2015

Brain inspired approach to computational face recognition

da Silva Gomes, Joao Paulo

Copyright statement

BRAIN INSPIRED APPROACH TO COMPUTATIONAL FACE RECOGNITION

by

JOÃO PAULO DA SILVA GOMES

A thesis submitted to Plymouth University in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

September 2014

João Paulo da Silva Gomes

Brain inspired approach to computational face recognition

Abstract

Face recognition that is invariant to pose and illumination is a problem solved effortlessly by the human brain, but the computational details that underlie such efficient recognition are still far from clear.

This thesis draws on research from psychology and neuroscience about face and object recognition and the visual system in order to develop a novel computational method for face detection, feature selection and representation, and memory structure for recall.

A biologically plausible framework for developing a face recognition system will be presented. This framework can be divided into four parts: 1) A face detection system. This is an improved version of a biologically inspired feedforward neural network that has modifiable connections and reflects the hierarchical and elastic structure of the visual system. The face detection system can detect if a face is present in an input image, and determine the region which contains that face. The system is also capable of detecting the pose of the face. 2) A face region selection mechanism. This mechanism is used to determine the Gabor-style features corresponding to the detected face, i.e., the features from the region of interest. This region of interest is selected using a feedback mechanism that connects the higher level layer of the feedforward neural network where ultimately the face is detected

to an intermediate level where the Gabor style features are detected. 3) A face recognition system which is based on the binary encoding of the Gabor style features selected to represent a face. Two alternative coding schemes are presented, using 2 and 4 bits to represent a winning orientation at each location. The effectiveness of the Gabor-style features and the different coding schemes in discriminating faces from different classes is evaluated using the Yale B Face Database. The results from this evaluation show that this representation is close to other results on the same database. 4) A theoretical approach for a memory system capable of memorising sequences of poses. A basic network for memorisation and recall of sequences of labels have been implemented, and from this it is extrapolated a memory model that could use the ability of this model to memorise and recall sequences, to assist in the recognition of faces by memorising sequences of poses.

Finally, the capabilities of the detection and recognition parts of the system are demonstrated using a demo application that can learn and recognise faces from a webcam.

Contents

Abstract **iii**

Acknowledgements **xv**

Author's declaration **xvii**

1 Introduction **1**

1.1 Problem formulation 2

1.2 Overview of the thesis 3

1.3 Tour of this thesis 6

2 Face recognition as a biological and computational process: literature review **9**

2.1 Face detection and recognition in computer vision 10

2.1.1 Face recognition process: from image capture to classification 10

2.1.2 Applications of face detection and recognition 15

2.1.3 Face detection 18

2.1.4 Face recognition 21

2.1.5 Dimensionality reduction and Feature extraction 25

2.1.6 Memory models 29
2.1.7 Classification 30
2.2 Face detection and recognition in neuroscience 35
2.2.1 Neural mechanisms for face detection and recognition 37
2.2.2 Computational models of biological face recognition 44
2.3 Face recognition in psychology 47
2.4 Conclusions 51

3 Modelling of the face features extraction and face detection 55
3.1 Model description 56
3.1.1 Methods 65
3.2 Model improvements 70
3.2.1 Parameters adjustment 71
3.2.2 Mechanism for region of interest segmentation 72
3.2.3 Multi pose face detection 73
3.2.4 A demo for face detection, memorisation and recognition . . . 74
3.3 Conclusions 91

4 Coding of face features 93
4.1 Face features 98
4.2 Face feature coding 98
4.3 Conclusions 103

5 Analysis of the quality of coding 105
5.1 Geometry of multidimensional coding space 107

5.1.1 Database of face images 107

5.1.2 Procedure to compare feature vectors 108

5.1.3 Results of feature vector comparison 111

5.2 Conclusions 116

6 Memory model approach for multi-pose face recognition 119

6.1 Model for memorisation of sequences 123

6.2 Proposed memory model for multi-pose face recognition 130

6.3 Conclusions 137

7 Summary 139

7.1 Contributions 142

7.2 Future work 143

Bibliography 145

Bound copy of published papers 163

List of Figures

2.1 RGB colour space geometric representation. 12

2.2 HSV colour space representation. 13

2.3 Examples of faces detected using the face Analysis online demo from the Aurora Computer Services. 17

2.4 Different kinds of features shown relative to the enclosing search window. 20

2.5 Face Bunch Graph. 23

2.6 ICA based face recognition. 28

2.7 Support Vector Machines. 33

2.8 Lift transformation. 34

2.9 Diagram of the pathway from the retina to the primary visual cortex. 36

2.10 Hierarchical architecture of the visual system. 39

3.1 Simplified representation of the four layer feedforward network. 58

3.2 Convolution kernels used in the model as edge detectors. 59

3.3 Examples of images used to train the network. 67

3.4 Examples of successfully detected faces at different poses, sizes and positions. 77

3.5 Visual representation of the weights between V1 complex cells and V4 cells layers after the training phase. 78

3.6 Visual representation of the weights between V1 complex cells and V4 cells layers after the training phase. 79

3.7 Visual representation of the weights between V1 complex cells and V4 cells layers after the training phase. 80

3.8 Examples of successfully and unsuccessfully detected faces at challenging lighting conditions and poses. 81

3.9 Visual representation of the weights between V1 complex and the first neuron of the V4 cells layer. 82

3.10 Visual representation of the weights between V1 complex and the second neuron of the V4 cells layer. 83

3.11 Visual representation of the weights between V1 complex and the third neuron of the V4 cells layer. 84

3.12 Face detected after pressing "next image". 85

3.13 Another face detected after pressing "next image". 86

3.14 Adding face to the gallery for posterior identification. 87

3.15 Example of live face detection. 88

3.16 Example of enrolment. The currently detected face is added to the local gallery of faces. 89

3.17 Example of live identification. The face detected is matched against the local gallery of faces and the label of the best match is displayed. 90

4.1 Coding schemes. 101

4.2 Feature extraction and coding process. 102

5.1 Ten subjects from the Yale Face Database B (Georghiades et al., 2001).108

5.2 9 poses per subject from the Yale Face Database B. 109

5.3 29 out of 64 lighting conditions per subject from the Yale Face Database B. 110

5.4 Correct identification rates for initial test. 113

6.1 Spiking neural network model. 124

6.2 Spiking neural network model. 132

6.3 Memorisation of a sequence of four poses for a single subject 133

6.4 Forwards recall of a sequence of four poses 134

6.5 Backward recall of a sequence of four poses 135

6.6 Memorisation of two sequences of poses. 136

List of Tables

3.1 Layers parameters. 68

3.2 Convolution kernels. 69

4.1 Hamming distances between different binary coded orientations with two bits per orientation. 102

5.1 Correct matching rates for 1 output neurons. 114

5.2 Correct matching rates for 3 output neurons. 115

Acknowledgements

I would like to dedicate this thesis to my wife Yushuo and my son Luís, for all the inspiration, love and support they have given me.

I cannot forget to thank my parents Jorge and Fátima, and my sister Sandra for all they have done for me during all my life.

I would like to thank my supervisor Roman Borisyuk, for all his help, his restless pursue of excellence, for his patience, tutoring, inspiration and great example. I want to thank my second supervisor, Tony Belpaeme, for always being available when I needed, and for the good advice and conversations we have had.

I want to thank Aurora Computer Services, for allowing me the time to finish this thesis, and for giving me the great opportunity of putting into practise the skills learnt during the PhD degree.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Committee.

Work submitted for this research degree at the Plymouth University has not formed part of any other degree either at Plymouth University or at another establishment.

This study was financed with the aid of a University of Plymouth Research Studentship.

Relevant scientific seminars and conferences were regularly attended at which work was presented; external institutions were visited for consultation purposes and two papers have been published in refereed proceedings and a journal, one as a leading author and the second publication as a co-author. Also research activities in a commercial environment have been carried out with Aurora Computer Services from the fourth year of the PhD onwards. One technical report has been published with Aurora during this period, as a co-author, and the number one position in the world in one of the most recognised benchmark tests for face recognition (Labelled Faces in the Wild) has been achieved, with a mean classification accuracy of 0.9324. The research carried out within Aurora Computer Services is not included in this

thesis due to its commercial nature and because it was carried out independently from the PhD studies.

Word count of main body of thesis: 31 000

Signed: _____

Date: _____

Referred conference proceedings article:

da Silva Gomes, João and Borisyuk, Roman (2012). "Biological brain and binary code: quality of coding for face recognition". In: *Proceedings of the 22nd International Conference on Artificial Neural Networks*. ICANN 2012. Lausanne, Switzerland: Springer, pp. 427 - 434.

Referred journal article:

Borisyuk, Roman and Chik, David and Kazanovich, Yakov and **da Silva Gomes, João** (2013). "Spiking neural network model for memorising sequences with forward and backward recall". In: *Biosystems* 112, pp. 214 - 223.

Technical report:

Heseltine, Thomas and Szeptycki, Przemyslaw and **Gomes, João** and Ruiz, Maria C. and Li, Peng (2014). "Aurora Face Recognition Technical Report: Evaluation of Algorithm "Aurora-c-2014-1" on Labelled Faces in the Wild".

Oral and Poster Presentations:

2012: Cognition Institute Research Day, Plymouth. **Poster presented: Gomes, João** and Borisyuk, Roman. "Biological Brain and Binary Code: A Hybrid Model for Face Recognition".

2012: International Conference on Artificial Neural Networks, Lausanne. **Oral presentation: da Silva Gomes, João** and Borisyuk, Roman (2012). "Biological brain and binary code: quality of coding for face recognition". **Special session co-organiser:** "Faces: How We Recognise Them". European Neural Network Society **travel grant Award received**.

Research visits to external institutions:

2010: Two weeks in the Institute for Neural Computation to work under supervision of PD Dr. Rolf P. Würtz, Bochum.

Review work:

2011: PLoS ONE.

2012: Cognitive Computation.

Other Conferences and Workshops attended:

2010: Computational Neuroscience Workshop, Plymouth and Bristol.

2010: Mathematical Neuroscience, Edinburgh

2010: Computer Vision and Pattern Recognition, San Francisco

2010: International Summer School on Pattern Recognition, Plymouth

2011: 2nd PLUS Advanced School on Computer Vision, Genoa

2012: Tsinghua University's Summer Semester Short Course on Biological Vision, Beijing

Chapter 1

Introduction

In computer science, biological systems have been inspiring scientists and engineers since the development of the earliest computers. References to brain-inspired artificial systems can be found in the literature as early as the 1950's, for example in the 1958 book "The Computer and the Brain" by one of the pioneers in computing, John von Neumann (Neumann, 1958). Some time later, David Marr published his classic book "Vision" (Marr, 1982), which took a biological approach to the specific problem of computer vision. Since then, many more bio-inspired computer algorithms have been developed in order to perform visual tasks such as object detection, segmentation and tracking. Despite great efforts and advances in developing such artificial visual systems, their limitations are obvious in comparison to the performance of real biological systems. In most cases humans are much better than even the best artificial algorithms at visual tasks such as recognition. There is therefore a huge ongoing research effort in this area, the aim of which is to further improve brain-inspired architectures and algorithms so that their performance is one day

comparable to that of humans. In addition to this motivation, the development of brain-inspired algorithms in computer vision may also accelerate our understanding of how the brain works. Since many details of the relevant biological systems are not known, a biologically inspired system will always have several simplifications and estimations replacing the missing details. These can be used as the basis of hypotheses about how certain brain mechanisms work.

Within this context, in this thesis we propose a brain-inspired approach to the problem of computational face recognition.

1.1 Problem formulation

The **aim** of this thesis is to develop a comprehensive computational framework for studying the face recognition that incorporates biologically inspired aspects of the visual system, by using algorithms which have parallel in the brain mechanisms of vision (Masquelier and Thorpe, 2007), and memory organization in the brain (Borisyuk et al., 2013). This brain-inspired framework should provide new approaches and algorithms that are able to recognise faces under challenging lighting and pose conditions. These new developments can be useful in the real world in automatic face detection applications such as image tagging in social network websites and photo cataloguing applications, as well as in automatic image tagging software (Zhang and Zhang, 2010). Ultimately the system should be able to store several faces in its internal memory and recognise a new face, according to the previously memorised faces. In order to store and compare faces a suitable representation should be used, which consists in a set of features extracted from the face and a

coding scheme which encodes such features in a discriminative way. This recognition process also has a huge array of real world applications such as passenger identification, in a secured environment such as an airport, for identification and access for an online secured application such as online banking, or even for organising personal photos based on the persons present in each photo (Senior and Bolle, 2002).

The **scope** of this study has been somewhat limited because the subject of face recognition is very broad. Existing algorithms vary greatly in both their goals and the domain within which they work. We therefore limit our system to greyscale, 8-bit, 2D images, with only one face per image. We also consider only human faces. Additionally, the neural models incorporated in our system contain several simplifications and approximations. When implementing a computational model of a large portion of the visual system many details must be left out or approximated. This is due to a lack of experimental data and to hardware constraints, i.e., computations should be performed quickly enough to respond to real-time video input.

1.2 Overview of the thesis

This thesis presents and analyses a computational face recognition framework inspired by biological visual processing mechanisms. The developed system demonstrates a very good performance on the level of the state of the art recognition rates when tested using the Yale B Face Database (Georghiades et al., 2001). These encouraging results were achieved while respecting the other goal of this project, which was to develop a model that was close to biological vision mechanisms. Therefore our study advances the understanding of key mechanisms for biological vision. This

is in contrast to the alternative approach of developing a purely mathematical model with the only goal of achieving a high recognition rate.

The starting point in developing the face recognition framework was to look at the relevant existing work from three communities: neuroscience, psychology and computer vision.

To summarise, our review of the psychology and neuroscience literature indicated that object recognition, and in particular face recognition, relies on several mechanisms:

1. Hierarchical processing, in which information from simple edge responsive neurons in the primary visual cortex flows through the ventral stream, where new neurons respond to increasingly complex features until reaching the IT where there is a single cell from a population responding to a face stimulus.
2. An attention mechanism which is both feed-forward and feedback driven, that enables the selection of the intermediate features corresponding to the face detected. The most salient parts in the image trigger the corresponding neurons in the lower levels of the feedforward network, and this signal is passed until reaching the face selective neurons in the higher levels. Then a feedback mechanism that connects the V4/IT cells to the V1 complex cells, highlights the region of interest in the V1 complex.
3. Highly efficient neural coding, in conjunction with very fast and hierarchical processing through the ventral stream, allows high level representation of faces.
4. Synaptic plasticity enables the visual system to adapt and learn new object representations, e.g., faces.

5. The brain learns the appearance of a face (or other 3D object) throughout an experience by observing the face at different points in time. Each time the face is seen its pose and illumination are different, yet the brain can construct a memory of the face based upon this information.

The framework presented in this thesis is biologically plausible and consists of four key elements: The first element is the ability to detect a face in an image. This is achieved by a face detection system based on low level visual features such as oriented bars, together with well-known brain mechanisms such as synaptic plasticity and feedback connectivity.

The second element corresponds to extracting the features of the face from the region of the image that contains it. The third element is the coding of such features. This is achieved by creating a feature space, where similar features are close to each other and the features that are less similar are located farther apart. Two alternative artificial binary codes are used, both of which code the preferred orientation in a small local region using 2 or 4-bit binary pattern. To allow good face recognition this code has to be discriminative (i.e. should easily separate different faces), therefore we analyse the quality of the binary code using multi-dimensional binary feature space techniques such as clustering, and also by computing matching rates for the same and different individuals in a well known face database.

Finally, the fourth element of our face recognition framework is the memory model, which is used to store familiar faces organised by pose. This is a theoretical model based in oscillatory neural networks. The design of such model was influenced by the results of our feature space work, which indicated that pose is a big challenge

for the face recognition framework.

1.3 Tour of this thesis

The thesis is organised as follows: Chapter 2 contains a literature review that covers several aspects of face recognition from the perspectives of neuroscience, psychology, and computer science. We review existing experimental results and computational models for face detection, facial feature discrimination, memory and classification.

Taking into consideration the information gathered during the literature review stage, Chapter 3 presents our face detection model, i.e. a mechanism of visual attention which is tuned for faces. This model is based in the representation of facial features by the means of Gabor filter responses, which are known to have similar properties to the cells in the primary visual cortex (Jones and Palmer, 1987).

Chapter 4 then presents a coding for the facial features that can be identified in the output of the face detection model. Such coding is intended to reduce the dimensionality of the vectors while keeping the separability of classes, therefore a sparse binary code which represents edge orientation is chosen.

In Chapter 5 a method of comparing faces using a simple similarity measure is presented. Several experiments were conducted in order to determine the suitability of this representation for the task of face recognition. The tests looked at intra- and inter-subject distances in feature space, the nearest neighbour for each face, and k-rank based matching (i.e. finding the nearest face in the top k matches). The results showed that our representation of faces manages to cluster together feature vectors

representing the same subject. These results where strengthened by a more elaborate test using clustering techniques, which took into consideration other factors such as pose and illumination. The face database Yale B (Georghiades et al., 2001) was chosen for performing the tests. The results of these tests are compared with state of the art methods (both biologically inspired and purely artificial).

In Chapter 6 a theoretical model for facial features memorisation based in an oscillatory neural network (Borisyuk et al., 2013) which proposes a memory organisation based in pose is presented.

Finally, in Chapter 7 we present our general conclusions, draw some directions for future work and summarise the contributions of this thesis.

Chapter 2

Face recognition as a biological and computational process: literature review

A literature review of methods and models related to face detection and recognition is presented in this chapter. This review covers several stages of the face detection and recognition process, including object segmentation, feature selection and extraction, memory organisation, storage and retrieval, and classification (object recognition). We have divided this chapter in three main sections: computer vision, neuroscience, and psychology. First, a review of the computer vision algorithms that take part in the face recognition framework in artificial systems is presented. Then, a review of the mechanisms studied by the neuroscience community regarding the face recognition framework and its biological mechanisms is done. Finally, psychological evidence of the way humans recognise faces is presented.

2.1 Face detection and recognition in computer vision

In this section a review of computer vision algorithms, processes, and applications related to face detection and recognition is presented. We start by introducing the typical process of recognising a face in an image, from the moment the image is captured until the classification stage where a tag is assigned to the face present in that image. Then we present several applications where face detection and recognition has been used. After this, a review of the face detection algorithms is presented, followed by a review of face recognition algorithms.

Finally, several computer vision algorithms that are incorporated by most face detection and recognition systems are presented in separated sections. We have divided these algorithms in several categories according to their main function in the face detection or recognition system: feature extraction, dimensionality reduction, memory models, and classification. Zhao et al. (2003) present a survey on face recognition advances.

2.1.1 Face recognition process: from image capture to classification

A typical face recognition process has several key components:

- Image capture by a sensor
- Colour and Image binary representation
- Segmentation of the face in the image (face detection)

- Feature extraction
- Dimensionality reduction
- Storage of the face representation in memory
- Classification

In this section, a connection between each component is made, in order to contextualise the role of each component in the face recognition process. Individual components are then analysed later in this chapter.

Let's start with the image capture process. A digital image is captured by an imaging sensor, which has a mechanism to generate images by converting the light captured into voltage, and then the voltage into a binary format. Most modern sensors are able to capture colour images. The mechanism used to capture the different colour channels is to use different filters to filter the light for each of the individual pixel sensors that are arranged in a grid inside the main sensor. The most common arrangement of the individual pixel sensors is to alternate colour channels. With this arrangement, at each pixel position only one colour channel is captured, therefore the values for the other two colour channels are interpolated using the neighbouring pixels values. There are alternative arrangements where there are three individual sensors for each pixel, but they are more expensive, thus less popular.

Then the colours are represented using colour-schemes (Schwarz et al., 1987). The way these colours are represented in digital images is using a positive integer value (typically 8-bits) for each colour channel, red, green and blue (RGB), therefore such images have a 24-bit depth. The RGB colour space is the most commonly used in computer vision applications (see Figure 2.1).

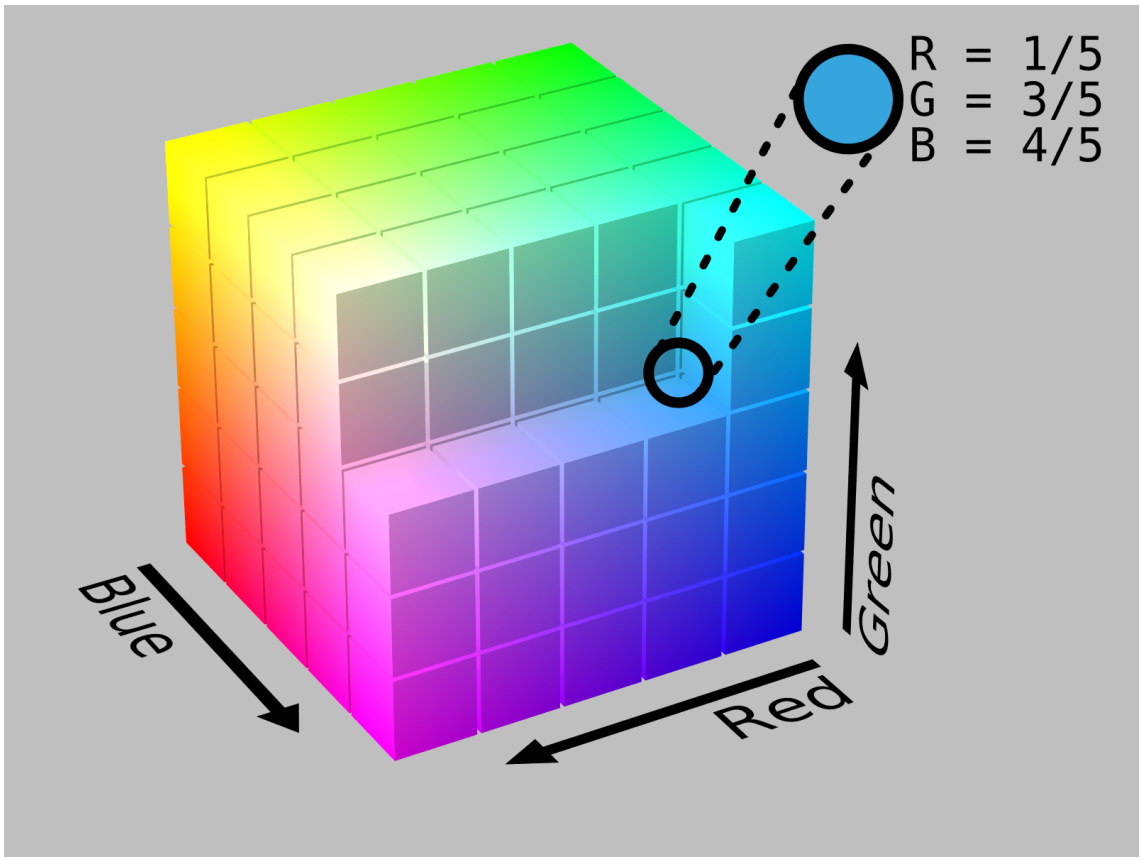


Figure 2.1: RGB colour space geometric representation: in one extreme, the origin (0,0,0), which is located at the vertex hidden in the image, the black colour is represented. Considering that each of the values red, green, and blue vary from 0 to 1, then the progressive variation resulting from the combination of the three colours is shown in the cube. In the opposite vertex of the cube, the point (1,1,1) which is also not shown in the picture, the white colour is represented. Illustration created by Michael Horvath, http://commons.wikimedia.org/.

An alternative colour representation is by hue, saturation and value (HSV). Hue is a characteristic of a colour which changes continuously, i.e., between two different tones of red there is an infinite number of hues. Saturation reflects the richness

or purity of a colour, therefore a colour with high saturation is very vivid, while a similar colour with a low saturation value tends to be closer to grey. Finally the value corresponds to the brightness of the colour. This HSV representation is more popular for applications where the user has to pick a colour from a palette or has to visualise the available colours, because it is an easy way to visualise colours (see Figure 2.2).

