U. Hasson, "The topography of high-order human object areas," *Trends in Cognitive Sciences*, vol. 6, no. 4, pp. 176–184, 2002.

[94] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: A module in human extrastriate cortex specialized for face perception," *The Journal of Neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[95] B. D. McCandliss, L. Cohen, and S. Dehaene, "The visual word form area: expertise for reading in the fusiform gyrus," *Trends in Cognitive Sciences*, vol. 7, no. 7, pp. 293–299, 2003.

[96] R. Epstein, A. Harris, D. Stanley, and N. Kanwisher, "The parahippocampal place area: Recognition, navigation, or encoding?" *Neuron*, vol. 23, no. 1, pp. 115–125, 1999.

[97] F. Mormann, S. Kornblith, M. Cerf, M. J. Ison, A. Kraskov, M. Tran, S. Knieling, R. Q. Quiroga, C. Koch, and I. Fried, "Scene-selective coding by single neurons in the human parahippocampal cortex," *Proceedings of the National Academy of Sciences*, pp. 46–53, 2017.

[98] J. Gomez, F. Pestilli, N. Witthoft, G. Golarai, A. Liberman, S. Poltoratski, J. Yoon, and K. Grill-Spector, "Functionally defined white matter reveals segregated pathways in human ventral temporal cortex associated with category-specific processing," *Neuron*, vol. 85, no. 1, pp. 216–227, 2015.

[99] T. F. Shipley and P. J. Kellman, *From fragments to objects: Segmentation and grouping in vision.* North Holland, 2001.

[100] M. Bolduc and M. D. Levine, "A real-time foveated sensor with overlapping receptive fields," *Real-Time Imaging*, vol. 3, no. 3, pp. 195–212, 1997.

[101] P. Artal, *Handbook of Visual Optics, Volume Two: Instrumentation and Vision Correction.* Taylor & Francis, 2017.

[102] J. M. Henderson and A. Hollingworth, "The role of fixation position in detecting scene changes across saccades," *Psychological Science*, vol. 10, no. 5, pp. 438–443, 1999.

[103] J. Kevin O'Regan, H. Deubel, J. J. Clark, and R. A. Rensink, "Picture changes during blinks: Looking without seeing and seeing without looking," *Visual Cognition*, vol. 7, no. 1-3, pp. 191–211, 2000.

[104] W. S. Geisler and J. S. Perry, "Real-time simulation of arbitrary visual fields," in *Proceedings of the 2002 Symposium on Eye Tracking Research and Applications (ETRA'02)*, 2002, pp. 83–87.

[105] M. Carrasco, B. McElree, K. Denisova, and A. M. Giordano, "Speed of visual processing increases with eccentricity," *Nature neuroscience*, vol. 6, no. 7, pp. 3263–3277, 2003.

[106] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, and T. Masquelier, "Deep networks can resemble human feed-forward vision in invariant object recognition," *Scientific Reports*, vol. 6, pp. 46–53, 2016.

[107] A. Farzmahdi, K. Rajaei, M. Ghodrati, R. Ebrahimpour, and S.-M. Khaligh-Razavi, "A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans," *Scientific Reports*, vol. 6, p. 25025, 2016.

[108] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao, "Deep learning based radiomics (DLR) and its usage in noninvasive idh1 prediction for low grade glioma," *Scientific Reports*, vol. 7, pp. 46–53, 2017.

[109] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413–420.

[110] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comp. Vis. Imag. Underst.*, vol. 106, no. 1, pp. 59–70, 2007.

[111] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision*

and pattern recognition, 2006 IEEE computer society conference on, vol. 2. IEEE, 2006, pp. 2169–2178.

[112] F.-F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 524–531.

[113] S. Hochstein and M. Ahissar, "View from the top: Hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, no. 5, pp. 791–804, 2002.

[114] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1233–1258, 1987.

[115] A. D. Jepson and M. R. Jenkin, "The fast computation of disparity from phase differences," in *Computer Vision and Pattern Recognition, 1989. Proceedings CVPR'89., IEEE Computer Society Conference on*, no. 7045. IEEE, 1989, pp. 398–403.

[116] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[117] D. J. Fleet, A. D. Jepson, and M. R. Jenkin, "Phase-based disparity measurement," *CVGIP: Image Understanding*, vol. 53, no. 2, pp. 198–210, 1991.

[118] D. J. Fleet and A. D. Jepson, "Stability of phase information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1253–1268, 1993.

[119] T. Aach, A. Kaup, and R. Mester, "On texture analysis: Local energy transforms versus quadrature filters," *Signal Processing*, vol. 45, no. 2, pp. 173–181, 1995.

[120] W. He and K. Yuan, "An improved canny edge detector and its realization on FPGA," in *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, vol. 2, no. 7045. IEEE, 2008, pp. 6561–6564.

[121] W. Gao, X. Zhang, L. Yang, and H. Liu, "An improved sobel edge detection," in *2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, vol. 5, no. 7045. IEEE, 2010, pp. 67–71.

[122] S. Marĉelja, "Mathematical description of the responses of simple cortical cells," *J. Opt. Soc. Am.*, vol. 70, no. 11, pp. 1297–1300, 1980.

[123] A. Hyvärinen and P. Hoyer, "Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces," *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[124] A. Hyvärinen, M. Gutmann, and P. O. Hoyer, "Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in v2," *BMC Neuroscience*, vol. 6, no. 1, p. 12, 2005.

[125] Y. Karklin and M. S. Lewicki, "A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals," *Neural Computation*, vol. 17, no. 2, pp. 397–423, 2005.

[126] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1794–1801.