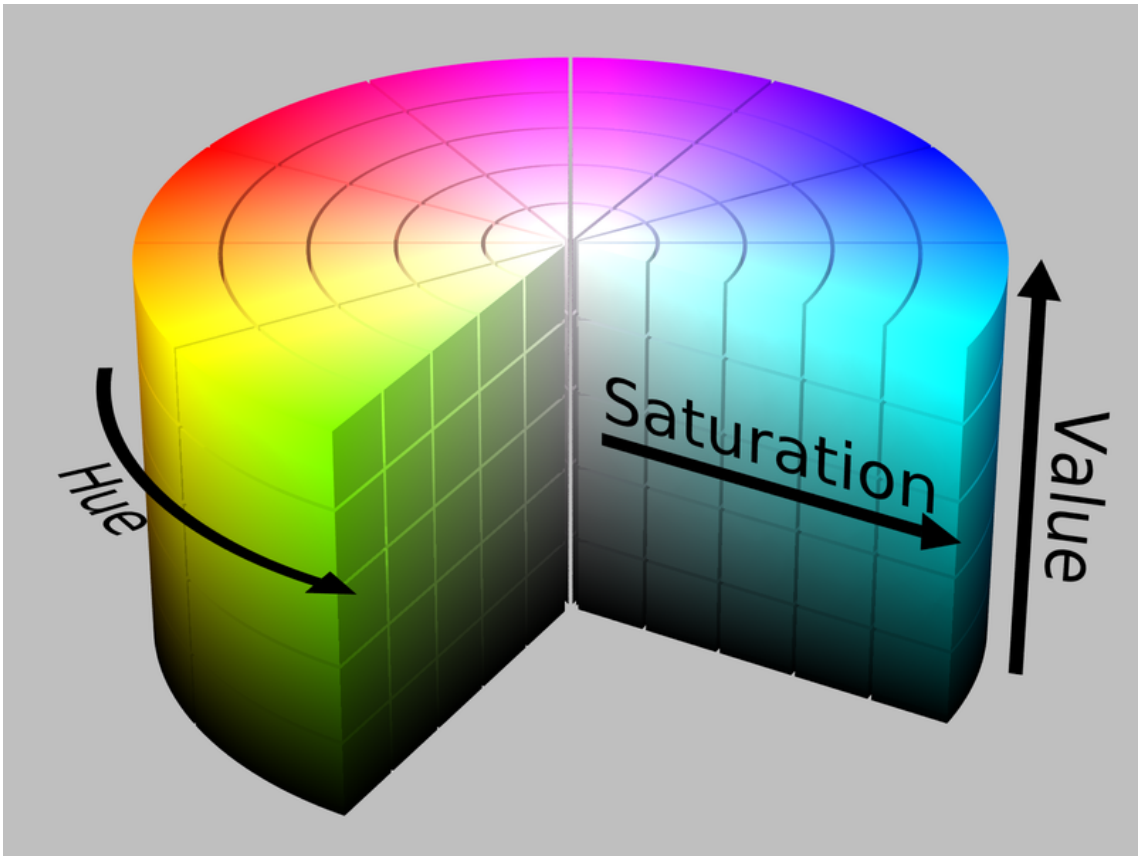


Figure 2.2: HSV colour space representation. Hue determines the colour, saturation determines the vivacity of the colour, and finally the value corresponds to the brightness of the colour. Illustration created by Michael Horvath, http://commons.wikimedia.org/.

After the image is captured the next stage in the face recognition process is to determine if there is a face in the image and the location of that face. This stage is usually called face detection or segmentation, and can be pose dependent, i.e., only faces with a certain pose can be detected, or pose invariant, where faces from different poses can be successfully detected by using view-dependent methods or invariant features to represent the faces.

Once this stage is finished, features are extracted from the face pixel values. These features can be pre-determined by the algorithm or learned from the available data. The combined set of features extracted from the face, form a feature vector, which is a representation of the face that is often more abstract and with a smaller dimensionality than the pixel representation (Huang and Yin, 2009).

In order to store several representations of the same and different faces and the corresponding labels, the face recognition process uses a memory module. This module has the role of storing and retrieving face representations (Barwinski, 2008).

Finally, a classification stage takes place to decide the identity of the face in the image. Classifiers can perform two different tasks: 1) identification, in which a classifier that can assign a label to the probe image, according to all the faces and labels represented in the memory. 2) verification, which consists in a binary match classifier, which can determine if two face representations are the same (Jafri and Arabnia, 2009). In fact the first kind of classifier can be reduced to the second, by matching the probe face individually with all the faces stored in memory.

2.1.2 Applications of face detection and recognition

Face detection and recognition applications are very important in today's world, with the proliferation of digital imagery and social media, and the need for the use of biometrics in security sensitive situations.

Starting from face detection (Zhang and Zhang, 2010), which is more than ever a hot topic, we see that it has become a standard feature in digital cameras, and mobile phones, because it allows the software to focus correctly the faces found in a picture in real time. Also some of these cameras go further by not only detecting the faces, but also analysing them, in particular, analysing if the face is smiling.

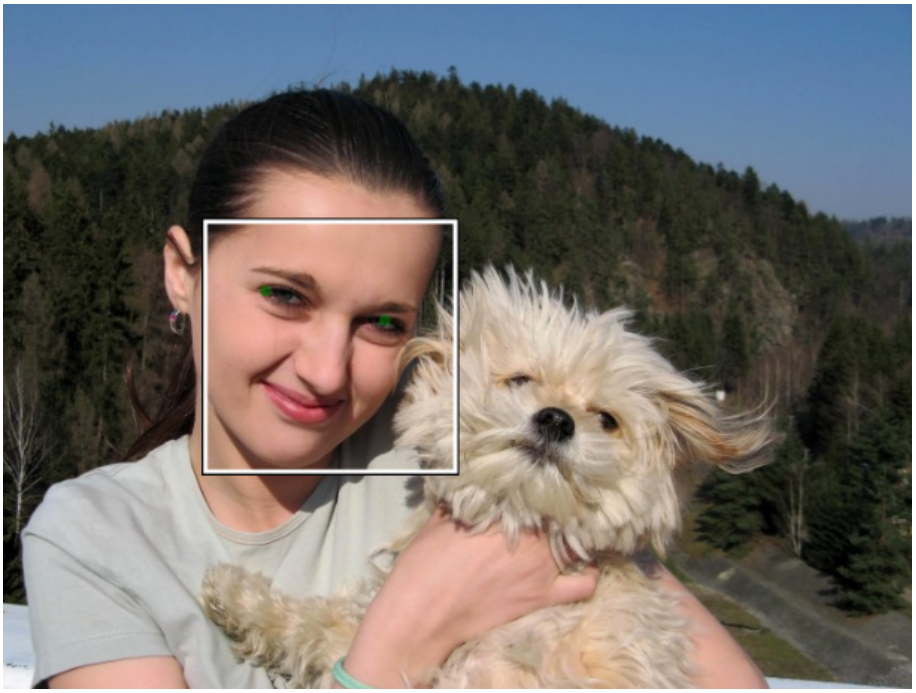
Moreover in social networks websites such as Facebook and Google+ this capability has been used in order to help the users of such websites to tag their friends more easily by showing boxes around the detected faces.

Finally, photo management software and websites have also been using this functionality in order to assist the users in tagging people, thus making the search based on the identity easier. An online demo of the face detection capabilities of Aurora Computer Services' system can be found online[1], and the output can be seen in Figure 2.3. In terms of algorithms, the de facto standard for many years has been the Viola–Jones algorithm (Jones and Viola, 2001), mainly because of its high computational efficiency which is one of the most important requirements in most applications above mentioned. This efficiency is achieved due to two main factors: the simplicity of the algorithm (looks for simple features) and the cascade structure which eliminates a large number of candidate faces in the early, more efficient stages

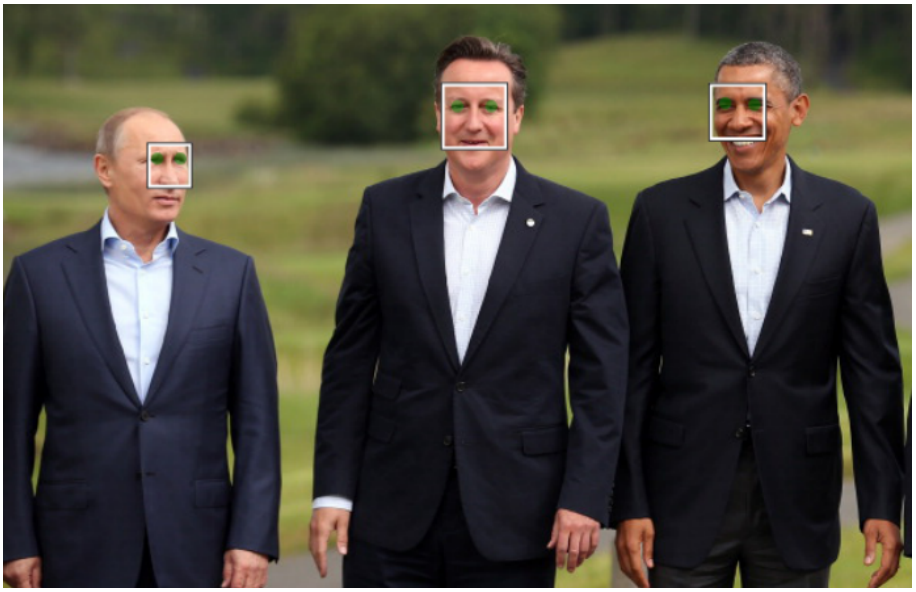
[1] https://auroracloudapi.azurewebsites.net/

of the cascade. More details of the algorithm will be introduced later in this chapter. Also, face recognition is a very important and popular topic nowadays (Senior and Bolle, 2002). It has been successfully used in security applications inside airports, such as matching between the passport photo and a live photo, used to allow or deny passengers to go through the passport control. Another example application is the enrolment of the passengers at the check in desks and the verification at the gate, in order to determine if the passenger that checked in and holds the boarding pass is the same that is entering the aircraft at the gate. One of the main limitations of such systems is the difficulty to deal with large variations of lighting conditions, which often occurs in open environments such airports. A common solution to overcome the illumination challenges is to use infra-red cameras and illumination (Kong et al., 2005). Using such technology face recognition systems can achieve very high recognition rates.

Another popular application in the field of face recognition is the recognition of social media type images. Unlike in the airport environment, in the social media images there is no degree of restriction in the environment in which the images were taken. They can have virtually any kind of lighting conditions, pose, rotation and the person is often wearing accessories such as hats, sun glasses, etc. Therefore the recognition in such conditions is much more challenging. To create a common ground to develop and compare face recognition algorithms that can be used in images from the web, a database containing thousands of this kind of images and a methodology to compare algorithms was proposed by the University of Massachusetts: the Labelled Faces in the Wild Face Database (Huang et al., 2007), and became the



(a)



(b)

Figure 2.3: Examples of faces detected using the face Analysis online demo from the Aurora Computer Services. (a) Example of successfully detected human face. Animal face is not detected. (b) Example of successfully detected human faces across different poses. Online demo can be found in https://auroracloudapi.azurewebsites.net/ .

standard test for this type of face recognition algorithms.

2.1.3 Face detection

Viola Jones algorithm

citetjones-fr-boosted-local-feat-2001 presented an algorithm that became the standard for face detection, mainly because of its computational efficiency. This algorithm has four main components: a set of features extracted from the input image, an intermediate image representation referred as integral image, a method for selecting the best features from the large pool of all possible features, and, finally, a cascade classification method to improve the speed of classification.

The simple rectangular features used in this approach are similar to those proposed by Papageorgiou et al. (1998). These features respond to the difference of intensity among rectangular regions. There are three kinds of such features see Figure 2.4. One that responds to differences between two adjacent rectangular regions. A second feature that responds to the difference between the sum of two outside rectangles and a third inner rectangle. Finally, the third feature responds to the differences of two pairs of diagonal rectangles.

One key element in the efficiency of this algorithm is the integral image. This representation contains in each position (x, y) the sum of all pixel values for the positions $x' \leq x$ and $y' \leq y$. With this representation, any of the features can be computed in constant time. For instance a rectangle defined by its top-left corner (x_1, y_1), top-right corner (x_2, y_2), bottom-left corner (x_3, y_3), and bottom-right corner (x_4, y_4) can be computed using the integral image I as follows: $I(x_4, y_4) + I(x_1, y_1) - [I(x_2, y_2) + I(x_3, y_3)]$.

Despite the low computational cost of evaluating each feature, the number of possible features for each search window is prohibited. For instance, for a window of 24x24 pixels, if we combine all the possible features with different sizes and positions, the total number of features to be computed exceeds 180,000. This is only for one window, which would be multiplied by the number of windows necessary to cover the entire image and also the number of different scales to be evaluated. For this reason, a method for selecting the most discriminative features was also presented, which is similar to the one presented by (Freund and Schapire, 1997). In order to select the features, each of them is used individually to classify the training data by applying a simple threshold to separate the faces and non-faces. This threshold is chosen individually in order to minimise the classification error. This error is weighted according to each training sample. The weights for each training sample are uniformly distributed at the beginning. Only the miss-classifications are added to the error. Finally, the classifier with the lowest weighted error is selected for the current round. Before moving to the next round, the weights of the misclassified samples are increased and the correctly classified are kept as they are. After a pre-determined number of rounds, or when the combined classification performance is satisfactory, the features (and corresponding thresholds) selection process is finished. Each feature/threshold combination is also referred as weak classifier. The final strong classifier weights each weak classifier according to the error it had and produces a final classification. Despite this weak classifier selection process, there are still a considerable number of features to be evaluated for each window, which is multiplied by the different scales and the number of windows necessary to

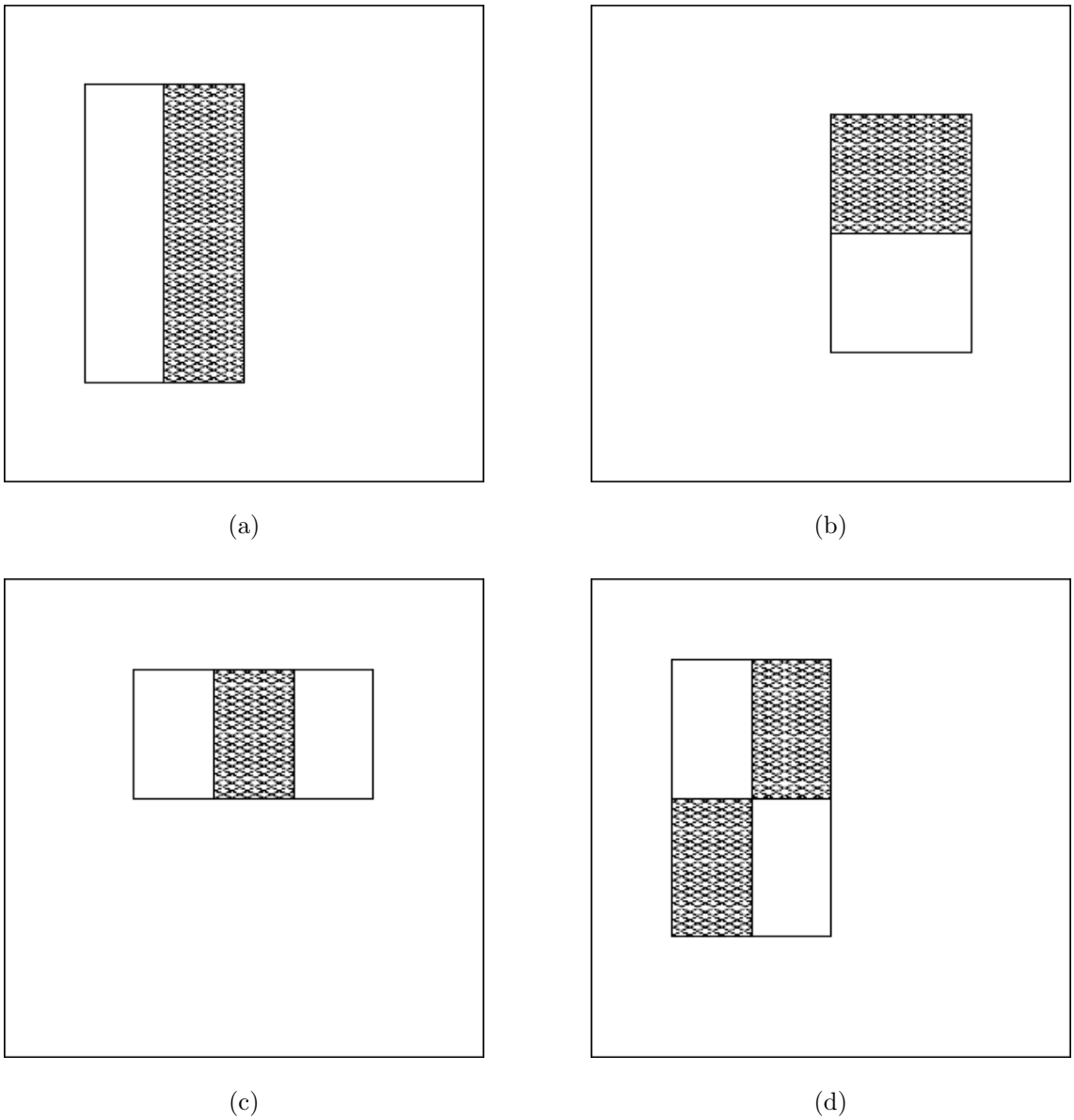


Figure 2.4: Different kinds of features shown relative to the enclosing search window. The values of the pixels from the grey rectangles are subtracted from the sum of the pixels in the white rectangles. This sum can be calculated efficiently using an integral image. (a) and (b) show two-rectangle features. (c) Three-rectangle feature. (d) Four-rectangle feature. Illustration from Jones and Viola (2001).

cover the whole image in a sliding window approach. Therefore a second stage is added to this algorithm, which is known as cascading. The idea is that we can used smaller groups of weak classifiers with low false acceptance rate at the beginning in order to quickly reject non-face windows. The selection strategy proposed consists in reducing the false positive rate and the detection rate at each stage, by setting a target for the false positive rate and a maximum decrease in detection and adding new classifiers to the stage until reaching the targets. New stages are added until the global target for false positive and detection rates is met.

2.1.4 Face recognition

Artificial neural networks

Artificial Neural Networks (ANN) have been used to develop face recognition systems.

Lawrence et al. (1997) propose a method that combines a self-organising map (SOM) neural network with a convolutional neural network to perform face recognition after local image sampling is performed. The authors also compared this approach with others where the SOM is replaced by a Karhunen–Lo'eve transform, which is a method for reducing dimensionality that has been used in several machine learning methods, and the convolutional network is replaced by a multi-layer perceptron (MLP), but the results show that the original approach achieves better performance, particularly in the case of the convolutional network comparing to the MLP.

Lin et al. (1997) present a successful usage of a probabilistic decision-based neural network in a face recognition system. This is a complete system which performs from face detection to recognition. For the recognition part, only a sub-area of the

face image is used, which includes eyes, eyebrows, and nose but excludes mouth.

A radial basis function as neural classifier for face recognition has also been proposed by Er et al. (2002). The features are extracted and their dimension reduced by using PCA and Fisher's linear discriminant analysis.

More recently, Taigman et al. (2014) and Sun et al. (2014) presented an approach that combines a 3D alignment of the face with the deep neural networks approach (Bengio et al., 2013). Firstly a 3D model of the face based on landmark localisation is built. Then a nine-layer deep neural network calculates a face representation from the raw Red, Green and Blue (RGB) pixel values, using a reduced number of convolutional layers, alternated with a max-pooling layer (Bengio et al., 2013), and followed by several other locally or fully connected layers. This approach improved the state of the art results considerably in both the Labelled Faces in the Wild (Huang et al., 2007) and YouTube Faces (Wolf et al., 2011) datasets.

Elastic bunch graph matching

The algorithm briefly described in this section was presented by Wiskott et al. (1997) and is broadly based in a previous paper from Buhmann et al. (1990).

The algorithm relies on two main concepts: Jets and Graphs.

Jets are based on a Gabor wavelet transform. The wavelet transformation is computed for a discrete set of orientations and spacial frequencies. The Jet is then defined as a set of wavelet coefficients which contains local edge information regarding different orientations and scale description at several levels Therefore a jet is a rich local feature descriptor. Phase information is often discarded because its coefficients vary greatly with small shift in the position, degrading the position invariance.

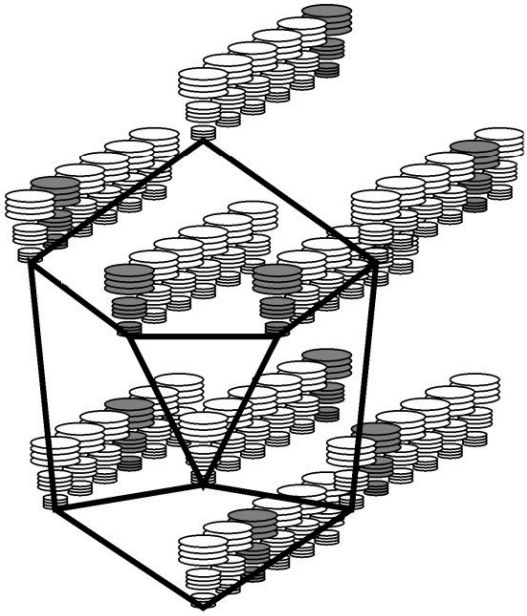


Figure 2.5: Face Bunch Graph. Each node has a set of Jets, each one corresponding to some variation in pose, expression, shape, etc. Illustration from Wiskott et al. (1997).

The Jets can be compared according to a similarity function (Wiskott et al., 1997). Graphs are usually labelled with an identity and represent a face by N nodes connected by E edges. Usually the nodes are placed in fiducial points like corners of the mouth, pupils, etc. In order to cover the wide range of possible variations caused by different expressions, positions and shapes of the facial features, a set of individual model graphs, called face Bunch Graph (FBG), is combined (see Figure 2.5).

Using these two concepts the algorithm is able to perform face detection and recognition. Face detection (Elastic Bunch Graph Matching) can be done automatically when there is a FBG large enough to cover all possible variation. The authors argue that this is achieved with 70 graphs. Face detection is then basically done by maximising the similarity between and image graph and the FBG over all positions.

Face recognition is performed very efficiently after the extraction of the model graphs from the gallery images and image graphs for the probe images, by comparing the image graph to all model graphs and selecting the one with higher similarity value. In Wiskott et al. (1997), the similarity function used is the average over the similarities between pairs of corresponding jets.

Several people have followed and improved this work. For instance, Günther and Würtz (2009) presented several maximum likelihood classifiers on this Gabor graphs and Müller and Würtz (2009) designed an automatic method for learning how to generalise over pose and Illumination.

Combination of 2D and 3D information

Some authors combined 2D with 3D information in order to attempt to improve the 2D-only traditional methods.

Wang et al. (2002) combine 2D Gabor filter responses with 3D information in the form of point signature (Chua and Jarvis, 1997). In this approach, both 2D and 3D feature vectors are projected into the corresponding PCA subspaces and combined to form a unique augmented feature vector which represents an individual face. The classification is then performed using either a similarity function or a support vector machine.

Bronstein et al. (2004) introduce a method of generating 2D illumination and expression invariant face representation from a combination of 3D geometric information with albedo of the face. From the generated 2D representation standard techniques can be then applied.

3D approach

A face recognition method relying only on 3D information was published by Lu et al. (2004). They introduce a method for building a gallery of 3D faces by combining information from 2.5D face scans from different viewpoints. This 2.5D face scans are basically augmented 2D representations where for each (x,y) position the depth information is associated. Furthermore a method for comparison of face representations based on interactive closest point algorithm is presented. This algorithm minimises the distance between two clouds of points.

Bronstein et al. (2005) present a method for 3D only face recognition in order to achieve expression invariance by assuming that it is possible to model the facial expressions as isometries of the face surface. Isometric surfaces preserve their length across deformations, i.e., do not stretch and do not tear, therefore preserving the surface metric. Other authors presented methods for face recognition based on 2.5D and 3D features such as curvatures (Gordon, 1992; Tanaka et al., 1998).

2.1.5 Dimensionality reduction and Feature extraction

Principal component analysis

The use of Eigenfaces is probably the most popular approach in face recognition (see Turk and Pentland, 1991). Despite the original system having been outperformed by many others, the core idea still applies in many modern face recognition algorithms. This approach relies on Principal Component Analysis (PCA) for extracting the principal components based on a set of training images. The calculation of the subspace is done in three main steps:

1. Calculate the covariance matrix of the input dataset over all dimensions of the features vectors
2. Calculate the orthogonal eigenvectors and eigenvalues of the covariance matrix
3. Sort the eigenvectors in the descending order of the corresponding eigenvalues
4. Select the principal components: the eigenvectors that keep a good description of the input data
5. Project the input data-points into the new basis

These principal components characterise a subspace with a smaller dimension than the original space and into which the test images are projected before classification. Pentland et al. extended the original approach by introducing a view-based multiple-observer Eigenspace technique to tackle the problem of pose variance (Pentland et al., 1994). In the same publication they also introduced a method to integrate features of the human face such as eyes, nose and mouth by means of a modular eigenspace description technique.

Most of the algorithms used in face recognition don't use colour information, even though Torres et al. (1999) studied the importance of this discarded information in the case of PCA and concluded that by including this extra information the recognition can be enhanced.

Later, Yang et al. (2004) extended the PCA to be able to handle directly 2D matrices representing the images rather than the traditional 1D vectors, i.e., the image matrix does not need to be converted in a 1D vector as a preliminary step for the covariance matrix computation. This new approach also results in a smaller covariance matrix.

Finally, Perlibakas (2004) tested 14 distance measures and some more modifications

of those that can be used with PCA to compare feature vectors in order to find which one gives best recognition rates.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) was applied to the problem of face recognition for the first time by Etemad and Chellappa (1997). This approach finds the vectors that characterise the subspace that best discriminates the different classes while keeping the within classes distances low. This is done by maximising the difference between class means while keeping the within class distances to a minimum.

This approach finds a linearly separable space, but unlike PCA, its basis vectors are not necessarily orthogonal.

LDA suffers from the small sample size problem, i.e., when the number of samples is much smaller than the number of dimensions of the input space the algorithm doesn't perform optimally. Several publications present possible solutions to overcome this problem (Cevikalp et al., 2005; Chen et al., 2000).

Liu and Wechsler (2002) applied a related approach named Fisher Discriminant Analysis to a Gabor feature space derived from the Gabor wavelet transform of the face images. The main advantage of this method is the robustness according to changes in illumination and facial expression.

Independent component analysis

An Independent Component Analysis (ICA) based face recognition method was published by Bartlett and Sejnowski (1997) and Bartlett et al. (2002). This is a statistical method which minimises not only the second order dependencies, like

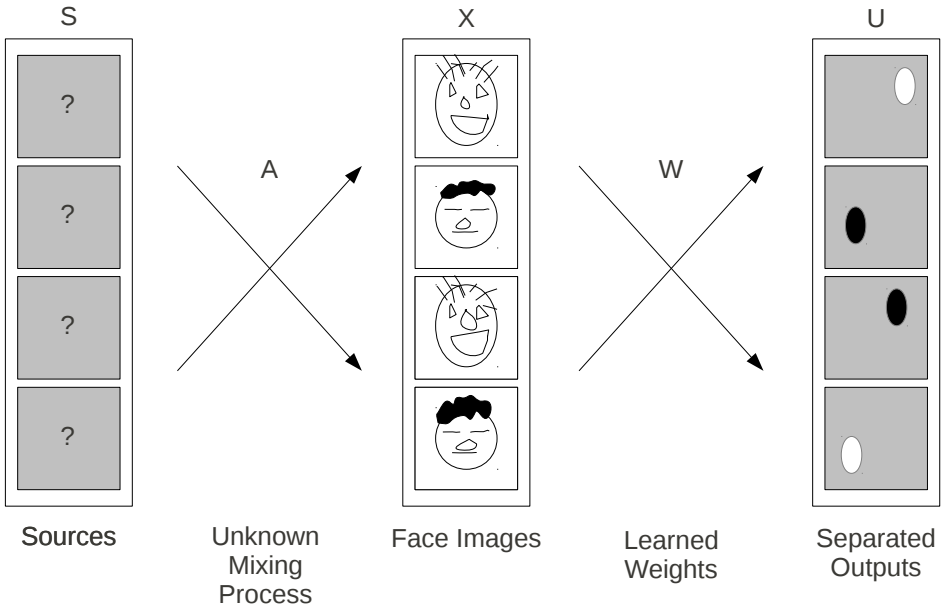


Figure 2.6: ICA based face recognition diagram showing the relationships between the source image S, the face images X, the mixing process A, separated outputs U and learned weights W (Bartlett and Sejnowski, 1997).

PCA, but also the higher order dependencies and finds a basis for representation of faces. The proposed method considers a set of images X which are the result of an unknown mixing process of the source images S and the mixing matrix A. The sources can be recovered by a matrix of filters W which is found using an unsupervised learning algorithm. These filters produce statistically independent output images U, i.e., $U = WX$ (see Figure 2.6). The rows of U are then used as the basis for face representation.

Liu and Wechsler (2003) extended this method by applying it to a vector of features extracted from the PCA reduced Gabor feature vector, instead of the original pixel vector.

2.1.6 Memory models

Barwinski (2008) presents a neurocomputational model of memory acquisition for novel faces in which each face is represented by a point in a multidimensional space. A given identity is defined by all the points representing different images of the same subject. A locally-linear embedding (LLE) algorithm is used for generating the face space and a linear mapping algorithm, graduated non convexity (GNC), to find an explicit function that expresses the dependence between input and output vectors. The face space can be either norm based or exemplar based. On one hand, the norm based face space has all the vectors placed on a hypersphere, i.e., all the faces are at the same distance from the centre. In the exemplar based face space, in the other hand, each vector is not normalised, so all the points are not contained in any particular structure. In terms of face recognition and novelty detection, both face spaces achieve good performance. These face spaces can then be used for face recognition and novelty detection. Based on the distances between a new test image and the existing images on the free space, one can evaluate if the new image has the same identity of some of the images present in memory or if it is new. One of the interesting results presented in his thesis is that the exaggeration of samples, or caricatures, improves recognition, which is in accordance to psychological evidence. The result of exaggerating all sample vectors is that the identity centre moves closer to the origin of the face space. This means that often the vector belonging to one class that best improves classification and false positive rates is not the middle one but some on the borders.

Borisyuk et al. (2013) present a memory model for memorising sequences based in

spiking neural networks, which is capable of solving some ambiguities in sequences containing repeated elements. This model pretends to be a generic memorisation mechanism that is biologically inspired and takes into account several mechanism well known in neurobiology, namely neural groups oscillations and synaptic-time-dependent plasticity. Therefore it is a good candidate for memorisation of faces, and can be adapted and used as a memory model for different appearances of the same individual. This model will be discussed and a theoretical application for face recognition presented in a separate chapter.

Finally, other models of memory and memory organisation for face recognition have been presented by DiCarlo and Cox (2007), Okada and Malsburg (2002), and Jitsev and Malsburg (2009).

2.1.7 Classification

Bayesian framework

Moghaddam et al. (1996, 1998, 2000) use the Bayesian Framework in the problem of face recognition. Their approach is based on a probabilistic similarity measure, which uses a Bayesian analysis of images differences. This approach takes into consideration the typical variations occurring between intra and inter-personal images, rather than a more traditional Euclidean nearest-neighbour matching used in the Eigenfaces technique which disregards these variations. Therefore the proposed similarity measure is expressed as follows:

$$S(I_1, I_2) = P(\Delta \in \Omega_I) = P(\Omega_I|\Delta) \tag{2.1}$$

where $P(\Omega_I|\Delta)$ is the a posteriori probability given by Bayes rule using $P(\Delta \in \Omega_I)$ and $P(\Delta \in \Omega_E)$ which are estimates of the likelihoods. These estimates are derived from the training data by using an method for density estimation of high-dimensional data (Moghaddam and Pentland, 1997).

Hidden Markov models

Nefian and Hayes (1998) presented a solution based on Hidden Markov Models (HMM). As they explain is this paper, "HMM consist of two interrelated processes: (1) an underlying, unsolvable Markov chain with a finite number of states, a state transition probability matrix and an initial state probability distribution and (2) a set of probability density functions associated with each state.". In this approach, the nodes chosen for the 1D HMM represent directly facial features as eyes, mouth, etc.

Support vector machine

Phillips (1999) introduced a Support Vector Machine (SVM) method for face recognition. Due to the binary nature of the SVMs, the face recognition problem was reformulated. The concept of difference space was introduced were the dissimilarities between two facial images are stored. The two classes that can be handled by the SVM are then formulated in this difference space: faces from the same subject and faces from different subjects.

(Guo et al., 2000) published a different approach to the binarization of the problem. They used a binary tree where in each level only a binary decision has to be made and the process continues in the winner sub-tree until the top of the tree where the

unique class appears.

Finally a comparison between different SVM approaches (holistic and components-based approach) was presented by Heisele et al. (2001). They concluded that the components-based approach outperformed the holistic approach. Support vector machines are linear classifiers which find a hyperplane that best separates positive and negative examples. Figure 2.7 illustrates the basic concepts of support vector machines, and their mathematical representation is:

- Data points x_i can only assume labels $y_i \in -1, 1$
- Support vectors: $x_i \cdot w + b \pm 1$, where w is separating plane's normal vector, and b is the distance from origin
- Distance between point and hyperplane: $\frac{\|x_i \cdot w + b\|}{\|w\|}$
- Margin: $\frac{2}{\|w\|}$

The task of the classifier (SVM) is to maximise the margin, with the constraint of correctly classifying all training data. Assuming that the data has been normalised, and the data is linearly seperable, this task can be formulated as follows:

$$\begin{aligned} &y_i = 1 \Rightarrow x_i \cdot w + b \geq 1 \\ &\qquad\qquad\qquad\qquad \rightarrow y_i^\top (x_i \cdot w + b) \geq 1 \\ &y_i = -1 \Rightarrow x_i \cdot w + b \leq -1 \end{aligned} \tag{2.2}$$

This is a quadratic optimisation problem, and can be solved with quadratic programming. This approach works well if the data is separable. But there are solutions for non-separable data.

The first solution is to use soft margins SVM (Cortes and Vapnik, 1995).

The second solution is to map the data in to a higher-dimensional space - non-linear SVM (see figure 2.8).

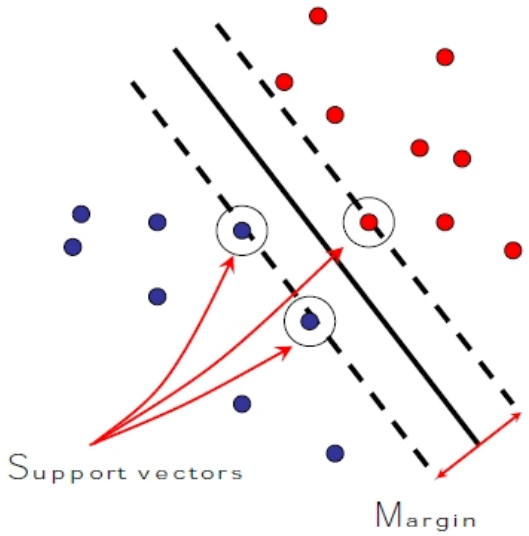


Figure 2.7: Support Vector Machines (Savarese S., PLUS School, Genoa, 2011). The dotted lines mark the boundaries of each class, which are determined by the support vectors (highlighted by black circles). The margin corresponds to the distance between the two boundanormalisedries, marked with the red double arrow.

The SVMs described above account only for two class problems. The usual solution for multi-class problems is to combine multiple two-class SVMs. For this purpose there are two alternatives:

- Classify one class against all others
  - Training: learn an SVM for each class vs. the others
  - Testing: apply each SVM to test example and assign to it the class of the SVM that returns the highest decision value
- Classify classes in pairs
  - Training: learn an SVM for each pair of classes
  - Testing: each learned SVM "votes" for a class to assign to the test ex-

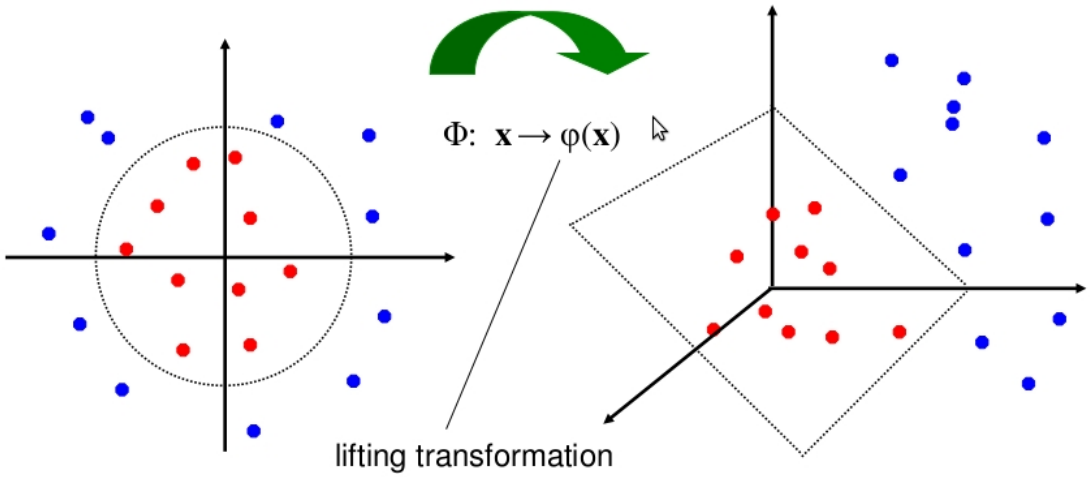


Figure 2.8: Lift transformation. The data x is mapped into a higher-dimentional space by the φ transformation. The goal of this transformation is to make the data separable in the new space. Illustration from Andrew Moore.

ample

Boosting

The AdaBoost algorithm was applied to the face recognition problem by Guo and Zhang (2001) and Guo et al. (2001). This algorithm performs a classification between two classes, so their approach is to divide a C class face recognition problem (C is the number of face images) into $C(C-1)/2$ two class problems. With this approach, a different set of features can be chosen for each pair, so only the most discriminative features are kept. After this step, the AdaBoost classifier is trained for each pair with the corresponding features. The test process is also done in two steps, first the test image is compared with all classes by a pairwise classification and then the results of all these classifications are combined to perform the final classification.

2.2 Face detection and recognition in neuroscience

The face detection and recognition process in the brain starts with a representation of the world which is based on the sensing of the light by the retina. This representation is generated from the responses of cones and rods, which are photosensitive cells in the retina that have the role of sensing the light. Due to their different sensibilities, cones and rods cells operate under different light conditions, daylight and dim light, respectively. Then neural impulses in the brain are triggered through the ganglion cells, the optic chiasm and the lateral geniculate nucleus (LGN).

An important property of this information pathway is that the responses in the visual cortex are organised topographically in the same way as the retina, hence these maps in the visual cortex are also called retinotopic maps. Image 2.9 shows the pathway from the retina to the visual cortex.

The processing of the visual information in the primary visual cortex is executed in a hierarchical manner. In the early stages there are cells that respond to very simple features, such as oriented bars, and are invariant to shifts, therefore their receptive fields are very well defined. Then, the complexity of the features and the invariance increases as the higher levels of the cortex are reached. In the higher levels of complexity there are groups of cells responding to faces (face detectors) and other groups of cells responding to particular faces (face identifiers). This kind of cells are often referred as grandmother cells and have been observed experimentally by Quiroga et al. (2005). This biological system achieves a remarkable degree of accuracy for familiar faces. People can recognise familiar faces under a very high

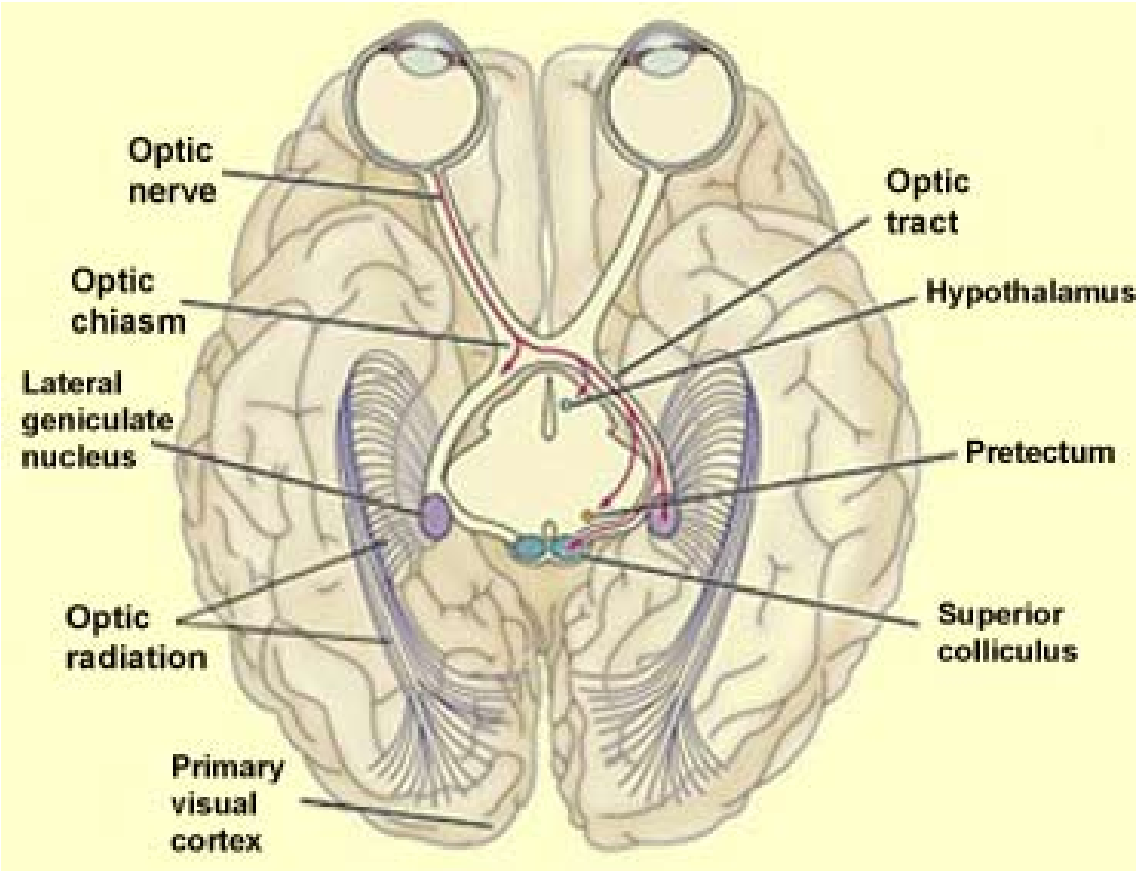


Figure 2.9: Diagram of the pathway from the retina to the primary visual cortex. Illustration from The Brain from Top to Bottom, http://thebrain.mcgill.ca/.

degree of variability of lighting, pose and age conditions, among others factors such as partial occlusion that occurs when wearing glasses, hats, etc. The Labelled Faces in the Wild Face database published the results of human performance for the task of face matching. For the funnelled images (which were preprocessed using an advanced alignment preprocessing method), the mean classification accuracy is 0.9920. This indicates the very high accuracy of humans in this task. It is important to note that the database contains mainly images of famous people, therefore the results reflect in many cases matches of familiar faces, while the results by artificial algorithms are generated with completely blind test sets. Despite recently several artificial

algorithms achieved similar results using deep neural networks (Sun et al., 2014; Taigman et al., 2014), the human capability in recognising faces still outperforms the artificial systems, particularly under difficult conditions. Therefore there are lessons to be learnt from the way our brain works in order to develop robust artificial face detection and recognition systems.

Some of the mechanism described in this section have a central role in the biological visual processing and therefore face detection and recognition. They have been well studied in the literature, thus in the next section these mechanisms will be described in more detail.

2.2.1 Neural mechanisms for face detection and recognition

Probably the most influential discovery from neuroscience with practical application in the development of computer vision algorithms, in particular artificial face recognition, was done by two Nobel prize recipients: Torsten Wiesel and David Hubel. By studying the cat's striate cortex they discovered that in the early stages of visual processing there are single cells tuned to respond to specific edge orientations and motion (Hubel and Wiesel, 1959, 1962, 1963a,b). These cells are the so called "simple cells", and they are organised in cortical columns where the orientation response changes smoothly from cell to cell. There are other cells with similar properties which are called "complex cells". The main difference is that these kind of cells have a high degree of spacial invariance, i.e., they respond to similar stimuli, but over a larger receptive field. This kind of responses can be simulated by the so called Gabor filters (Jones and Palmer, 1987), which makes this an interesting and widely used biologically inspired feature extractor for computer scientists.

These low level orientation features are only the first stage of what is believed to be a complex hierarchical and distributed object recognition mechanism in the brain. This system is distributed in the sense that the representation of different features or characteristics (like colour, texture, and context) takes place in different areas of the brain and only in a later stage there is a convergence of all these cues, which contribute to the final classification. It also seems that there is a hierarchical representation of features. As mentioned before, in the primary visual cortex the features are as simple as edge orientations, but when moving towards higher cortical areas the complexity increases as well as the size of the receptive fields so there are groups of neurons responding to a class of objects. Figure 2.10 illustrates this hierarchical architecture. A good review about these mechanisms underlying visual object recognition was written by Palmeri and Gauthier (2004). Back in 1981, ambitious work was published by Baron (1981) where a computational model of the whole system behind face recognition is presented. In this model "Several fundamental processes are implicated: encoding of visual images into neural patterns, detection of simple facial features, size standardisation, reduction of the neural patterns in dimensionality, and finally correlation of the resulting sequence of patterns with all visual patterns already stored in memory.". Although this is a very complete work and the author tries to find the correspondences of the several artificial neural networks used in this system and the brain anatomy, he admits that numerous questions remain unanswered, which is not surprising, since 30 years after this publication the scientists keep looking for many answers in this field.

Many other face recognition computational models inspired by the brain have been

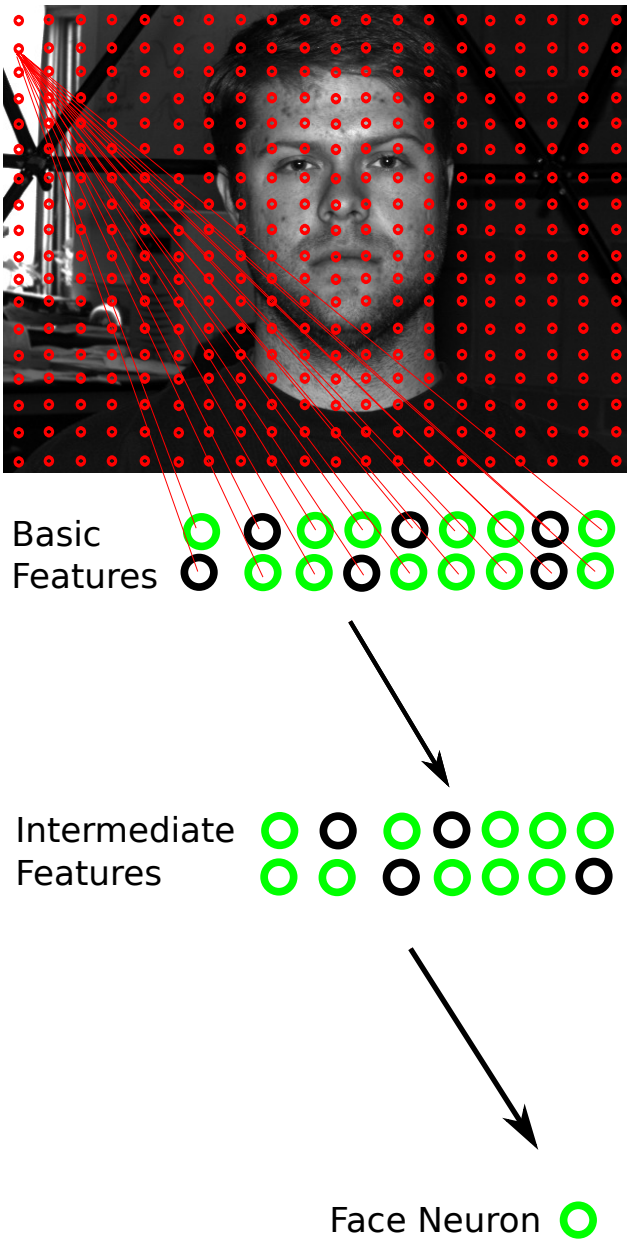


Figure 2.10: Hierarchical architecture of the visual system, illustrating a possible face detection/recognition neural network. From the input stimulus at the top of the figure, basic features such as edges are extracted. Then by combining several basic features, more complex features such as corners and other more complex shapes emerge. Finally, in the higher levels of the network there are groups of neurons that can respond to high level concepts such as faces. Face Image from Georghiades et al. (2001).

developed but in this review only three of them are included due to their higher biological plausibility. A face recognition algorithm was presented by Wiskott et al. (1997), which is based in the mapping of a low level representation similar to a hyper-column-like activation pattern onto a higher order representation which maintain the 2D relative relations and spatial values. This is known as the Elastic Bunch Graph Matching algorithm. Biederman and Kalocsais (1997) used this model to compare the differences between face and general object recognition.

The second model was published by Delorme and Thorpe (2001) and is based on the idea of rapid feed-forward spike propagation suggested by the short response latencies of the face selective neurons in the inferotemporal cortex. The authors present a network of three layers organised in a retinotopic fashion using a synaptic time dependent plasticity learning rule which can be trained to recognise faces.

Finally, Jitsev and Malsburg (2009) present a generic recognition system which was applied to the human face recognition problem. This system is based on the evidence that the visual cortex encodes, stores and retrieves objects in a parts-based fashion, although some authors believe that in the case of face recognition a more holistic approach takes place in the brain (Farah, 1996). The hierarchical memory model presented supports storage and recall of the parts' representations and relies in a process based on a slow bidirectional synaptic plasticity together with a homoeostatic unit activity regulation, both relying on a fast activity dynamics and a winner-take-all mode modulated by an oscillatory rhythm.

One of the most asked questions from scientists is which brain cells respond to a face stimulus. Perrett et al. (1982) present a study where several neurons from the

fundus of the superior temporal sulcus (STS) in the cortex where recorded. The results suggests that these neurons integrate a system specialised in facial features coding, relating then the damage of this system to prosopagnosia. Rolls and Baylis (1986) also studied the responses of these neurons located in the STS to face stimulus, and the invariance to size and contrast of these neurons. They concluded that the neurons show a high degree of variance according to these two parameters. The same author also studied the primate temporal lobe cortical visual areas' role in invariant face and object recognition. He says that temporal cortical visual areas have similar properties in terms of neuronal populations as the primate temporal lobe. So the author argues that these populations code for objects and faces, so the study of both sets of neurons is giving some help in solving a very hard computational problem which is invariant object recognition (Rolls, 2000). More recently, Rolls (2008) presented a more general model that tries to connect different areas of face processing. This model covers from the neurons' invariant responses to position, size, view and spacial frequency in the inferior temporal visual cortex (for face and other objects stimulus), to brain regions responsible for face processing such as the orbitofrontal cortex and amygdala. The author argues that the invariant representations provided by the neurons in the inferior temporal visual cortex have ideal properties to be used as input to the other areas referred before. In a different approach, Haxby et al. (2002) state that the visual recognition and perception of faces are performed in different areas. In the occipitotemporal regions in extrastriate visual cortex the analysis of the faces' visual features is performed while in other areas such as the amygdala the analysis of more emotional properties related to

faces is carried out.

Recently, Quiroga et al. (2005) have shown evidence of single cells located in the medial temporal lobe (MTL) which responds for individual faces. More precisely, it is suggested that this neurons respond to the identity of a certain person, so stimuli like the written name of the subject will activate the same neuron as a face image of this person. Perrett et al. (1998) considers a population of cells instead of single cell's activity in the temporal cortex. The authors related the time course of this activity to the variation of the speed of recognition from different viewing perspectives, arguing that this variation is due to a specialisation of the visual mechanisms for the most familiar views and not to a metal transformation from the canonical view.

A different mechanism involved in face recognition, memory, has been studied by Haxby et al. (1996). The authors found that the human neural systems that perform the coding of memories for faces is dissociated of the system used for recall of the same faces.