[127] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[128] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, "Learned-norm pooling for deep feedforward and recurrent neural networks," in *Machine Learning and Knowledge Discovery in Databases*, vol. 5, no. 7045. Springer, 2014, pp. 530–546.

[129] X. Sun and W. Xu, "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves," *IEEE Signal Processing Letters*, vol. 21, no. 11, pp. 1389–1393, 2014.

[130] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *J Machine Learning Tech.*, vol. 2, no. 1, pp. 37–63, 2011.

[131] N. Rasiwasia and N. Vasconcelos, "Holistic context modeling using semantic co-occurrences," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 1889–1895.

[132] N. Ras and N. Vas, "Scene classification with low-dimensional semantic spaces and weak supervision," vol. 53, no. 1, pp. 1–6, 2008.

[133] J. Liu and M. Shah, "Scene modeling using co-clustering," in *2007 IEEE 11th International Conference on Computer Vision*, vol. 2525. IEEE, 2007, pp. 1–7.

[134] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via plsa," in *European conference on computer vision*, vol. 2. Springer, 2006, pp. 517–530.

[135] A. Alameer, G. Ghazaei, P. Degenaar, J. A. Chambers, and K. Nazarpour, "Object recognition with an elastic net-regularized hierarchical MAX model of the visual cortex," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1062–1066, 2016.

[136] H. Liang, X. Gong, M. Chen, Y. Yan, W. Li, and C. D. Gilbert, "Interactions between feedback and lateral connections in the primary visual cortex," *Proceedings of the National Academy of Sciences*, vol. 59, no. 3, pp. 323–346, 2017.

[137] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of The 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.

[138] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[139] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision.* Springer, 2014, pp. 818–833.

[140] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses

in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[141] J. Kubilius, "Partially occluded figures," 2016. [Online]. Available: 10.6084/m9.figshare.2114191.v1

[142] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. Von Der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on computers*, vol. 42, no. 3, pp. 300–311, 1993.

[143] G. W. McConkie and K. Rayner, "The span of the effective stimulus during a fixation in reading," *Attention, Perception, and Psychophysics*, vol. 17, no. 6, pp. 578–586, 1975.

[144] E. M. Reingold, L. C. Loschky, G. W. McConkie, and D. M. Stampe, "Gaze-contingent multiresolutional displays: An integrative review," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 45, no. 2, pp. 307–328, 2003.

[145] D. J. Parkhurst and E. Niebur, "Variable-resolution displays: A theoretical, practical, and behavioral evaluation," *Human Factors*, vol. 44, no. 4, pp. 323–346, 2002.

[146] J. M. Henderson, K. K. McClure, S. Pierce, and G. Schrock, "Object identification without foveal vision: Evidence from an artificial scotoma paradigm," *Attention, Perception, and Psychophysics*, vol. 59, no. 3, pp. 323–346, 1997.

[147] W. S. Geisler and J. S. Perry, "Variable-resolution displays for visual communication and simulation," *SID Symposium Digest of Technical Papers*, vol. 30, no. 1, pp. 420–423, 1999.

[148] P. M. van Diepen and M. Wampers, "Scene exploration with fourier-filtered peripheral information," *Perception*, vol. 27, no. 10, pp. 1141–1151, 1998.

[149] A. M. Larson and L. C. Loschky, "The contributions of central versus peripheral vision to scene gist recognition," *Journal of Vision*, vol. 9, no. 10, pp. 46–53, 2009.

[150] P. K. Kaiser, *The joy of visual perception.* York University, 2009.

[151] W. S. Geisler, J. S. Perry, and J. Najemnik, "Visual search: The role of peripheral information measured using gaze-contingent displays," *Journal of Vision*, vol. 6, no. 9, pp. 195–212, 2006.

[152] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 977–992, 2001.

[153] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network: Computation in Neural Systems*, vol. 14, pp. 391–412, 2003.

[154] M. R. Greene and A. Oliva, "The briefest of glances the time course of natural scene understanding," *Psychological Science*, vol. 20, no. 4, pp. 464–472, 2009.

[155] D. D. Coggan, L. A. Allen, O. R. Farrar, A. D. Gouws, A. B. Morland, D. H. Baker, and T. J. Andrews, "Differences in selectivity to natural images in early visual areas (V1–V3)," *Scientific Reports*, vol. 7, pp. 3263–3277, 2017.

[156] S. S. Miguel Thibaut, Thi Ha Tran and M. Boucart, "The contribution of central and peripheral vision in scene categorization: A study on people with central vision loss," *Vision Research*, vol. 98, pp. 46–53, 2014.

[157] M. Boucart, C. Moroni, M. Thibaut, S. Szaffarczyk, and M. Greene, "Scene categorization at large visual eccentricities," *Vision Research*, vol. 86, pp. 35–42, 2013.

[158] J. Lee, J. Wang, D. Crandall, S. Šabanović, and G. Fox, "Real-time, cloud-based object detection for unmanned aerial vehicles," in *Proceedings of IEEE International Conference on Robotic Computing (IRC)*, 2017, pp. 36–43.

[159] A. Alameer, G. Ghazaei, P. Degenaar, and K. Nazarpour, "An elastic net-regularized HMAX model of visual processing," in *Proc. The 2nd IET Int. Conf. Intelligent Sig. Process. (ISP)*, 2015, p. 4.

[160] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[161] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in Neural Information Processing Systems*, 2015, pp. 1135–1143.

[162] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[163] F. Ronquist and J. P. Huelsenbeck, "Mrbayes 3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[164] O. R. Joubert, G. A. Rousselet, D. Fize, and M. Fabre-Thorpe, "Processing scene context: Fast categorization and object interference," *Vision Research*, vol. 47, no. 26, pp. 3286–3297, 2007.