Another interesting finding is that by simply reversing the contrast polarity of the input image, the recognition is severely impaired despite all edges and spatial frequencies remaining the same (George et al., 1999). The reason for this is unknown but George et al. (1999) showed that "bilateral posterior areas in fusiform gyrus responded more strongly for faces with positive than with negative contrast polarity. An anterior, right-lateralized fusiform region is activated when a given face stimulus becomes recognisable as a well-known individual".

Haan et al. (2002) show evidence that newborns have some predefined preference

to face stimuli which could indicate that and their cortical face processing systems are subject to specialisation mechanisms during the development. Moscovitch et al. (1997) also shows evidence about the specialisation for faces by studying a man with object agnosia and dyslexia caused by a closed-head injury. They conclude that face recognition relies on two systems. One system is face-specific and performs recognition in a holistic manner, and a second system that is not face-specific. The second system is shared by other recognition tasks and relies on the parts rather than the whole, i.e., relies on the internal object features individually.

DiCarlo and Cox (2007) give an overview of the current state of brain inspired object recognition. Their aim is to put together several ideas into a unified framework, which they believe is the core of any future brain inspired object recognition system. To do so, the authors explicitly ignore some mechanisms that may interfere with recognition such as tracking, similarity estimation, texture recognition, etc. It is claimed that the key point on recognition is to understand the neural coding from the ganglion cells through the ventral stream. In particular how this high-dimensional representation of the input stimulus, can in each stage of the visual processing be untangled, until reaching a high level representation in the IT area, where every object can be approximated by a manifold or a single point in the high-dimensional space where the separation between different identities is more clear. This is a purely theoretical work, therefore much can be done in order to implement models of the idea of untangled feature space.

2.2.2 Computational models of biological face recognition

David Marr's work about the visual system (Marr, 1982) integrates all the aspects of vision like colour, shape, motion, and object representation in a single frame of models. Probably the most important contribution of this work is a hierarchical system of representation of visual scenes. This system is divided into three main stages. First the low level "raw primal sketch" which represents the retinal image. Then a "2.5D sketch", where depth and orientation information is integrated. Finally at the higher level the objects are represented by a simplified 3D model which is used for scene interpretation (for example, face recognition). Although the details of the actual brain mechanisms might be different from David Marr's model, a hierarchical representation of visual features, with increasingly higher complexity, seems to be one of the key mechanisms underlying the human vision.

Marr's book is an incomplete work due to his short life. Despite there are some inconsistencies with the actual knowledge of the brain's visual system, nevertheless his work is still an important reference for those who develop brain inspired visual systems.

Brunelli and Poggio (1993) compares two commonly used approaches on face recognition, geometrical features and template matching. The authors concluded that the second more holistic approach leads to better results than the first parts-based approach. But interestingly 10 years later, Heisele et al. (2003) concludes the exact opposite: parts-based approaches are better than holistic for face recognition.

A common problem in face recognition is that the face can be recorded from an infinite number of different views. Beymer (1994) present a possible approach to

solve this problem by basically representing different poses by the corresponding templates in the gallery. To classify a new face the pose is estimated first and then the classification procedure takes place taking into account the actual pose of the input. A different solution for this problem by Lam and Yan (1998) only needs one training image.

Jeng et al. (1998) proposes a method for face feature detection and a geometrical face feature model which could be used as basis for building face recognition systems based of facial features.

Recently, Meyers and Wolf (2008) have shown that it is possible to use biologically inspired features (i.e. features which are similar to those extracted by neurons in the visual cortex) and still perform as good as some of the most successful artificial face features representations such as local binary partterns, and histogram of gradient features. This observation motivates biologically inspired face recognition systems, such as the one presented in this thesis.

Riesenhuber and Poggio (1999) describe and analyse an hierarchical model of the visual processing and object recognition in the cortex. Their model is based on physiological data from the inferotemporal cortex, which is responsible for visual processing in the brain. Namely two observations are taken into account. Firstly, the model builds upon the simple and complex cells concept (Hubel and Wiesel, 1962). The neighbouring simple cells feed into a single complex cell, which results in the complex cell being phase-invariant. Secondly, the observation that in the macaque inferotemporal cortex (IT) cells that respond to a specific view of objects such as faces are thought to have an important role in invariant recognition was

also incorporated into the model (Bruce et al., 1981). The network presented has an hierarchical structure starting with a layer of simple cells (S1) that receives input directly from the input stimulus (image), followed by a complex cell layer (C1). Then, a second layer of simple cells (S2) receives input from the previous layer (C1), which then feeds another complex cells layer (C2). Finally, an array of view-tuned cells finishes the hierarchy. This last layer of view-tuned units receives input from the C2 layer. There is the possibility that some connections skip certain layers, for instance, some neurons from C1 can feed directly to C2 neurons. The simple cells perform a weighted sum operation. While the complex cells perform a max operation. The view invariance is achieved with view-tuned units while the scale and translation invariance is achieved by the architecture of the network. The authors compare the use of max and sum operations in the complex cells and concluded that the nonlinear max operation allows the cell to respond to the most salient feature, which is a good way to pool the afferent responses and achieve a good invariance and seems to be in line with neurophysiological data. Furthermore they investigate if the proposed model could achieve a degree of selectivity and invariance similar to the findings from the physiology. To do so, a network trained with 21 view-tuned units was tested with samples rotated in three dimensions, scaled and translated around the preferred view. The responses to the test samples were compared to those responses generated by distractors, and the results show a clear degree of invariance and selectivity.

Delorme and Thorpe (2001) also propose a feed-forward network capable of performing face identification which is inspired in biological and psychophysical studies that

suggest that a very fast, automatic, feed-forward mechanism can produce highly selective responses in the visual system. This network uses spiking neurons and a rank order coding (Gautrais and Thorpe, 1998) which is in line with previous findings that suggest that the highly complex recognition tasks can be perfomed under 150 ms by the visual system (Thorpe et al., 1996), and provides a very efficient way of encoding information. The network consists of three layers of retionotopic maps of integrate-and-fire neurons. The three layers are a simplification of the primate visual system. The first layer corresponds to the retina, the second layer to the V1 region and the third and last layer corresponds to the V4/IT region. Therefore the network as an increasingly complex structure in terms of features represented at each layer. The network presented is simulated using SpikeNet (Delorme and Thorpe, 2003; Delorme et al., 1999) which is an efficient software package that can simulate a large number of asynchronously firing integrate-and-fire neurons. The results show that the network is capable of determining the identity of a person from views that are not present in the training data, and the network shows good robustness to noisy and low-contrast inputs.

2.3 Face recognition in psychology

In psychology there are several studies related to face recognition which give some insights about the mechanisms and features that are important to humans for performing such task. These insights can be an important source of inspiration in building artificial face recognition systems. Perhaps these studies are not so suitable as neuroscience's in order to give an almost direct guidance to the structure

and architecture of such artificial systems, but they do help by providing benchmarks for artificial systems, and specific examples of limitations of the human and primates performance in face recognition, as well as in which cases humans excel such task. This data can therefore be used as a comparison benchmark for artificial systems, and also help develop and improve such systems by providing a better understanding of the variables that improve or disrupt the performance of a face recognition system. These studies cover many aspects of human and primates face recognition. Yin (1970) suggest that face recognition is a special case of object recognition and concluded about different mechanism for recognising upright and inverted faces by comparing performance on recognising faces between patients with no brain damage and others with brain damage. These differences between recognition of inverted and upright faces are presented in several other publications (Eimer, 2000; Tanaka and Sengco, 1997). Probably the most studied area are the features used by humans on face recognition. Patterson and Baddeley (1977) studies the effect of disguises (changes on features appearance), pose and expression in the face recognition performance. The authors concluded that changes in pose and expression don't compromise the recognition but major disguises can severely impair the recognition. It has been concluded that the usage of personality characteristics improves the recognition performance. This last conclusion will not be taken into account in this thesis because of the complexity of incorporating such personality features in any artificial system. Later, Ellis et al. (1979) studies the effect of internal and external features by comparing the recognition of familiar and unfamiliar faces using internal and external features. The authors argue that both features

are used on recognition but the recognition of familiar faces rely more heavily on internal features whereas recognition of unfamiliar faces rely more on external features. Configuration of the features seems to have an important role on recognition according to Tanaka and Sengco (1997). In their experiments, the authors found that facial features are better recognised when placed in the original configuration than when their configuration is changed (for instance when eyes are more close together or more separated). The recognition performance is even more degraded when the features are presented isolated. Haig (1984) also concluded that the position of the face features is highly important for recognition, namely "the vertical positioning of the mouth, followed by eyes, and then the nose, as well as high sensitivity to close-set eyes, coupled with marked insensitivity to wide-set eyes".

A developmental perspective of face recognition can be also found in the literature. Evidence that newborns have a preferred attention to faces and that this is due to the fact that infants are born already with the information about the face structure has been shown (Morton and Johnson, 1991). Also Nelson (2001) agrees that face recognition is different from other recognition tasks, when he clearly states that "Evidence from fields as diverse as cognitive, evolutionary, and developmental psychology, as well as cognitive neuroscience, has increasingly pointed to the 'special' nature of face recognition". In the same paper the author also argues that there are evidences that faces start being seen as different, when comparing to any other objects, in 6 months old infants. Carey et al. (1980) also studies the development of face recognition between 6 and 16 years old children and he concludes that the performance is improved between 6 and 10 years of age and then stabilises before

improving again at the age of 16. This study suggests again that face recognition is a special case of object recognition. The special nature of face recognition has been concluded by Farah (1996) as well. By studying brain damaged and normal subjects, the author concludes that the mechanisms and brain areas used for face recognition differ from the ones used in general object recognition. Furthermore, it is claimed that the representation of objects and faces are different. Having individual features is more importance in object recognition and a more holistic approach is beneficial in face recognition (Farah, 1996).

Other authors studied the face recognition problem with a more broad approach. For instance, Bruce and Young (1986) presented a theoretical model of face recognition and perception. According to their work, recognition of familiar faces is done by matching what they refer as structural encoding products with previously stored ones. The authors argue that the cognitive system has a main role on the first decision about the actual identity of a new face or if its only possible to identify as a resemblance. To finalise this section, two books on the psychological perspective of face recognition and perception are introduced. Young (1998) deals with a broader concept than face recognition - face perception, which is the process of interpreting the face not only in terms of identity, but also emotions, race, gender and other attributes. Nevertheless face recognition is approached in this work by a collection of research review papers and gives some insights on the brain mechanisms behind recognition, recall of faces, and errors in disguised face recognition. Also, Li and Jain (2005) present several chapters with a psychological perspective on face recognition, from which two main conclusions can be drawn. Firstly, the dynamic nature of face

recognition and face processing is presented in one of the papers, which indicates that motion has a important role in enhancing the amount of information available, therefore improving and enabling better face recognition and interpretation. Secondly we can conclude that human face recognition is very robust and invariant to several factors such as pose, age gap and lighting. This robustness is visible only when recognising familiar faces. When unfamiliar faces are used to test the human capabilities, the overall recognition performance is much worst, and some artificial algorithms can even outperform humans.

2.4 Conclusions

In this literature review we have covered the most important aspects of face detection and recognition, from three different perspectives: psychology, neuroscience and computer vision.

The main conclusions from the psychological point of view are that face recognition is a very important cognitive function for humans from birth, and there is actually evidence that we have some tendency to look at faces and we are able to recognise familiar faces from early days of life. There is also evidence that we have a mechanism for face recognition separated from the general object recognition, and we rely of internal and external features for the recognition according to the familiarity of the subject. Finally the human performance in this task is very high, and robust to changes in pose and expression.

From a neuroscience point of view, the evidence in the literature indicates that the face detection and recognition is implemented in the brain by a hierarchical mech-

anism of increasing feature complexity. In the first stages of this hierarchical system are the simple and complex cells of the V1 region of the primary visual cortex, which are selective to oriented bars. In the end of the hierarchy neurons responding to very complex features such as faces and with a high degree of complexity have been experimentally observed.

Finally, from the computer vision algorithms related to face detection and recognition, which were analysed in this review we concluded that these algorithms follow mostly this sequence of steps: a feature extraction method, followed by a dimensionality reduction step, then a storage and recall system (memory), and, finally, a classifier.

To perform the feature extraction, several methods for extracting discriminative features from the image pixel values are presented. These features combined generate highly dimensional feature vectors representing the faces, therefore several methods to reduce this dimensionality, and often increase the separability of the feature space, are also presented. Then, we introduced several memory models for storing, organising and retrieving feature vectors, and, finally, several classification algorithms were presented.

The main contributions from this chapter are:

- A review of the state of the art in face detection and recognition algorithms, models and methods from psychology, neuroscience and computer vision
- Identification of the expected behaviour of a biologically inspired face detection and recognition model, from a psychological point of view
- Identification of the main blocks of a biologically inspired face detection and

recognition model

- Identification of the typical sequence of algorithms and methods used in an artificial face detection and recognition algorithm.

Chapter 3

Modelling of the face features extraction and face detection

In this chapter we start by presenting a biologically inspired model of face features extraction developed by Masquelier and Thorpe (2007). This model is biologically plausible and is capable of learning the features which can discriminate faces from other classes of objects, therefore it can be used as a face detector. Although the model performance is good, a mechanism to determine the region of interest corresponding to a face is missing. Furthermore it required several improvements. For example, the size of the receptive fields is rather small and has to be adjusted. Therefore we have studied the model properties and modify them to improve the model performance to the level which is required for reliable Face recognition.

In section 1 the original model by Masquelier and Thorpe (2007), is described in detail. This model has been implemented in Matlab by the original authors and the code has been used for the experiments presented in this thesis. In section 2

we present the original contribution from this thesis, which consists in modifications and improvements of the previous model. The conclusions of this chapter are presented in the last section.

3.1 Model description

As we have seen in the literature review, there are several approaches that are commonly used in many object recognition tasks and, in particular, in face detection. We have also considered several artificial face recognition algorithms widely used by the Computer Vision and Image Processing communities that actually implement approximations of some of these popular mechanisms. Taking into consideration the literature review, a biologically inspired model for learning visual features introduced by Masquelier and Thorpe (2007) was chosen as a starting point for building a face detector. It is based on a well-known hierarchical visual processing model known as HMAX (Riesenhuber and Poggio, 1999) which is a feedforward hierarchical convolutional network with four layers. It relies on a synaptic-time-dependent plasticity rule (STDP) for adjusting the intermediate connections strengths. The network can learn input patterns in a completely unsupervised manner, by repeatedly displaying many examples of one category, such as faces. After being trained, the network can respond very quickly when stimulated with an input from a category previously learned, which is in line with findings that indicate that after only 100ms the neural responses can discriminate the nature of the input (Hung et al., 2005).

The four layers can be seen as representing four different types of cells found through the ventral pathway. V1 simple cells, V1 complex cells, V4 cells and V4/IT cells.

Figure 3.1 shows a schematic representation of the neural network for face detection. The first layer is composed of four groups of orientation selective simple cells. In this thesis, these are referred to as V1 simple cells since they have properties similar to the real V1 simple cells, i.e., they are selective to specific orientations and they have a good degree of invariance regarding changes in illumination. The V1 simple cells are implemented as simplified versions of Gabor filters, more precisely, they are convolution kernels of 5x5 pixels which are basically edge detectors (see Figure 3.2). The Gabor filter has wavelength of 5 pixels and effective width of 2. There are four preferred orientations: $\pi/8$, $3\pi/8$, $5\pi/8$, and $7\pi/8$. The orientation $\pi/8$ was chosen as a starting point in order to avoid focusing on horizontal and vertical edges which are rarely informative, since they occur very often in nature. Figure 3.2 displays the four kernels used. The responses of the Gabor filters are converted to spike latency according to the following formula: $l = |r|^{-1}$, where l is the latency and r is the convolution value. This conversion is made because the order or rank of each spike is the main coding element of this network.

The latency of each cell (Gabor filter) is inversely proportional to the activation strength, i.e., the closer orientation a given neuron is presented to, regarding its own preferred orientation, the stronger it will respond, and therefore the earlier it will fire.

At this stage, there is a competition mechanism in place that limits the length of the spike train by enforcing a winner-take-all mechanism at each location, by setting all other responses values to zero, i.e., only the neuron which has the preferred orientation closer to the stimulus will fire.

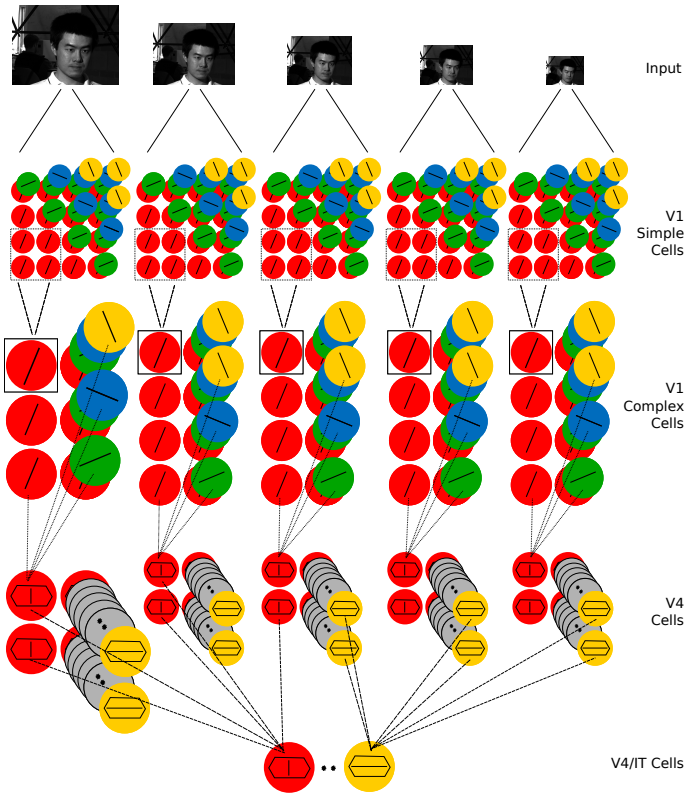


Figure 3.1: Simplified representation of the four layer feedforward network capable of learning face features. The input image is greyscale and the resolution is 640x480 pixels. The input is replicated at 5 different scales. The first three layers are replicated for each processing scale. The V1 simple cells layer contains four maps corresponding to different orientations of the Gabor filters. The V1 complex cells layer performs a max operation of a square region from the previous layer (in the picture a 2x2 area is shown just for illustration purposes). The V4 layers take inputs from maps corresponding to different orientations, therefore resulting in a more complex features, such as faces. In our model we use either 1 or 3 V4 cells. The last layer integrates input from all the processing scales, in order to be more robust to scale variations. In each layer, different colours represent different kinds of neurons. Each of which responds to a particular feature (bars of a particular orientation in the lower layers, and face features in the higher layers). Face images from the Yale Face Database B (Georghiades et al., 2001).

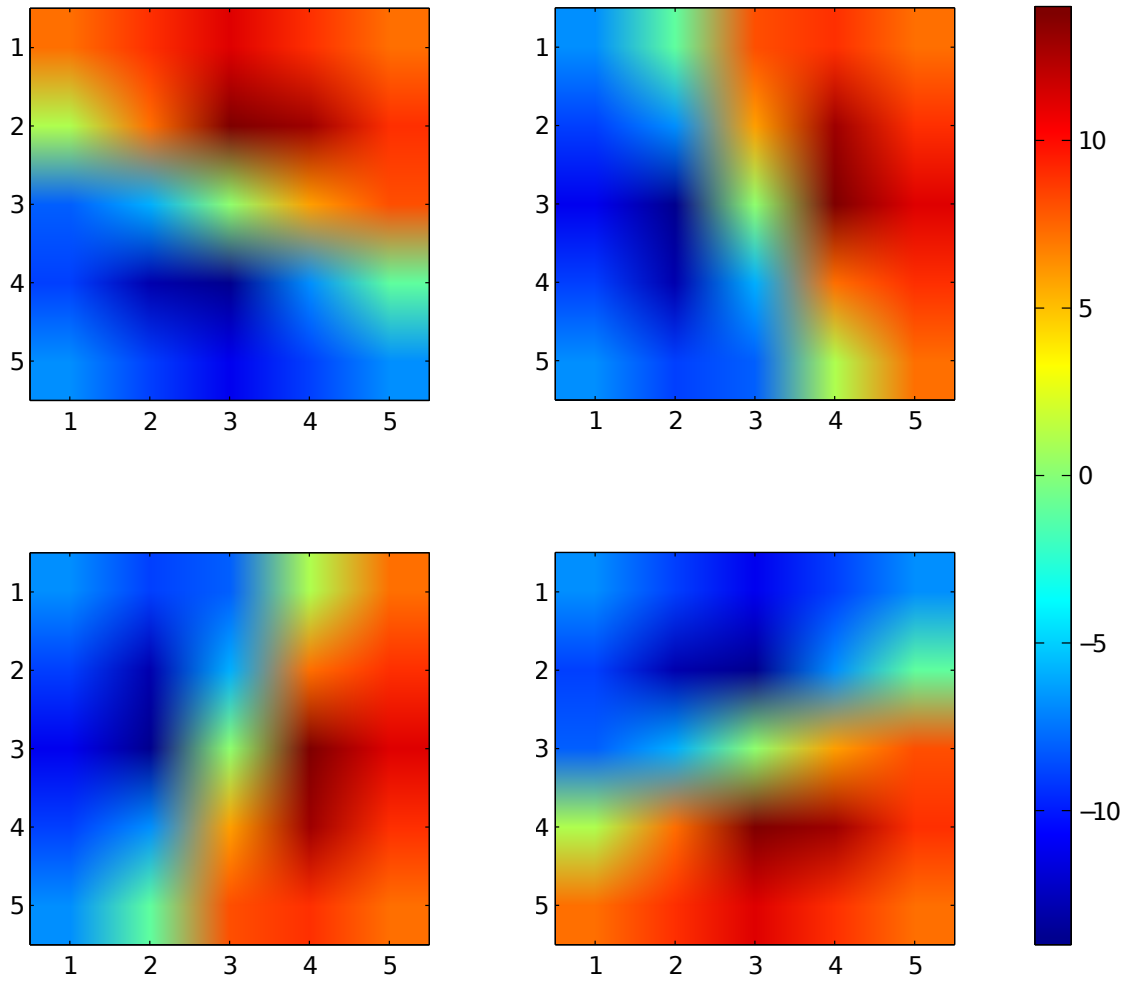


Figure 3.2: Convolution kernels used in the model as edge detectors. These kernels are a rough approximation of the real part of Gabor filters. Gabor filters can be expressed by the following equation: $g_{\lambda,\theta,\varphi,\sigma,\gamma}(x,y) = exp(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2})cos(2\pi\frac{x'}{\lambda}+\varphi)$ with $x' = xcos\theta + ysin\theta$ and $y' = -xsin\theta + ycos\theta$, where λ is the wavelength, θ is the orientation, φ the phase offset, σ is the standard deviation of the Gaussian factor, and *gamma* the aspect ratio. The kernels shown in the image approximate Gabor filters with $\lambda = 5$, $\theta \in \{\frac{\pi}{8}, \frac{3\pi}{8}, \frac{5\pi}{8}, \frac{7\pi}{8}\}$, $\varphi = -90$, $\gamma = 1.0$ and width=2 (which is related to the ratio between σ and λ).

The second layer of this network contains neurons corresponding to the complex cells in the primary visual cortex, which are spatially invariant up to a certain degree. For this reason we refer to this layer as the V1 complex cells layer. Each cell from this layer performs a max-pooling operation, i.e., each cell propagates only the maximum response from its receptive field, which is basically a square region of 7×7 simple V1 neurons, and takes only the spike from that region as input, i.e., the spike with the lowest latency. This means that the V1 complex cell map sub-samples a V1 simple cell map. There is an overlap of 1 neuron between each 7×7 region. This mechanism reduces the dimensionality of the input data by a factor of 36 (6 x 6 pixels, considering the overlap) as well as makes the system shift invariant, while being biologically plausible (Riesenhuber and Poggio, 1999). Furthermore, at the same level, lateral inhibition is introduced, which is another mechanism which is also inspired by the brain. A cell that is exited inhibits the neighbouring neurons with the same preferred orientation in a 11×11 region. The scale of this inhibition is proportional to the distance from the exited neuron, and can vary linearly from 5% inhibition, for the farthest neurons, to 15% inhibition, for the adjacent neurons. This mechanism ensures the suppression of regions dominated by a particular orientation, because the cells with the same preferred orientation inhibit each other.

The third layer corresponds to the V4 cells. The cells from this layer are selective to features of intermediate complexity such as a combination of different orientations which can be seen as individual features such as the eyes and mouth, or as the whole or parts of the face. In our experiments we used either one map of V4 cells (corresponding to the whole face) or three kinds, corresponding each one to a slightly

different region/different scale of the face. We would like to notice that in this model the first three layers are replicated for five different input scaled images, in order to introduce scale invariance to the system. These five scales are respectively: 100%, 71%, 50%, 35%, and 25%, therefore there are 4 x 5 = 20 S1 maps, the same number of C1 maps, and either 5 x 3 = 15 V4 maps or 5 x 1 V4 maps, depending on the configuration. The weights of the connections between the V1 complex cells and the V4 cells are shared among the five different maps. Each map of V4 cells receives spikes from the corresponding processing scale in the V1 complex cells maps only, i.e., it receives spikes from four maps corresponding to the four orientations represented in the previous layer, for the same processing scale. The receptive field of the V4 cells in this model is a square region of 16×16 cells, which corresponds to receptive fields with various sizes in the original image because of the different processing scales. Note that the synaptic connections between V1 Complex cells and V4 cells are modelled by using weights, and all other connections simply transmit the pre-synaptic neurons. In this model there is no need for a leakage mechanism because the spike waves are propagated one by one and the potentials are reset before each wave. The threshold used for the V4 cells is 64, which is one quarter of the number of afferent cells ($16 \times 16 \times 1/4$), and considering that the initial weights are randomly generated with mean 0.8 and standard deviation 0.05. At this stage a winner-take-all mechanism is adopted, therefore only one replicated cell can be excited, i.e., the first cell for each map among all processing scales to fire is the winner. Furthermore a mechanism to balance the number of of cells that fire among all processing scales is introduced in order to avoid that only neurons from certain

scales would be selected. This is implemented by limiting the number of cells firing at each stage to k (a k-winner-take-all approach). This is an important mechanism during the learning phase in order to avoid that in the beginning only patterns from certain scales are selected. In addition, a local inhibition mechanism is activated in this layer during the learning period in order to avoid that different prototype cells (cells from different maps) learn the same pattern. If the inhibition area was too large the different prototype cells could't learn different parts of the same object, and if too small the inhibition wouldn't work as a deterrent for learning the same pattern. Therefore, a local area corresponding to half of the height and width of the receptive field is used. This archived by preventing other prototype cells from firing in a region of 8×8 around the first prototype cell to fire.

The learning of the face patterns is done by presenting images, one by one, i.e., by stimulating the first layer. When the spikes that have been propagated from the first layer through the whole network reach the V4 cells layer, a STDP mechanism is triggered. This mechanism is defined according to the following rules:

$$\begin{cases} \Delta w_{ij} = a^+ \cdot w_{ij} \cdot (1 - w_{ij}), & \text{if } t_j - t_i \leq 0 \\ \Delta w_{ij} = a^- \cdot w_{ij} \cdot (1 - w_{ij}), & \text{if } t_j - t_i > 0 \end{cases} \tag{3.1}$$

where i and j are respectively post- and pre-synaptic neurons, t_i and t_j are the spike times for the neurons i and j, the synaptic weight adjustment is given by Δw_{ij}, and the change factors for increasing and decreasing the connection strength are, respectively, a^+ and a^-. These two values, a^+ and a^-, are respectively positive and negative values. Furthermore, a^+ and $|a^-|$ are increased as the STDP learning process evolves. The initial value of a^+ is 2^{-6}, and this value is doubled every 400

post-synaptic spikes until it reaches 2^{-2}. The value of a^- is set and proportionally adjusted in order to keep the ratio a^+/a^- fixed at $(-4/3)$. The reason for starting with a very small increase and decrease in the STDP learning process is to avoid learning quickly when the weights are random (in the beginning) which could lead to learning erroneous patterns. In this way, only when the weights have already been modified according to the input stimulus, they will be increased in order to converge quickly to the preferred pattern. The weights are kept in the range $[0, 1]$ because of the term $w_{ij} \cdot (1 - w_{ij})$, which has soft-bound like properties: when the weight value approaches zero or one, the weight adjustment Δw_{ij} tends to zero. One simplification in this model is the fact that the weight adjustment is not proportional to difference in time from the pre- and post-synaptic spikes, despite this is a very important mechanism in the brain, because it allows to distinguish between spikes from different events. Nevertheless, this temporal information is implicit through the spike order. As defined in the formula above, the change in this model is related to the sign of difference $t_j - t_i$, therefore only the order of the spikes is taken into consideration, not the precise timing. This simplification is possible because in this model only one spike per neuron is propagated, and it is assumed that the time to spike is fairly short (in the region of 20-30ms), therefore the decreasing effect in the STDP rule is negligible. There is also a long-term depression to the weights of synapses coming from pre-synaptic neurons that never fire. This mechanism reduces the initial noise introduced by the random weights.

Finally, the last and fourth layer corresponds to the V4/IT region in the brain. Its cells are selective to faces or parts of faces, like in the third layer (of V4 cells), but

in this case there is a degree of spacial and scale invariance. In order to achieve this degree of scale and location invariance, each of the cells in this layer take the maximum response (the first spike) from each of the maps from the previous layer, therefore performing a max-pooling operation.

This model was trained and tested using the face and background datasets from the California Institute of Technology[1]. Both face and background images have a resolution of 896 x 592 pixels, have jpeg format and have been converted to greyscale. The face images are not segmented and contain 27 unique people, with different lighting, expressions and backgrounds. Figure 3.3 shows some examples of both sets of images.

Both sets were split in two parts, one for training and another for testing. These tests showed a very good capacity for learning face features, which is the most relevant feature for the purpose of this thesis. To be more precise, the training was performed by repeatedly presenting half of the face dataset images in random order until the weights converge. Then the system is tested by turning off the STDP mechanism and presenting faces and background images. Finally the classification is based on the number of output neurons that fired. The threshold for the classification is set by running the network on a new set of images of faces and non-faces and selecting a value where the percentage of wrongly accepted images (false acceptance rate) is equal to the percentage of wrongly rejected images (false rejection rate). With this method, 96.5% of the images are correctly classified as face/non face. This figures can be improved by using more advanced classification methods as shown by Masquelier and Thorpe (2007), but throughout this thesis, the first method will

[1]Images are available at http://www.vision.caltech.edu/html-files/archive.html.

be used because its performance is satisfactory and it is more biologically plausible than the alternatives.

3.1.1 Methods

Table 3.1 summarizes the parameters for each of the layers, and table 3.2 contains the convolution kernels used to approximate the Gabor filters with 4 preferred orientations.

STDP parameters

The synaptic connections between V1 Complex and V4 cells are modelled using weights. Initially these weights are randomly generated with mean 0.8 and standard deviation 0.05. During training the weights are adjusted according to a rule similar to STDP:

$$\begin{cases} \Delta w_{ij} = a^+ \cdot w_{ij} \cdot (1 - w_{ij}), & \text{if } t_j - t_i \leq 0 \\ \Delta w_{ij} = a^- \cdot w_{ij} \cdot (1 - w_{ij}), & \text{if } t_j - t_i > 0 \end{cases} \tag{3.2}$$

i and j are respectively post- and pre-synaptic neurons. t_i and t_j are the spike times for the neurons i and j. Δw_{ij} is the synaptic weight adjustment. a^+ and a^- are the change factors for increasing and decreasing the connection strength are, respectively. The initial value of a^+ is 2^{-6}, and this value is doubled every 400 post-synaptic spikes until it reaches 2^{-2}. The value of a^- is set and proportionally adjusted in order to keep the ratio a^+/a^- fixed at $(-4/3)$.

Other Parameters

In this model only one spike per pixel is propagated. The first three layers are processed at 5 different scales: 100%, 71%, 50%, 35%, and 25%. The V4 cells threshold = 1/4 of the number of cells in the receptive field. At the V4 cells layer there is a winner-take-all mechanism among all replicated cells, and also there is a limit to the number of of cells that fire among at each scale.

(a) (b)

(c) (d)

(e) (f)

Figure 3.3: Examples of images used to train the network. Background images such as (a), (c) and (e) were used as negative examples and face images like (b), (d) and (f) were used as positive examples during training. More examples of images from these datasets can be found at http://www.vision.caltech.edu/html-files/archive.html.

Table 3.1: Layers parameters.

Name	Size	Receptive field	Neuron type	Inhibitions	Firing mechanism
Input	640x480;454x341 320x240 ; 224x168;160x120	-	-	-	-
Simple V1	4x640x480;4x454x341 4x320x240;4x224x168;4x160x120	5x5	Spiking	-	Latency inverse to Gabor response
Complex V1	4x107x80;4x76x57 4x53x40;4x38x28;4x27x20	7x7 (1 overlap)	Spiking	Same orientation 11x11 (5-15%)	Latency Max over RF
V4	3x5 scales or 1x5 scales weight sharing among scales	32x32 or 64x64	Spiking	Other prototype cells 8x8 (suppress)	Integrate and fire (no leakage)
V4/IT	3 or 1	All scales and all positions	Spiking	-	Max over RF

Table 3.2: Convolution kernels.

Orientation	Kernel				
	7	9	11	9	7
	1	7	14	13	9
$\pi/8$	-8	-6	0	6	8
	-9	-13	-14	-7	-1
	-7	-9	-11	-9	-7
	-7	-1	8	9	7
	-9	-7	6	13	9
$3\pi/8$	-11	-14	0	14	11
	-9	-13	-6	7	9
	-7	-9	-8	1	7
	-7	-9	-8	1	7
	-9	-13	-6	7	9
$5\pi/8$	-11	-14	0	14	11
	-9	-7	6	13	9
	-7	-1	8	9	7
	-7	-9	-11	-9	-7
	-9	-13	-14	-7	-1
$7\pi/8$	-8	-6	0	6	8
	1	7	14	13	9
	7	9	11	9	7

3.2 Model improvements

We have modified the original model developed by Masquelier and Thorpe (2007) with four main goals:

- To behave as a face detector, by having one output neuron behaving as a grandmother cell
- To have a number of intermediate features large and representative enough to be used for face recognition
- To define a Region of Interest (ROI) corresponding to a face, more precisely, to be able to determine the area of V1 complex cells layer corresponding to face and used ROI for face recognition
- To be able to detect faces at different poses and know which particular pose has been detected

The modifications detailed in this section can be summarized in four points. Firstly, a set of model parameters were modified and the model was re-trained with the new parameters. Secondly, a mechanism to determine the region of V4 complex cells layer has been introduced in order to use this region as input for a face recognition algorithm. Thirdly, several versions of the model have been trained to respond to a particular pose, resulting in a multi-pose detector. Finally, the model (excluding the training phase) has been re-implemented completely using C++ in order to demonstrate its capabilities in real time.

3.2.1 Parameters adjustment

With above mentioned goals in mind, we have modified the parameters of the model in the following manner. First we tested different configurations regarding the number of output V4/IT cells O, and the receptive field size of V4 cells S. In the first case we used the values $O = 1$ and $O = 3$, which means that the output layer will respond to a single face feature (the whole face), or to three different features (different parts of the face or different levels of detail), respectively. The first set-up is the most natural since the output neuron corresponds to the whole face region, therefore can be seen as simple binary face detector, or as a grandmother cell for faces. The second set-up, with three output neurons, was introduced in order to measure the effect of redundancy in the face recognition phase. With the three output neurons we expect to have some redundancy in terms of the area covered by each of the output neurons. This has been observed in our experiments, as shown in Figure 3.4, where different output neurons learn to detect the whole face at different scales and different centres, bringing in this way several levels of detail to the final representation. The effects of these two different representations will be analysed in the following two chapters. Another motivation for using three output neurons comes from the psychological studies, which indicates that we look at parts of the face at a time (Tanaka and Sengco, 1997), which can be represented by the different output neurons, each one corresponding to a face feature such as eyes, nose or any other. Although, as shown in 3.4, this was not observed experimentally in our model. Instead, what we observe is that the three different neurons would represent the face at a different level of detail.

Another modification to the original model was the increase of the receptive field size of the V4 cells, in order to obtain a higher number of intermediate features to be later used for face recognition. The receptive field size we use either 32×32 or 64×64 grid of V1 complex cells, instead of the original 16×16 in order to obtain a more detailed intermediate representation, which is useful for discriminating faces in the face recognition stage. Figure 3.5 shows the weights resulting from the set-up with one output neuron and 32×32 or 64×64 grid of V1 complex cells. Figure 3.6 shows the weights resulting from the setup with three output neurons and 32×32 grid of V1 complex cells. Finally, Figure 3.7 shows the weights resulting from the setup with three output neurons and 64×64 grid of V1 complex cells.

3.2.2 Mechanism for region of interest segmentation

We also introduce a mechanism to determine the regions of V1 complex cells corresponding to activated neurons of the output V4/IT cells layer, i.e., the Region Of Interest (ROI) corresponding to the face, or face features in a case if there is more than one output neuron. This can be seen as a feedback connection that drives the attention mechanism to the face area. The implementation of this mechanism is done by re-using intermediate variables from the original model which are saved in memory and indicate the position in the V1 complex cells map which was ultimately responsible for firing the output neuron. We then added a mechanism to re-calculate the maximum orientation for each position in this region of the V1 complex cells map, and a mechanism to save the winning orientations of the region of interest to disk, which can then be used for further processing, in particular as input for a face recognition algorithm. This improvement of the detection algorithm is crucial for

the integration of the face detection system with the face recognition mechanism, because it allows us to use only responses of V1 complex cells which relate to the face stimulus, instead of using responses from all input neurons, therefore reducing the dimensionality and increasing the shift and scale invariance.

Figure 3.8 shows examples of challenging successfully detected faces in extreme conditions such as bad lighting and side pose. It also shows unsuccessful detections. There are successful and unsuccessful examples for both configurations, with one and three output neurons.

3.2.3 Multi pose face detection

As a result of the analysis of the quality of the coding presented in a later chapter of this thesis, we concluded that the variation of pose is a challenging problem. Therefore we improved the model in order to detect specific poses. This pose specific detection could be useful to better organise the memory of faces and therefore improve the recognition rates.

In order to create the pose-specific face detector, we have retrained 9 different networks, corresponding to the 9 different poses from the Yale Face Database B (Georghiades et al., 2001). Using cardinal directions the poses can be approximately described as:

- Frontal
- North
- North West

- West
- South West
- South
- Pronounced North West
- Pronounced West
- Pronounced South West

For each network we have selected images from a single pose for the positive training set and non-face images for the negative training set. Each V4 cell in all of the networks receives connections from 32×32 cells from the previous layers and the number of output neurons in all of the networks is set to $O = 3$. The resulting training weights for each of the specific poses, can be seen in Figures 3.9, 3.10 and 3.11, where each pose has been clearly learned for each of the 7 three output neurons, respectively. The output neurons from this pose-specific model have parallels with biology, and they have been proposed in the literature as view tuned cells in the top of the hierarchical feedforward network (Riesenhuber and Poggio, 1999).

3.2.4 A demo for face detection, memorisation and recognition

Two applications were developed in C++ in order to demonstrate the detection capabilities of the model presented in this chapter as well as the recognition capabilities of the model when combined with the coding scheme presented in the next chapter. The reason for the two different versions is that the first one uses the full model for

face detection, which requires a considerable amount of computational power, therefore it can run at about 3 frames per second on a desktop PC, therefore a second version was developed which uses the Viola-Jones detection algorithm for the face detection (Jones and Viola, 2001), and uses the features and coding presented in this thesis for recognition. With this combination it is possible to achieve a real time detection and recognition.

Detection and recognition demo

The entire model from Masquelier and Thorpe (2007) and the coding of face features presented in the next chapter has been re-implemented in C++ and OpenCV in order to build a live demo which shows the detection and recognition capabilities of the model and coding scheme. A demo application has been written in C++ and using the QT library. The entire project contains around 4000 lines of code.
The application has three main options:

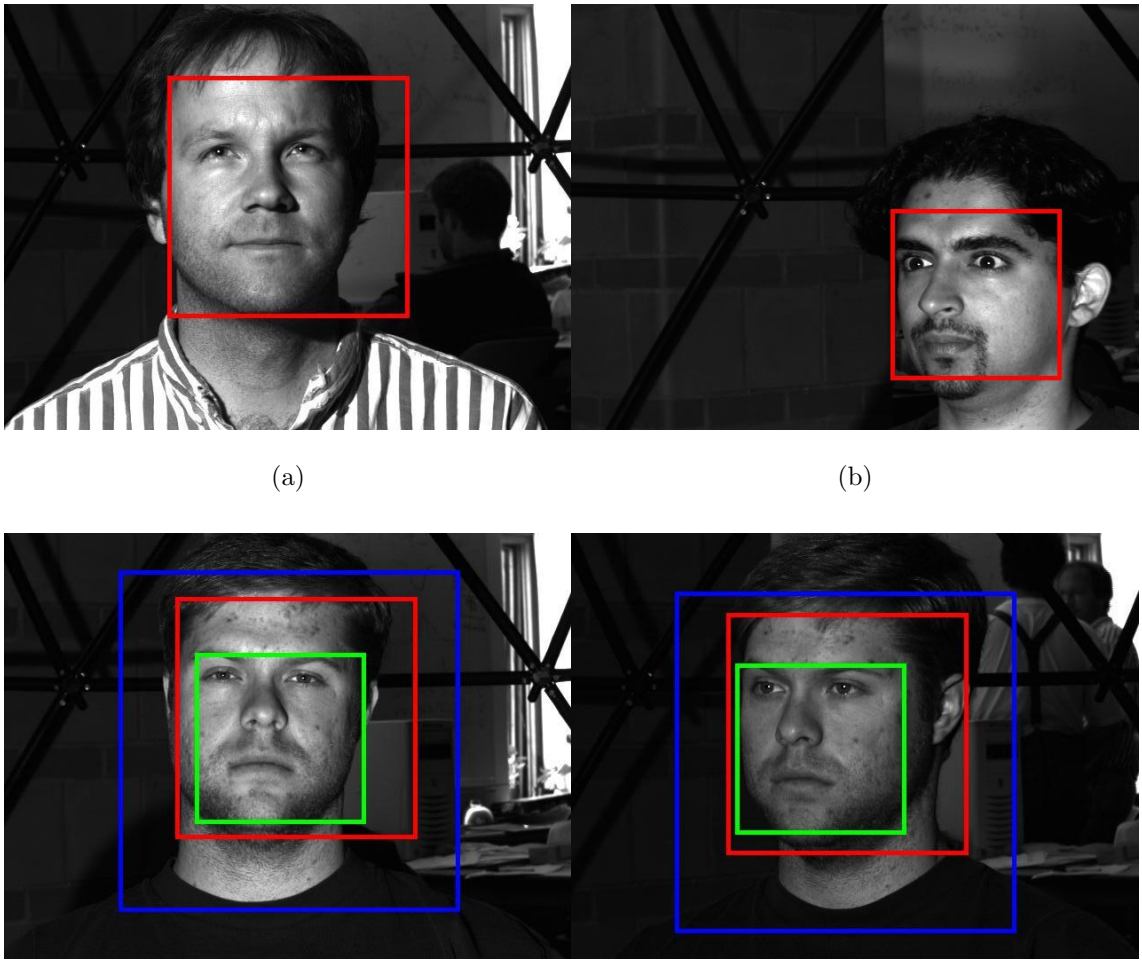
- next image - this option captures another frame from the live video feed, uses the c++ model for face detection and displays the image captured with the face detected (Figure 3.12)

- add to gallery - this option adds the currently detected face to the gallery by using the feature coding scheme presented in this thesis and prompting the user to enter the person name (Figure 3.14)

- identify - this option compares the face currently detected to those saved in the gallery and displays the name of the closest match

Recognition only demo (with third party detection)

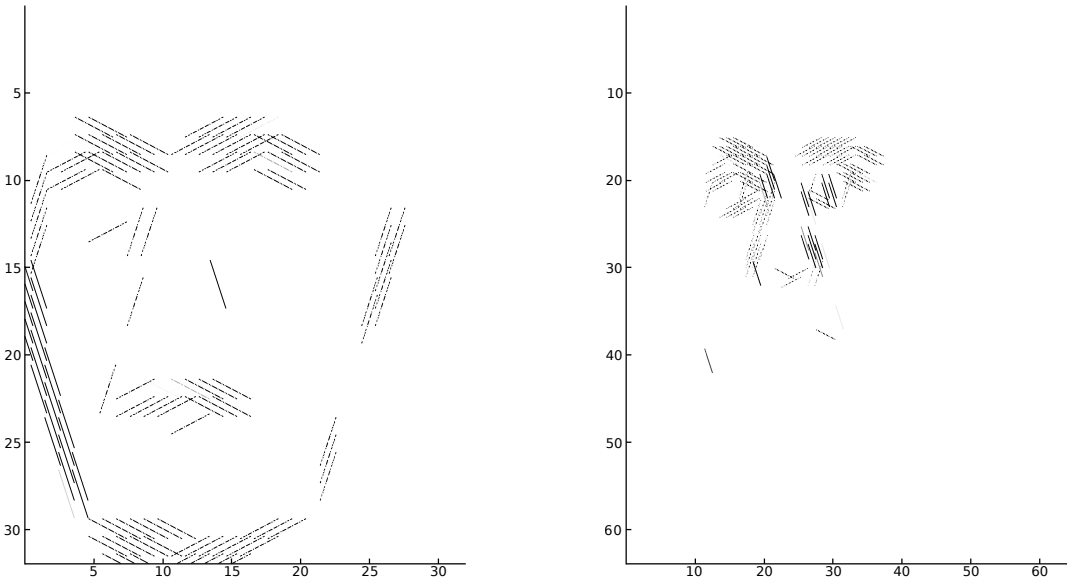
A second version of the application described above has been developed using C++, QT and OpenCV. This version differs from the first one essentially because it uses the OpenCV implementation of the Viola-Jones face detector (Jones and Viola, 2001) for detection. For the recognition part, it uses the features and coding scheme presented in the next chapter. This modification improves the speed of detection, which can be performed in real time using a standard PC. Furthermore a mechanism for multi-image enrolment and verification has been added in order to improve the recognition performance, i.e., when a face is added to the gallery, several images of the face are saved, and not only a single frame as in the previous demo. Also, during the identification process, several images are used for comparison with the gallery. Figure 3.15 shows the face capture process, Figure 3.16 shows the dialogue for adding a new person into the gallery, and Figure 3.17 shows an example of a successfully identified person.



(a) (b)

(c) (d)

Figure 3.4: Examples of successfully detected faces at different poses, sizes and positions. The size of each detection box reflects the corresponding detection scale. (a) Frontal face detected with 32×32 V1 complex cells receptive field, $O = 1$. (b) Side pose face detected with 32×32 V1 complex cells receptive field, $O = 1$. This face is at a non-central position in the image and the size of the face is smaller when compared to (a), which demonstrates a certain degree of position and scale invariance (c) Frontal face detected with 32×32 V1 complex cells receptive field, $O = 3$. (d) Side pose face detected with 32×32 V1 complex cells receptive field, $O = 3$. Face Images from the Yale Face Database B (Georghiades et al., 2001).



(a) (b)

Figure 3.5: Visual representation of the weights between V1 complex cells and V4 cells layers after the training phase. The weights are represented in shades of grey between 0 (white) and 1 (black). The corresponding orientation is drawn as a small line. Weights are shared among different scales. (a) Weights for the network with 32×32 V1 complex cells receptive field, and $O = 1$. (b) Weights for the network with 64×64 V1 complex cells receptive field, and $O = 1$.

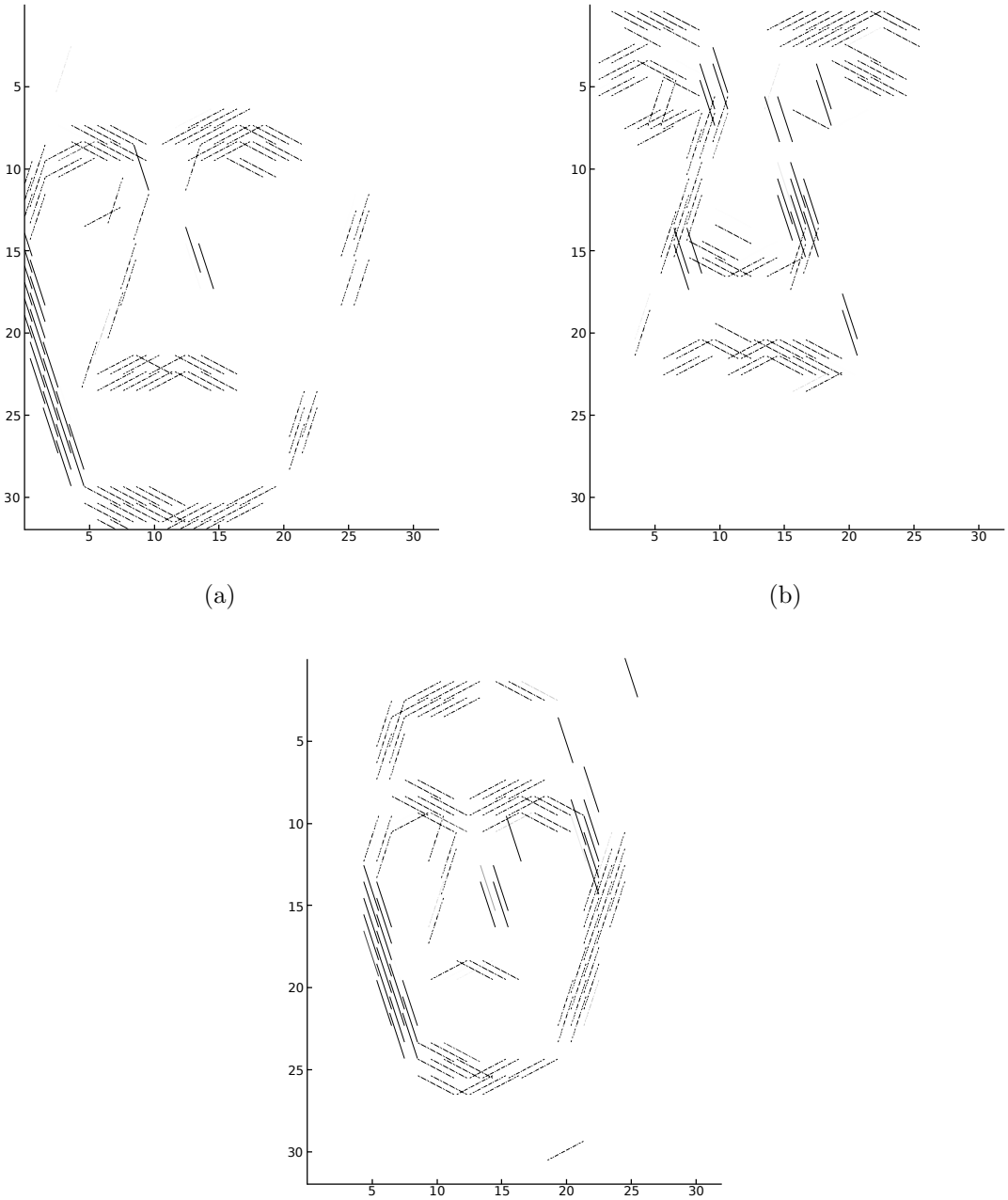
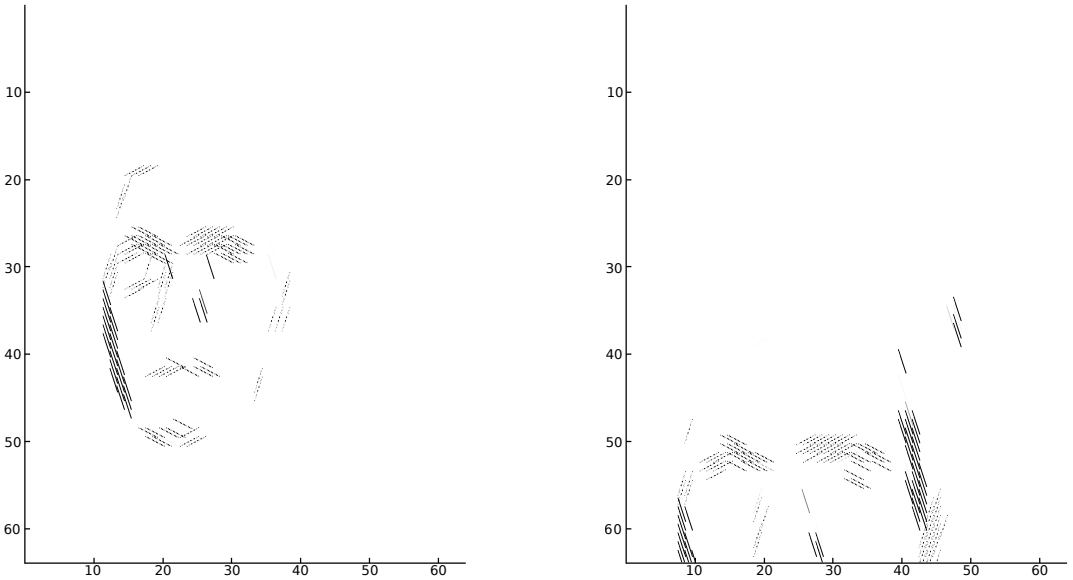


Figure 3.6: Visual representation of the weights between V1 complex cells and V4 cells layers after the training phase, for the network with $O = 3$ and 32×32 V1 complex cells receptive field. The weights are represented in shades of grey between 0 (white) and 1 (black). The corresponding orientation is drawn as a small line. Weights are shared among different scales. (a) Weights for the first V4 cell. (b) Weights for the second V4 cell. (c) Weights for the third V4 cell.

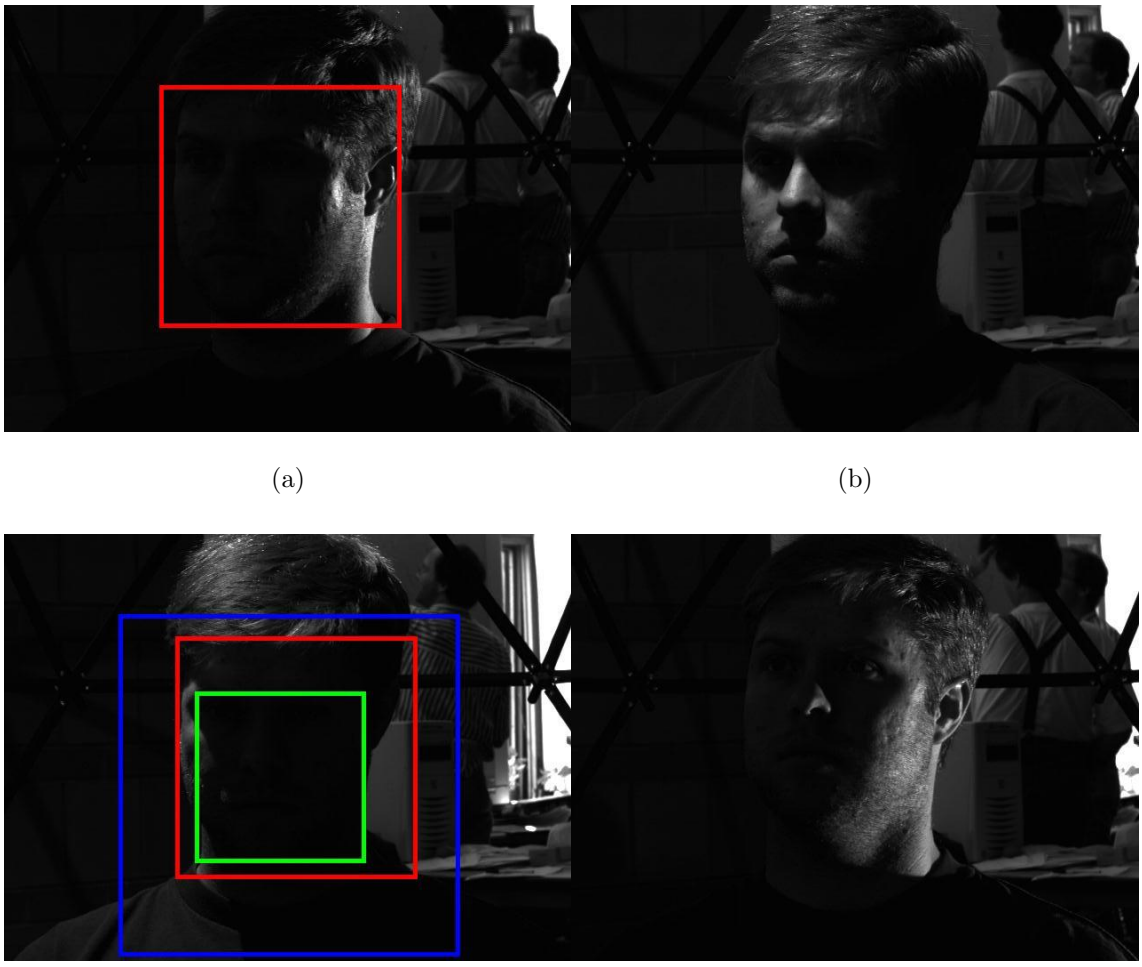


(a)

(b)

(c)

Figure 3.7: Visual representation of the weights between V1 complex cells and V4 cells layers after the training phase, for the network with $O = 3$ and 64×64 V1 complex cells receptive field. The weights are represented in shades of grey between 0 (white) and 1 (black). The corresponding orientation is drawn as a small line. Weights are shared among different scales. (a) Weights for the first V4 cell. (b) Weights for the second V4 cell. (c) Weights for the third V4 cell.



(a) (b)

(c) (d)

Figure 3.8: Examples of successfully and unsuccessfully detected faces at challenging lighting conditions and poses. (a) Side pose with poor lighting face detected with 32×32 V1 complex cells receptive field, $O = 1$. (b) Side pose face with poor lighting not detected with 32×32 V1 complex cells receptive field, $O = 1$. (c) Side pose successfully detected in extreme poor lighting conditions with 32×32 V1 complex cells receptive field, $O = 3$. (d) Side pose face with poor lighting not detected with 32×32 V1 complex cells receptive field, $O = 3$. Face Images from the Yale Face Database B (Georghiades et al., 2001).

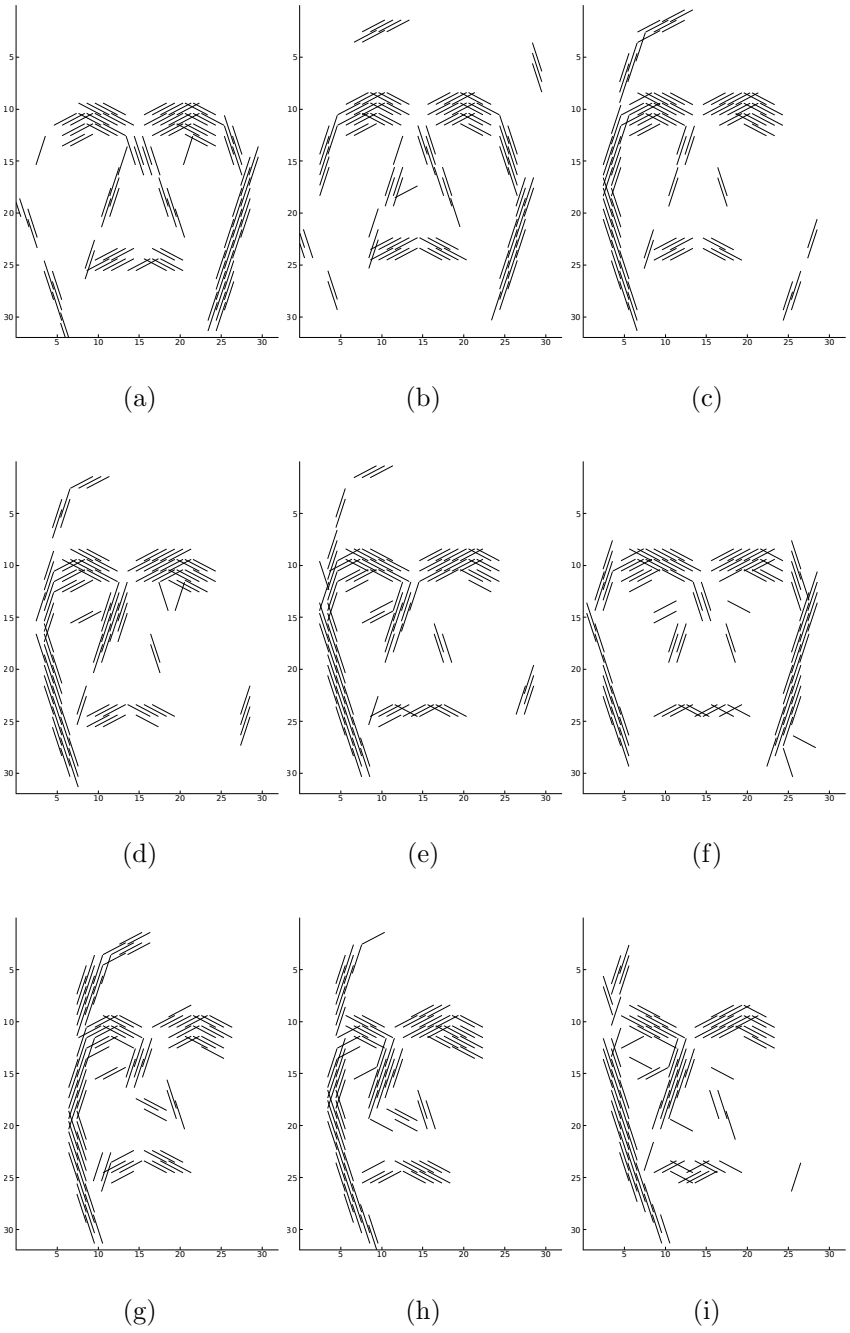


Figure 3.9: Visual representation of the weights between V1 complex and the first neuron of the V4 cells layer after the training phase, for each network corresponding to a particular pose, which has 32×32 V1 complex cells receptive field. (a) Weights for the pose 1. (b) Weights for the pose 2. (c) Weights for the pose 3. (d) Weights for the pose 4. (e) Weights for the pose 5. (f) Weights for the pose 6. (g) Weights for the pose 7. (h) Weights for the pose 8. (i) Weights for the pose 9.

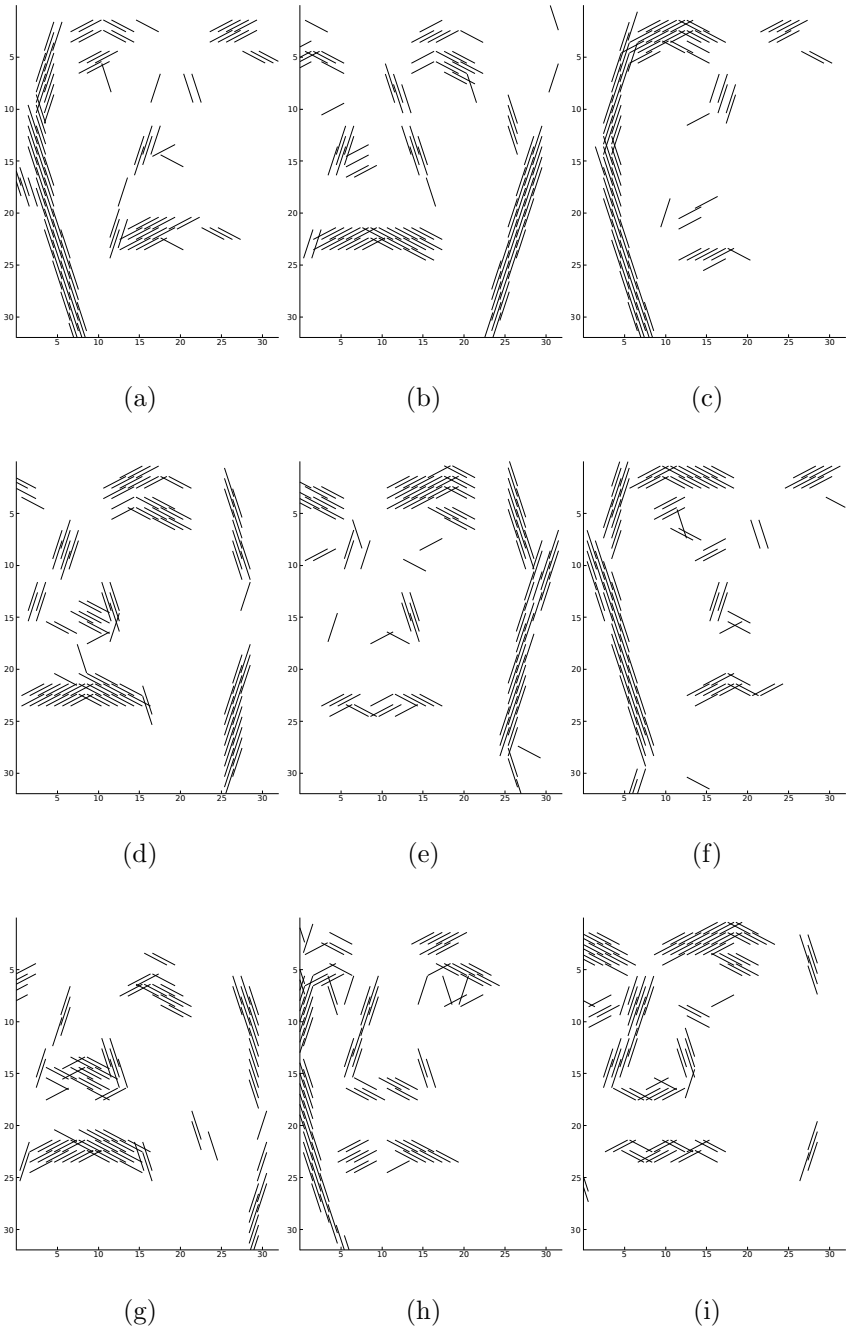
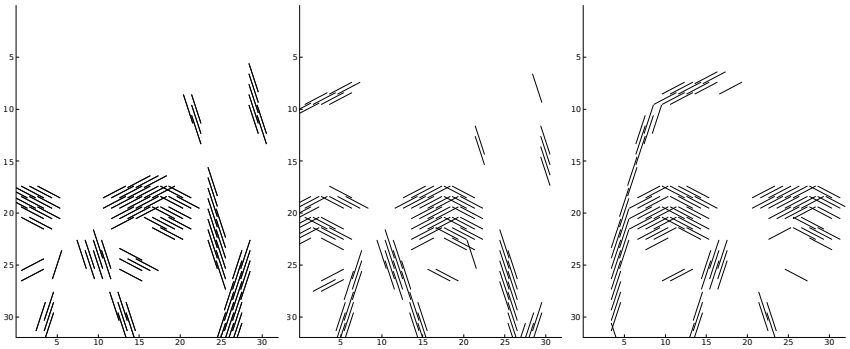
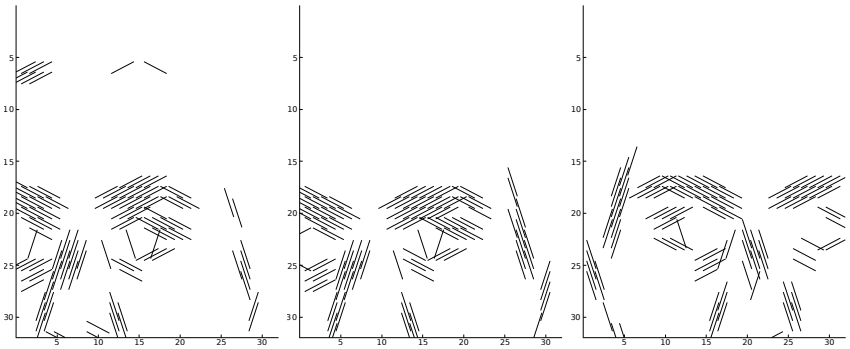


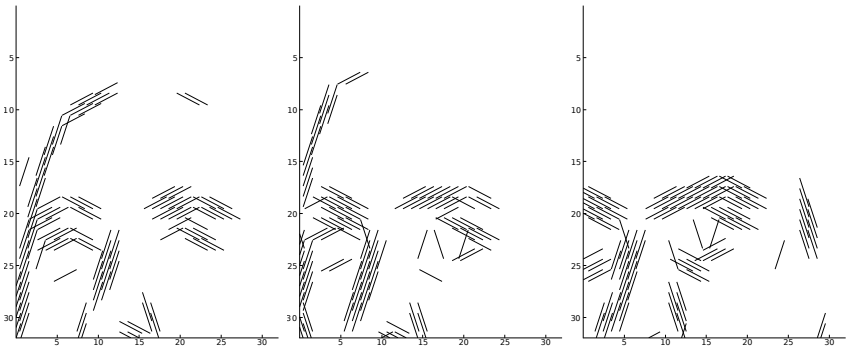
Figure 3.10: Visual representation of the weights between V1 complex and the second neuron of the V4 cells layer after the training phase, for each network corresponding to a particular pose, which has 32×32 V1 complex cells receptive field. (a) Weights for the pose 1. (b) Weights for the pose 2. (c) Weights for the pose 3. (d) Weights for the pose 4. (e) Weights for the pose 5. (f) Weights for the pose 6. (g) Weights for the pose 7. (h) Weights for the pose 8. (i) Weights for the pose 9.



(a) (b) (c)



(d) (e) (f)



(g) (h) (i)

Figure 3.11: Visual representation of the weights between V1 complex and the third neuron of the V4 cells layer after the training phase, for each network corresponding to a particular pose, which has 32×32 V1 complex cells receptive field. (a) Weights for the pose 1. (b) Weights for the pose 2. (c) Weights for the pose 3. (d) Weights for the pose 4. (e) Weights for the pose 5. (f) Weights for the pose 6. (g) Weights for the pose 7. (h) Weights for the pose 8. (i) Weights for the pose 9.

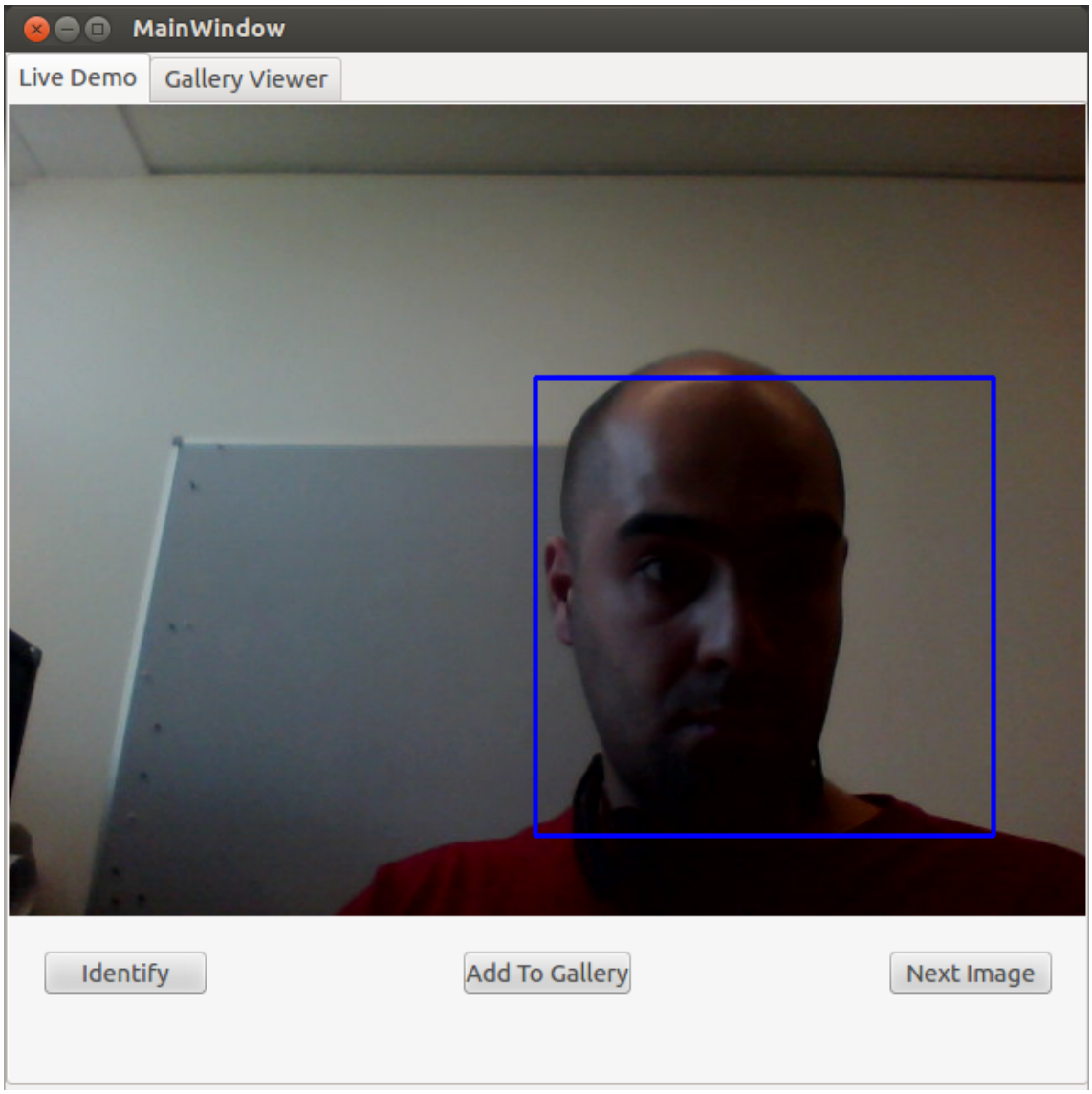


Figure 3.12: Face detected after pressing ”next image”.

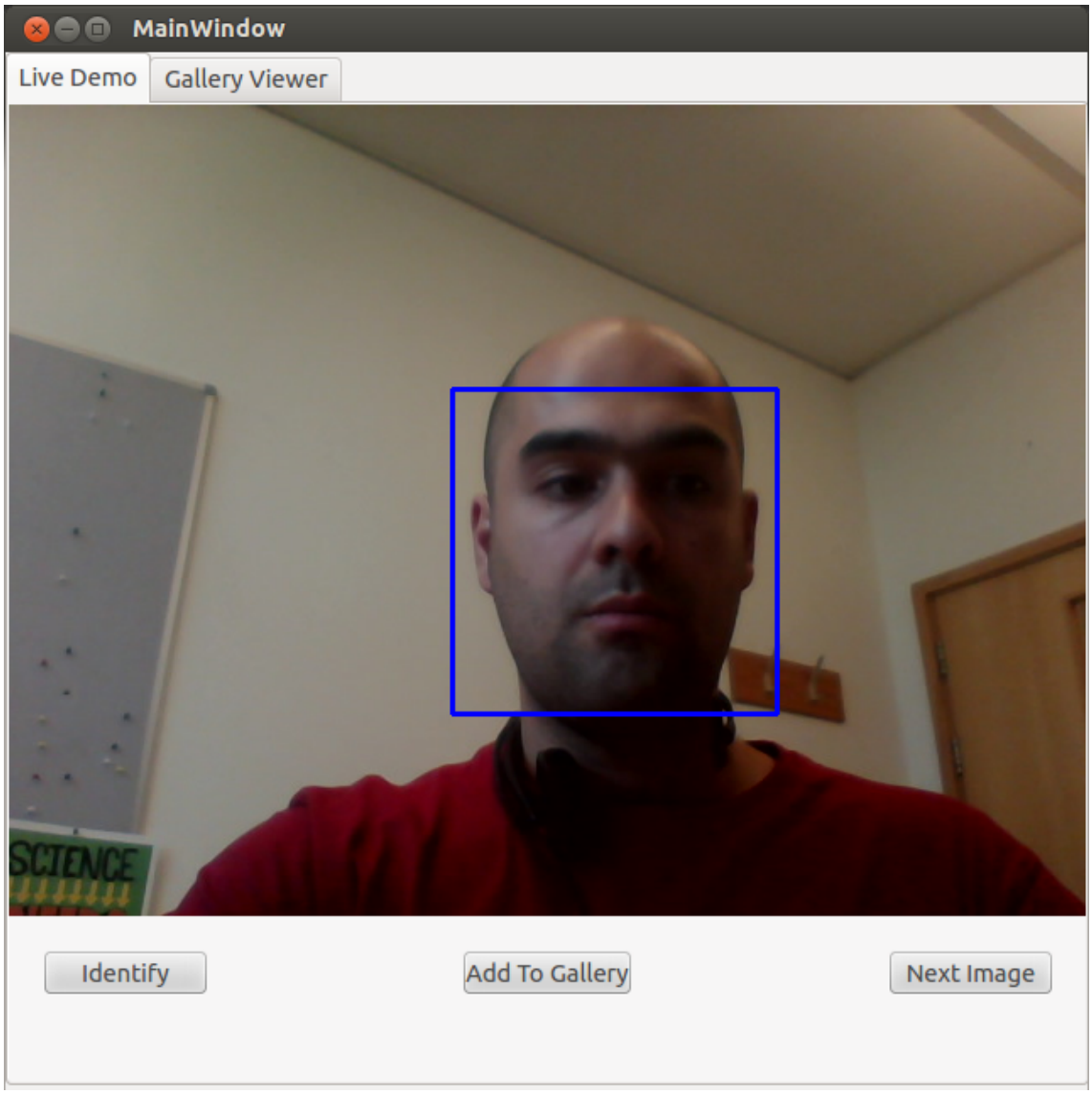


Figure 3.13: Another face detected after pressing "next image".

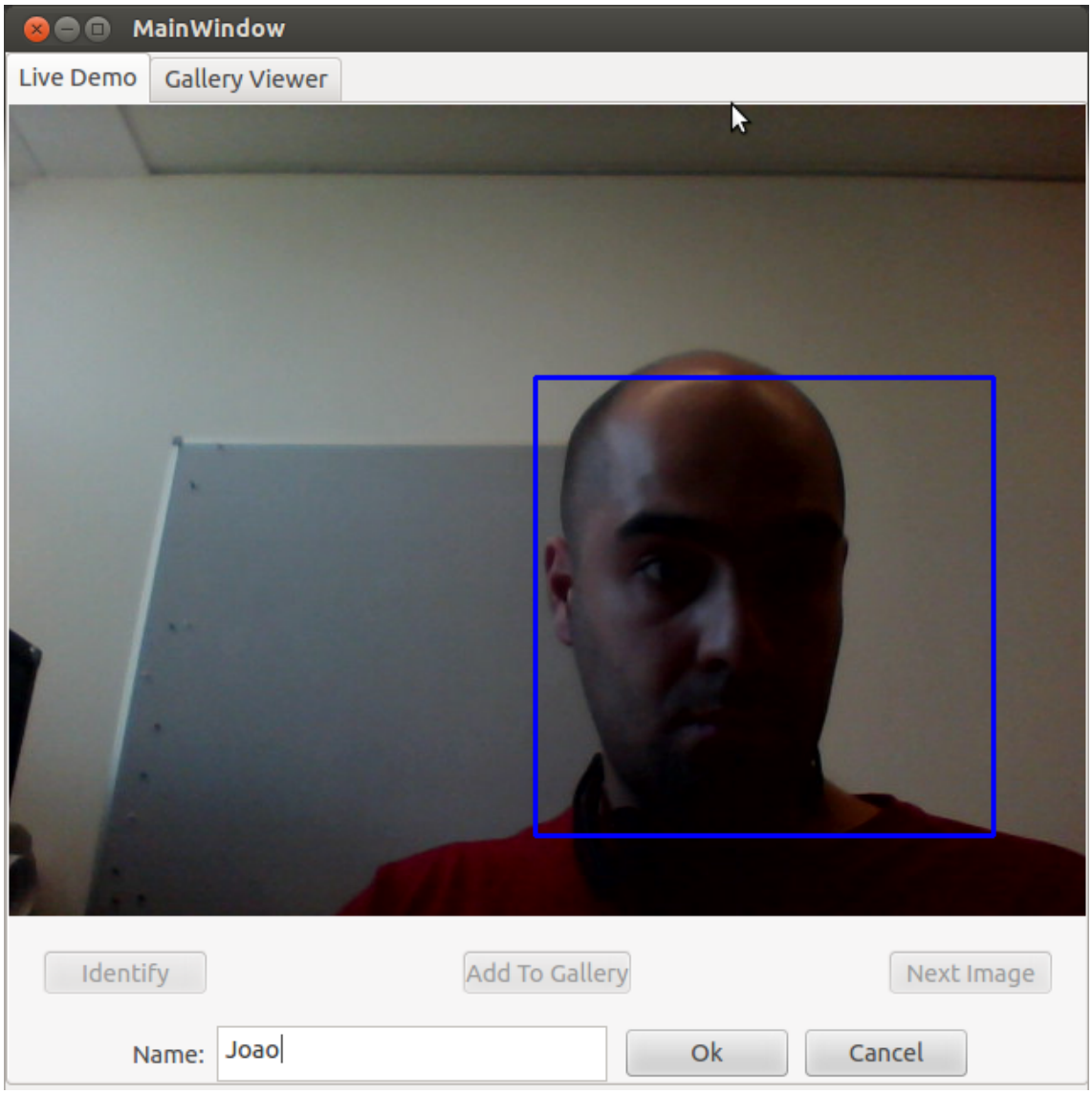


Figure 3.14: Adding face to the gallery for posterior identification.

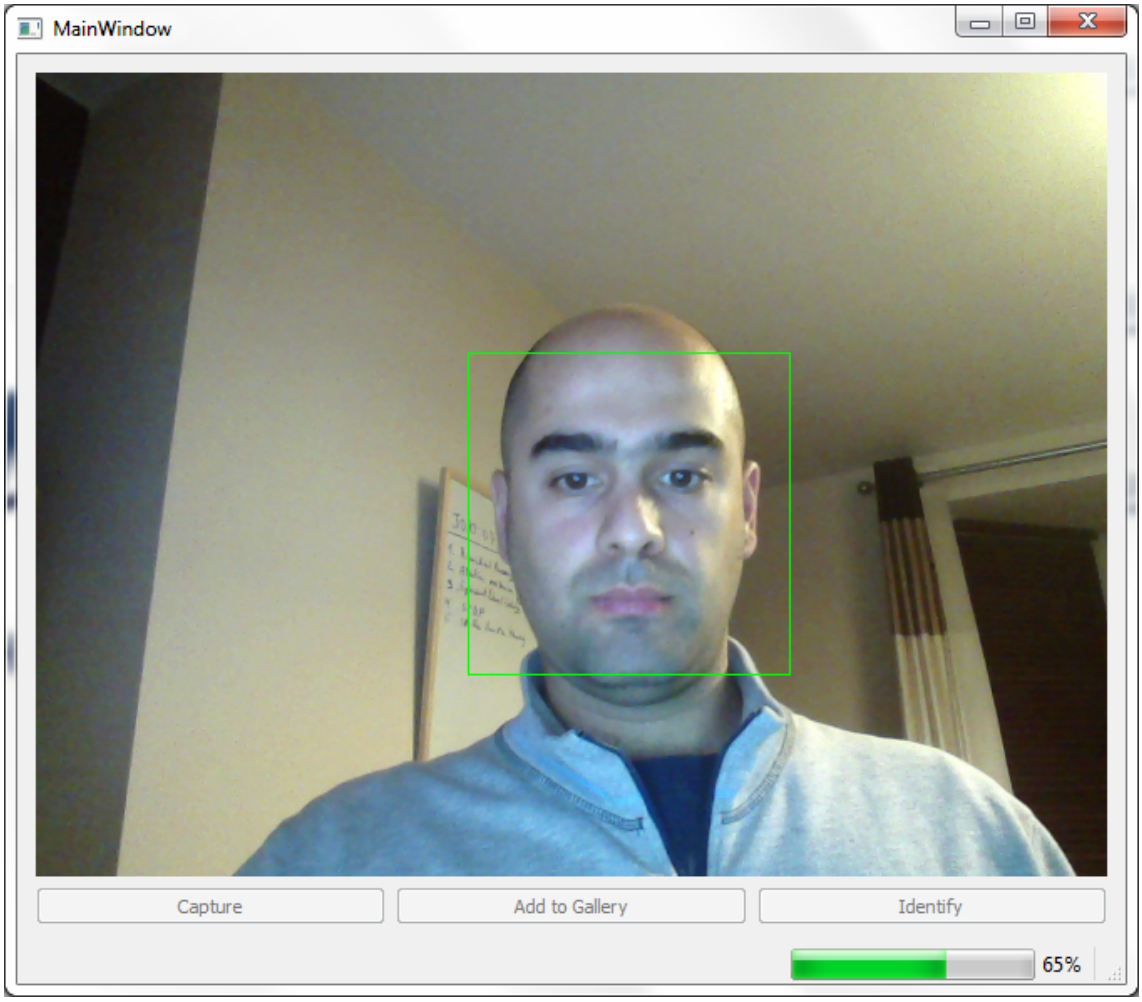


Figure 3.15: Example of live face detection.

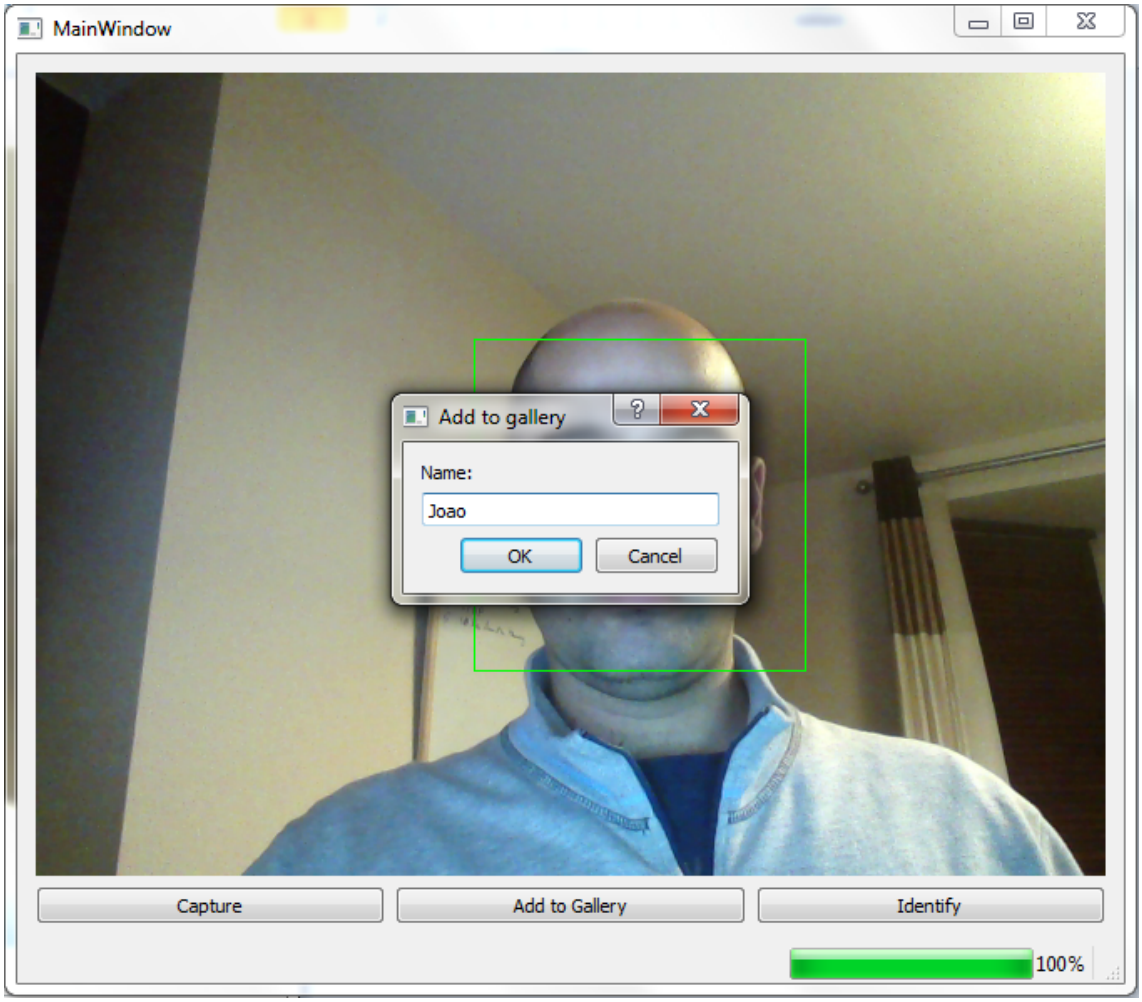


Figure 3.16: Example of enrolment. The currently detected face is added to the local gallery of faces.

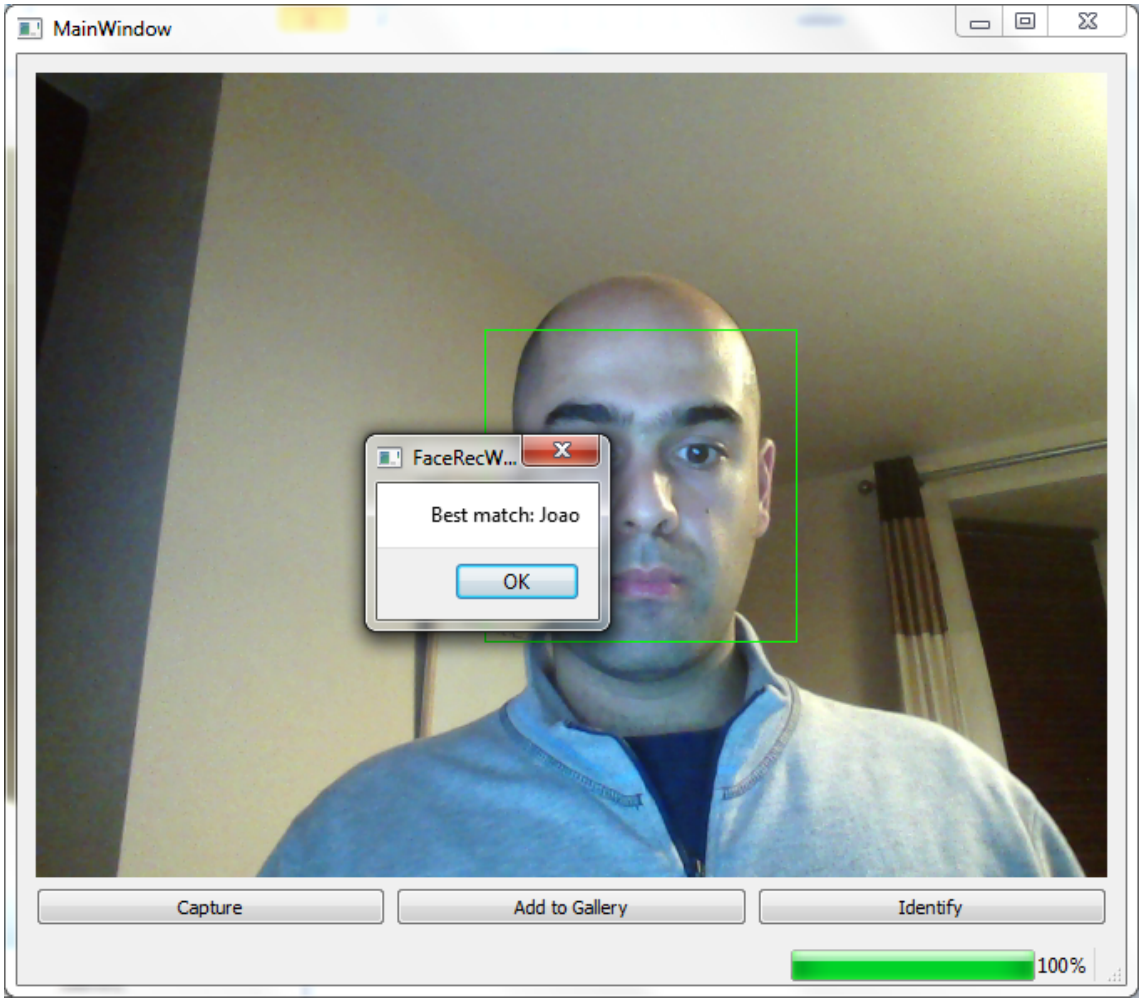


Figure 3.17: Example of live identification. The face detected is matched against the local gallery of faces and the label of the best match is displayed.

3.3 Conclusions

In this chapter a model inspired by the hierarchical structure of the visual system capable of learning face contours was presented. This model is in line with the biological findings regarding the visual system and also the state of the art face recognition and detection algorithms based in the deep convolutional networks, which are a hot topic at the current moment (Sun et al., 2014; Taigman et al., 2014).
The model was improved in order to be used as a face detector and suitable for integration in a face recognition system. The modifications and main contributions of this chapter can be summarised as follows:

- Show that the modified model is able to detect faces using a single output neuron performing as a grandmother cell.
- Retrain the model with three output neurons which leads to a smaller representation when compared to the ten neurons used in the original model from Masquelier and Thorpe (2007), but keeps some of the redundancy and different level of detail.
- Retrain the model where the level of detail of the intermediate features is higher, by using 32×32 and 64×64 as the receptive field of the V4 cells, which is useful to discriminate faces in the face recognition stage.
- A feedback mechanism that connects the V4/IT cells to the V1 complex cells, which in fact highlights the region of interest in the V1 complex cells layer that was responsible for firing the output cells, which makes the improved model a real face detector, since it can determine the location of the face, and not only

if there is a face or not.

- A C++ application based in this model that is a near real time web-cam live face detector with face recognition capabilities.

In the next chapter a coding for the Gabor-like features extracted by the face recognition model introduced in this chapter is presented.

Chapter 4

Coding of face features

In the previous chapter, the process of detecting a face was described in terms of the brain mechanisms and the computer algorithms involved. A model that implements these mechanisms and performs face detection was presented. It was shown the importance of the face detection mechanism in the face recognition process, i.e., every artificial or biological face recognition process has to be associated with a face detection mechanism in order to determine the region of the image or the cortical region corresponding to the face features (Kanwisher and Yovel, 2006; Tanaka and Gordon, 2011; Tsao and Livingstone, 2009). Therefore we have modified the original model in order better integrate the face detection process in a broader face recognition framework.

Now that we have a mechanism for detecting faces and extract low-level features from those faces, a method for coding such features relevant to face processing is needed. In this chapter we present the coding chosen to represent the features in this thesis. Two alternative coding schemes are proposed, one that uses for bits

to encode each feature (orientation) and another that uses two bits. In the next chapter, an analysis of the best combination of coding scheme and model parameters is presented.

The topic of feature representation and coding has been studied by experts in neuroscience, psychology and computer vision. The main questions pursued are which features better represent the face, which ones are actually used by the brain in order to represent the face, and, finally, which ones are more discriminative, in order to enable the brain to tell faces apart and ultimately recognise faces (Tanaka and Farah, 1993).

We can divide the features in two groups. Firstly we have the high level features which have a more direct connection to the human perception of human faces. These features can be seen as face regions or face parts such as eyes, nose, ears, chin, eyebrows, among others. Not always correspond these regions uniquely to a single part but they can be a more abstract region of the face covering one or more parts, or they can be a section of a part.

We are now going to explain the meaning of high level features in each of the fields of study.

From a psychological point of view, the face recognition process involves consideration of different face features. The process by which we look at different features starts with several eye saccades that result in changes of our gaze direction and, therefore, results in looking at different features or regions of the face (Yarbus, 1967). These regions correspond to the above mentioned high level features.

In neuroscience the high level features are linked to a well-studied mechanism of

saliency maps. The saliency maps are in fact cortical regions that contain sub regions that fire more actively than others, according to the input stimulus. The most active regions are therefore the most salient, hence the name of saliency maps (Soltani and Koch, 2010). In the case of faces, the idea is that certain regions or features are therefore more salient in terms of the response in the the primary visual cortex. Then, these most salient regions will serve as input for the object detection mechanism.

Finally, in computer vision, the high level features are broadly used and have been used mainly in algorithms that are non-holistic, i.e., use only a particular sub-set of the data available for a given face (Wiskott et al., 1997). Non-holistic methods rely on certain features or sections of the face to generate representations around them and then perform classification. But sometimes the actual representation of the face is holistic, in the sense that it uses information from all the pixels covering the face region, but a non-holistic method relying on face features is used as a preceding stage, where the alignment and sometimes normalisation of the face is done based on such features. In both cases the high level features are present, and represent usually facial landmarks such as tip of the nose, eye corners, etc (Heisele et al., 2003; Wiskott et al., 1997).

The second kind of features are the low-level features. These are mostly studied and used in neuroscience and computer vision, and not so much of interest for the psychology community (Cox and Pinto, 2011; Hubel and Wiesel, 1959). Low-level features are fine-grade representations of very small areas of the face, therefore contain much less information than the high level features and usually are much less

invariant. Nevertheless, their importance is critical, because by combining a large number of such small features leads to a very informative representation.

In neuroscience these local features have been broadly studied, and are detected in the primary visual cortex. They are essentially oriented bars, which when combined generate more and more complex features, which is believed to lead to the high level features such as face features or whole faces. Therefore, not only the individual face features are important when combined together, but they are also an essential part of the hierarchical process that takes part in the primary visual cortex.

In computer vision, these local features are also the basic element of most detection algorithms. There are many different representations of such simple features, from the most basic pixel intensity values, and the more complex Gabor filters, which have properties similar to the orientation selective simple cells in the primary visual cortex, to more artificial binary representations like local binary patterns, which have been proven to be very effective for the face recognition task (Ahonen et al., 2004, 2006). As in the brain, these artificial local simple features are combined together to generate a higher level representation and further classification of faces.

The high level features are represented in our model by a single V4/IT neuron that has a receptive field covering the whole face, or alternatively, three neurons covering different parts of the face, with different levels of detail. These features are not discriminative enough to differentiate different faces, therefore we will focus on the low level features in order to create a representation of the face suitable for face recognition. Nevertheless, let's not forget that the high level features have two very important roles in the whole system. First, they are ultimately face detectors, or

face-part detectors in case of three output neurons. These detectors are an essential part of a complete face recognition mechanism. Secondly, in the case of the three output neurons, the three slightly different regions represented by those neurons can be seen as result of the movement of the gaze direction over several saccades, which results in analysing different parts of the face with different levels of detail.

From the face detectors we have introduced a feedback mechanism that goes back to the V1 complex cells in order to determine the region of interest corresponding to the face, i.e., in order to highlight the low level features representing a particular face. This region is the region of interest which will be used for the face recognition stage.

The reason for choosing the edge detectors represented by the V1 complex cells for the low level features, are explained in the next section, together with their details. Once the features are known, an important question in neuroscience and computer vision is how to make sense of such features, and how are they coded in the brain or in a binary representation in the computer memory.

The neural coding problem, and how to efficiently encode features in a information theory perspective have been important topics of discussion in both communities, therefore we approach this problem in the third section of this chapter. In the same section we also present two alternative coding schemes for the face features used in this thesis.

4.1 Face features

The low level features chosen to represent faces in this thesis are the responses of the V1 complex cells, which respond to oriented edges in the visible light spectrum. This choice was based in the biological plausibility and extensive studies about these features, which indicate a broad consensus that these cells are selective to oriented bars, and they are very likely the first stage of a hierarchical visual system, capable of processing and analysing visual scenes. Furthermore, these neurons' behaviour have some very interesting properties that are important for achieving a robust recognition system. In particular, these oriented edge cells, implemented as Gabor filters, have a good degree of lighting invariance, i.e., small changes in the light or colour of the input stimulus would have no effect, or a very small effect in the response of the cell. Since we are using complex cells, we can achieve a certain degree of invariance regarding small local shifts, because these cells integrate the maximum response over a small receptive field of the V1 simple cells.

In the brain and in computer algorithms, there is a particular coding for features. The coding is a complex and well-studied topic. In the next section we will explain the coding used in this thesis, and in the next chapter we will analyse the suitability of this features and coding for face recognition.

4.2 Face feature coding

One important question in the research of visual recognition algorithms and brain mechanism is the coding used to represent features.

In neuroscience the big question is how the brain codes the neural responses, in other words, how to make sense of the huge number of cell activity that takes place in an interconnected, highly dynamical brain. There are two main interrelated ways of analysing this activity that are widely used by the neuroscience community. The first way is to analyse the brain activity in terms of firing neurons. The event of firing a neuron occurs when a quick variation of the electrical potential of the cell rises and falls in a standard manner. This event happens when the incoming potential reaches a certain threshold. This firing can be triggered by some event, and can be seen as a binary response (fire/non-fire). The second kind is the time, in particular, the relative time in which two spikes take place. This relation between spikes in two neurons will influence their synaptic connections. As we have seen in the previous chapter, the entire face detection model presented in this thesis relies on the time, and consequently in the firing of individual neurons as well. The remaining question is it how to represent the features from the V1 complex cells layer, which were chosen for the face recognition task, in order to achieve an efficient and informative representation of the face.

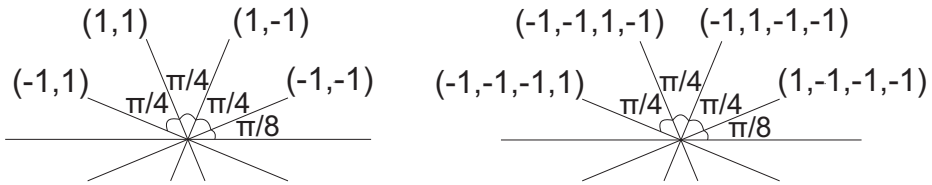
In computer vision a coding process usually takes place in order to transform the basic features into some representation which is mode suitable for the recognition task. The basic features can be pixel values, Gabor filter responses, or of a different nature, but what they have in common is that when all the individual features that cover the whole face region are combined, the resulting feature vector is of very high dimensionality. Therefore the coding method has often the goal of reducing the dimensionality and increasing the inter-subject separability while keeping or

reducing the intra-subject distance.

We took into consideration both computer vision and neuroscience perspectives when looking for the coding for the edge oriented features from the V1 complex cells layer. Only the area corresponding to the face stimulus is coded, which is the area determined by the propagation of the activity from the output V4/IT layer to the V1 complex cells layer. These V1 complex cells respond to different bar orientations. The orientation with the strongest response at each position of the grid is then coded in a vector of binary values in order to reduce the dimensionality, keeping the information relative to the neurons that fired. Thus, a winner take all approach is used to prescribe the winning orientation at each pixel. Two coding schemes are considered: coding a single orientation by 2 or 4 binary values (see Figure 4.1).

The binary encoded orientations for each position are concatenated per row in order to form the feature vector representing a face (see Figure 4.2).

The first coding scheme with 4 bits correspond to individual binary neurons, i.e., they translate directly the activity of the oriented edges neurons. In this scheme only the bit corresponding to the preferred orientation fires. The other scheme that uses two neurons takes advantage of the fact that only one neuron fires, because of the winner-take-all approach, therefore it can encode the firing information of four cells with only two. This gives a 50% dimensionality reduction. The actual encoding is -1 -1, 1 -1, 1 1, and -1 1 for the orientations $\pi/8$, $3\pi/8$, $5\pi/8$, and $7\pi/8$ respectively. Also the closer orientations have a closer Hamming distance, while further orientations have a higher distance. This is an important characteristic for



(a) (b)

Figure 4.1: (a) First coding scheme: two binary values are used to encode a particular orientation. The values chosen are such that immediate neighbour orientations (for instance $\pi/8$ and $3\pi/8$) differ only in one of the values, while the number of different values between binary representations of orientations that are not neighbour is two. (b) Second coding scheme: 4 binary values are used to encode a particular orientation. In this case only one of the 4 binary values can be 1 at a time, defining the orientation being coded. In this case the number of different values between any two orientations is two.

a recognition algorithm, since it discriminates different features by setting them apart in the feature space and and keeps similar features closer to each other. Table 4.1 illustrates the two-bit coding for each orientation and the hamming distances between orientations. The choice for -1 as non-spike and 1 as spike value instead of 0 and 1 is related to the way the feature vectors are compared in an efficient manner. The details of the comparison method are presented in the next chapter.

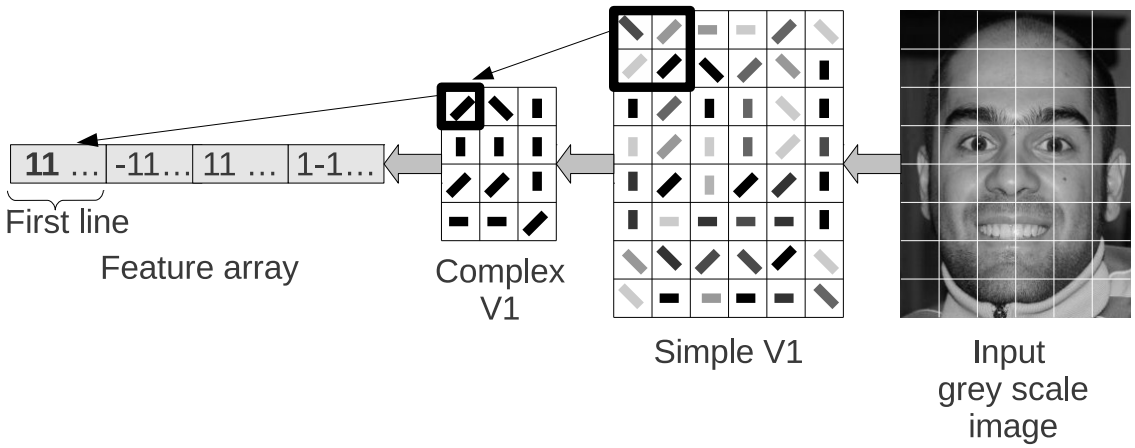


Figure 4.2: Feature extraction and coding process. The area shown in all layers corresponds to the region of interest and the scale determined by one V4/PIT cell, through the feedback process. *Simple V1* shows only the direction with highest value for each position (darker directions have higher values). *Complex V1* shows the winner orientations from the *Simple V1 layer*, which are coded and concatenated to form the feature array.

Table 4.1: Hamming distances between different binary coded orientations with two bits per orientation.

	$\pi/8$ (-1 -1)	$3\pi/8$ (1 -1)	$5\pi/8$ (1 1)	$7\pi/8$ (-1 1)
$\pi/8$ (-1 -1)	0	1	2	1
$3\pi/8$ (1 -1)	1	0	1	2
$5\pi/8$ (1 1)	2	1	0	1
$7\pi/8$ (-1 1)	1	2	1	0

4.3 Conclusions

In this chapter we presented the features selected for representing the face, which are based in Gabor filters. We also present two alternative coding schemes for those features as hypothesis to be tested in the next chapter.

Both the features (responses to oriented edges from the V1 complex cells layer) and the binary coding schemes are biologically plausible. Also, the coding scheme has very interesting properties that are desirable when building a complex recognition system, since in one of the variants they automatically reduce the dimensionality by half and separate the basic features nicely in the feature space.

The contributions from this chapter are summarised in the following list:

- A feature coding scheme for the oriented edges that uses 4 bits to represent directly the activity of the V1 complex cells.
- A feature coding scheme for the oriented edges that uses 2 bits to represent the winning orientation and has interesting properties in terms of building a face recognition system, such as class separability and dimensionality reduction.

In the next chapter an analysis of the quality of the current coding regarding its suitability for separating different individuals, while keeping close different representations of the same individual in the high dimensionality feature space is presented.

Chapter 5

Analysis of the quality of coding

In the previous chapter we have presented two alternative coding schemes for the features extracted and selected using the model introduced in chapter 3. One coding scheme codes the one of four possible orientations at each C1 complex cells map using 4 bits (only one is active), and the second alternative coding scheme uses only two bits to code the winning orientation. Both alternatives are biologically inspired. The first coding scheme implements a winner-take-all mechanism, which has been broadly used in brain inspired models (Fukushima, 1988; Marr and Poggio, 1976). The second coding scheme also have some similarities with brain mechanisms because the codes used for orientations that are close to each other are closer in terms of hamming distance than those codes used for orientations that are further apart from each other. In this chapter we analyse the quality of the coding schemes combined with different configurations of the face detection and feature extractor model. The quality of the code is analysed regarding the suitability of such coding for face recognition, and compared with other approaches in the same test and

training data, in order to determine if this coding is discriminative enough for a face recognition system.

The goals of this analysis can be summarised as follows:

- Determine if the feature vectors resulting from the proposed coding are discriminative, i.e., if the feature vectors from different subjects have a larger distance between them than the distance between feature vectors from the same subject.
- Analyse the performance of this coding scheme with regards to variations in the lighting conditions
- Analyse the performance of this coding scheme with regards to variations in the pose of the subject
- Analyse the effects of having in the output one grandmother cell or three output neurons corresponding to different levels of detail of the face in the overall performance of algorithm
- Analyse the effects that the variation of the receptive field of the V4 neurons have in the overall performance of the algorithm
- Determine which alternative coding achieves the best recognition performance: 2- or 4-bits coding given that the 2-bits coding separates better individual features

In the next sections the test results and set-up which was used to access the quality of the code are presented.

5.1 Geometry of multidimensional coding space

Here we investigate the geometrical structure of multidimensional space of vectors (V) representing faces using the coding scheme presented in the previous chapter. We expect that a set of vectors corresponding to faces of the same person under the variation of pose and illumination can be clustered together and has no or small overlap with a set of vectors corresponding to faces of another person. To test this hypothesis we study the geometrical properties of the vector space for different schemes of face representation. The scheme of face representation by two binary values coding the orientation in the grid 32×32 provides coding vectors of length 2,048 (the smallest dimension of the vector space). The highest dimension (49,152) is provided by a scheme with four binary values coding the orientation in the grid 64×64, and there are three such grids corresponding to three output neurons.

5.1.1 Database of face images

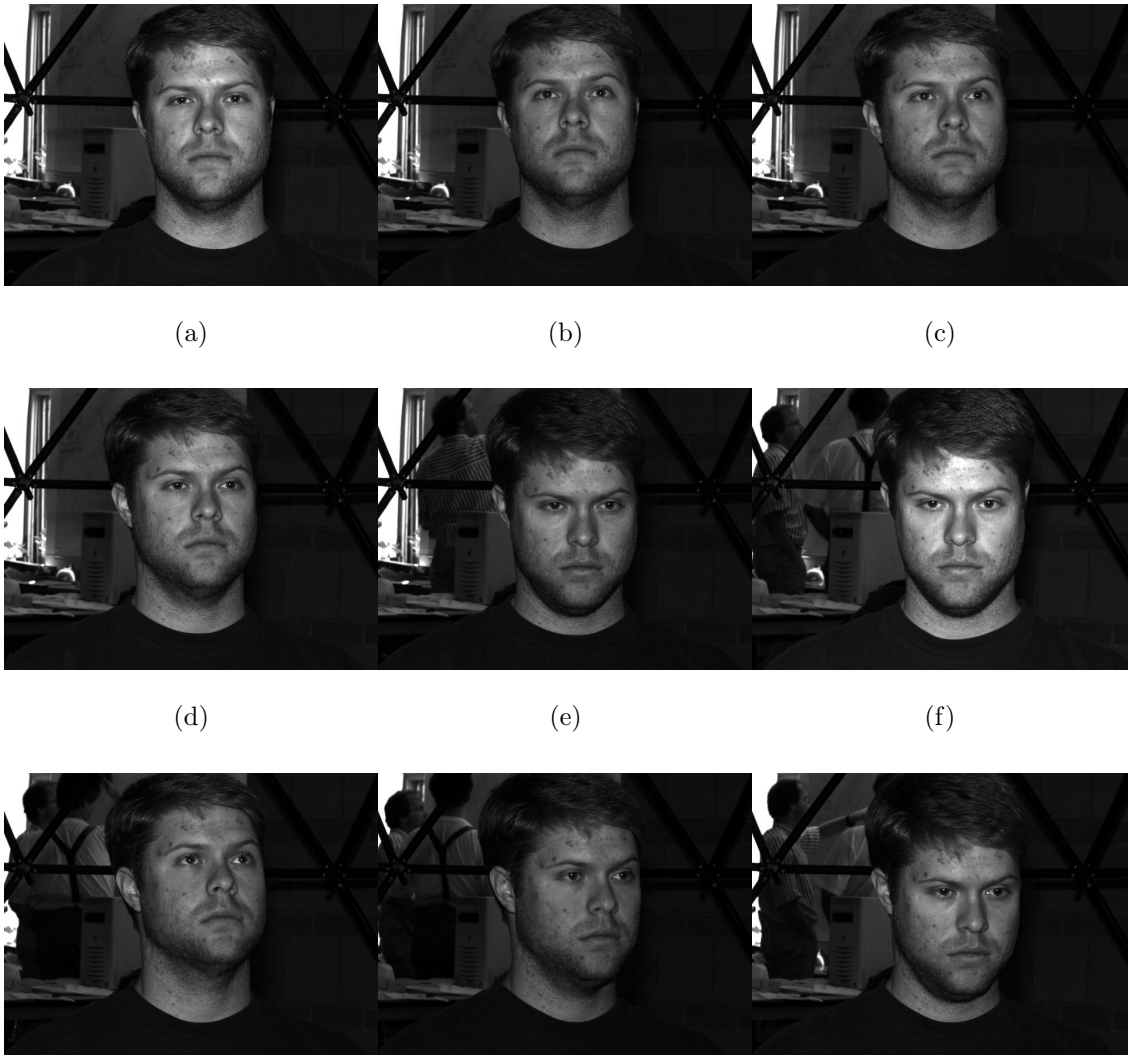
The database chosen to investigate the spaces V was the Yale Face Database B (Georghiades et al., 2001). This database contains a large variation of pose and illumination conditions for each face. The database has 5,696 images of 10 subjects. Figure 5.1 shows all the subjects in the database. For each subject there are 64 images with different illumination settings (see Figure 5.2) for every of the 9 different poses (see Figure 5.3). Therefore there is a total of 576 images per subject, except one of the subjects which only has 512.

Figure 5.1: Ten subjects from the Yale Face Database B (Georghiades et al., 2001).

5.1.2 Procedure to compare feature vectors

Given any subset of images from the database, for instance, any of the training subsets presented in the next section, the matrix G is constructed. We will call this subset the gallery. Each row of G contains a binary representation of the winning orientations of the V1 complex cells for a given face. The ROI used is determined by the face finding algorithm presented in chapter 3. The matrix G is of size $M_G \times N$, where M_G is the number of images in the gallery and N is the length of the binary vector representing a face. In a similar fashion, a matrix R is constructed based on images which are not included in the gallery. The number of coinciding components is a similarity measure for comparison of two face representation vectors (respectively, the distance between two vectors is the number of non-coinciding components). Let us assume that vector y does not belong to the gallery. To compare this vector with gallery vectors, we use the following formula:

$$z = Gy^T \tag{5.1}$$



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.2: 9 poses per subject from the Yale Face Database B (Georghiades et al., 2001). (a) Frontal. (b) North. (c) North West. (d) West. (e) South West. (f) South. (g) Pronounced North West. (h) Pronounced West. (i) Pronounced South West.

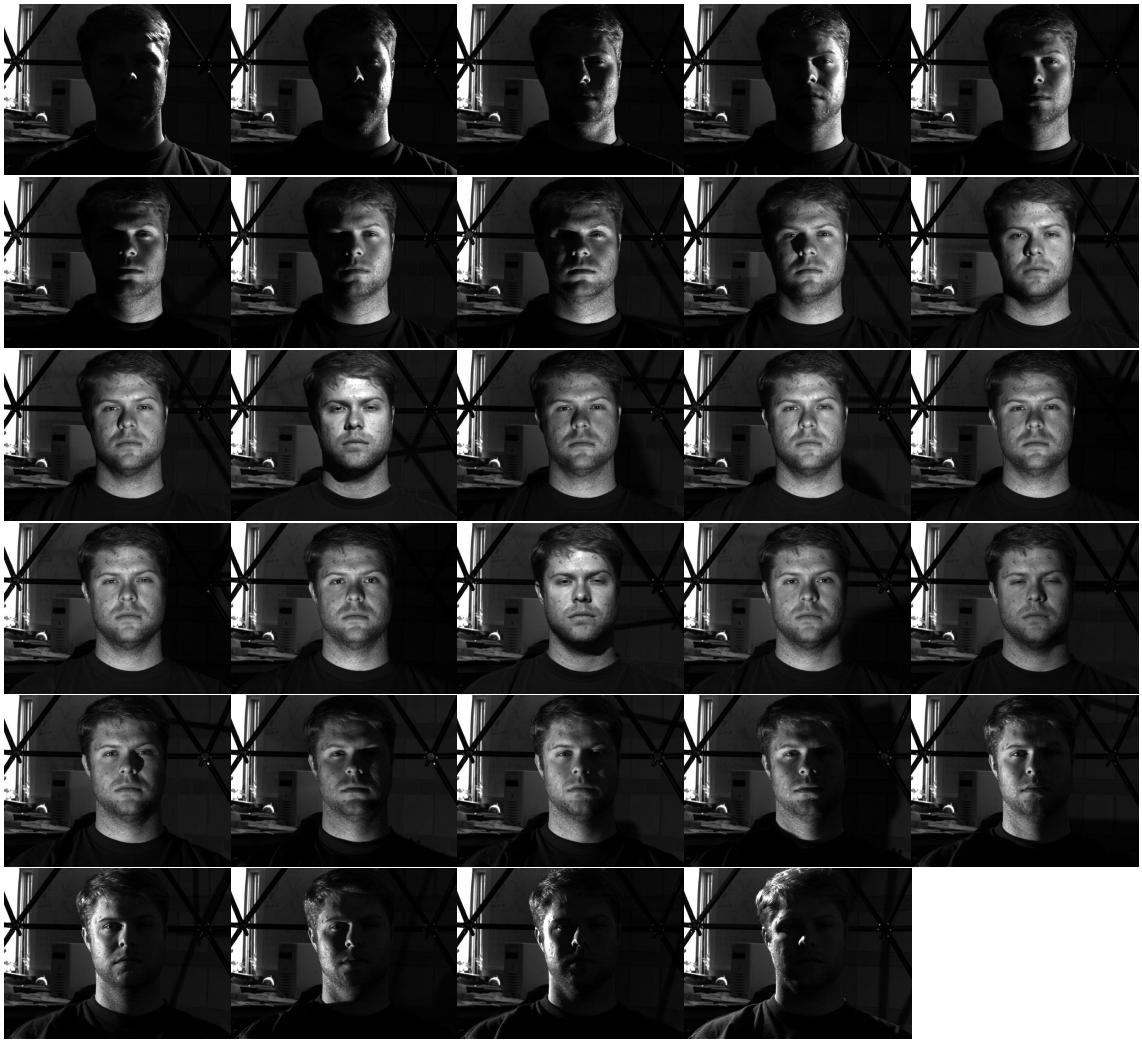


Figure 5.3: 29 out of 64 lighting conditions per subject from the Yale Face Database B (Georghiades et al., 2001). This particular subset shows the variation in the azimuth of the light source. The rest of the lighting conditions are due to variations in elevation.

The index k of the largest component of the vector z corresponds to the row of the matrix G which is the most similar to y and therefore this index also corresponds to the image in the gallery. Thus, the procedure for comparison of vectors can be expressed in terms of manipulations with matrices (multiplication and finding index of maximum element) which drastically accelerates computations. It takes only 5.62

seconds to compare 5,696 face vectors of size 2,048 to a gallery of 5,696 faces using Matlab on a desktop computer.

5.1.3 Results of feature vector comparison

For the scheme with vector length 8,192 (2 bit feature encoding, three output neurons and a receptive field of 64×64), we tested the hypothesis that for each image from the database, the best match is an image of the same person. The following procedure was used: select image from the database; compare this image with all other images, find the most similar image, and verify that the identities of selected image and the best match are the same. Repeating this procedure for all images from the database, it was found that the rate of correct "identification" is 0.9896. This rate means the in the large majority of cases the nearest neighbour of any give feature vector is also a feature vector from the same subject, but in some other cases the nearest neighbour is a feature vector from a different subject, resulting in an identification error. After this encouraging result, we use a more sophisticated procedure for investigating the coding space. In order to study further the properties of the vector space in relation to images taken under different illumination and pose conditions we used two set-ups for defining the gallery and matching data:

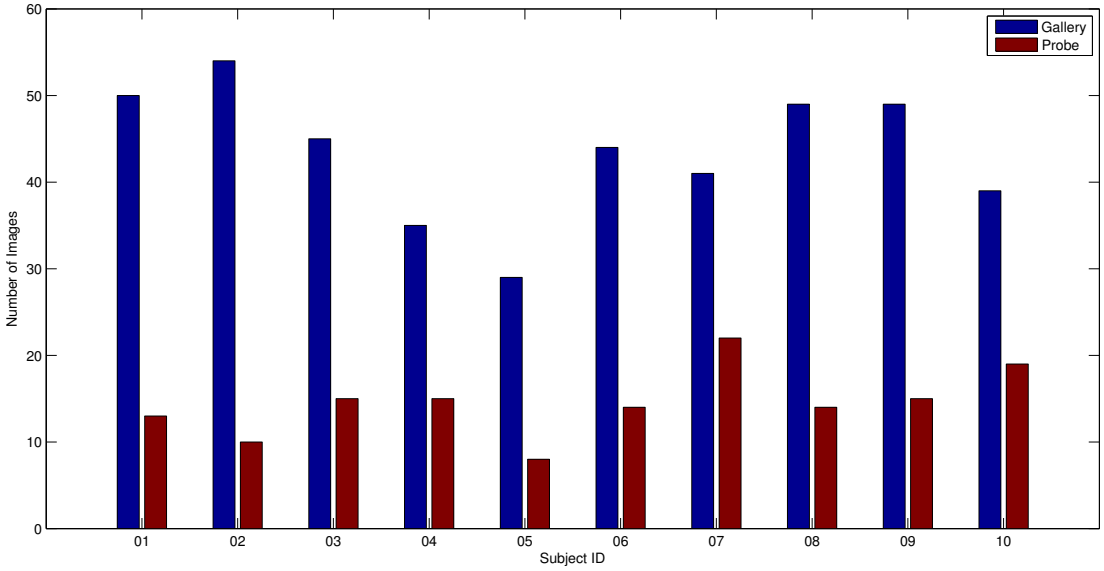
1. Illumination (il): The frontal pose [1] is fixed and a fraction F of the images with this pose but different illumination was randomly (uniformly) selected for the gallery. The remaining images for the same pose are used for identification.
2. Pose (po): The frontal illumination[2] is fixed and a fraction F of the images

[1] Frontal or pose '00' according to Georghiades et al. (2001) specifications.

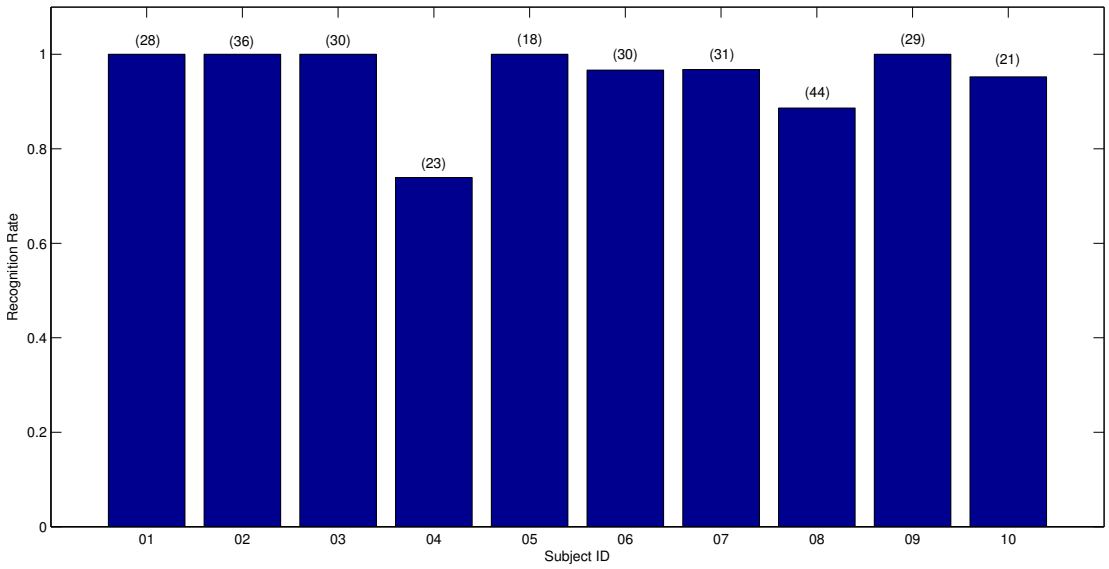
[2] Frontal illumination or illumination source direction with respect to the camera axis is at 0 degrees azimuth and 0 degrees elevation according to Georghiades et al. (2001) specifications.

with this illumination but different pose was randomly (uniformly) selected for the gallery. The remaining images for the same illumination are used for identification.

We studied a variety of cases corresponding to different values for the faction of images used for the gallery F (1/2, 3/4, and 7/8), V1 complex cells receptive field S, number of output neurons O and the number of binary values for each orientation. For each case a gallery was generated and the matching procedure described earlier was repeated for 50 runs.The gallery is generated independently for each run. Figure 5.4 shows one example how these images are distributed across all subjects for the case where 3/4 of the images chosen for the gallery and 1/4 for the probe set, as well as the correct identification rate per subject for another example, where 1/2 of the images chosen for the gallery and 1/2 for the identification set. The mean correct identification rate (m_{ir}) over 50 runs and the corresponding standard deviation (σ_{ir}) are shown in Tables 5.1 and 5.2, for the cases where we have $O = 1$ and $O = 3$, respectively. The result shown in Tables 5.1 and 5.2 evidences that a simple comparison with the gallery provides a good face "identification" which is comparable with results of Jaiswal et al. (2011) and Vu and Caplier (2011) in the same dataset. The best results regarding the pose variations are are achieved when the number of output features is 3, as opposed to a single feature covering the whole face. In the illumination test cases, the improvements are not significant. A possible reason for this observation is that, during the training phase, each of the three different output neurons became more specialised in a certain pose. Therefore, the detection across different poses can be more accurate, and this helps in the



(a)



(b)

Figure 5.4: (a) Number of images selected in one of the runs for the gallery and identification sets, for each subject. In this example 3/4 of the images were chosen for the gallery and 1/4 for the identification set. (b) Example of correct identification rate for each of the subjects. The values in parentheses are the number of identification images for a given subject. In this example 1/2 of the images were chosen for the gallery and 1/2 for the identification set.

		$S = 32 \times 32$		$S = 64 \times 64$	
		2-bit	4-bit	2-bit	4-bit
1/2	il	0.961 (0.0083)	0.962 (0.0125)	**0.967** (0.011)	0.963 (0.0116)
	po	0.629 (0.0639)	0.627 (0.0832)	**0.847** (0.051)	0.803 (0.0803)
3/4	il	0.964 (0.0124)	0.969 (0.0098)	0.97 (0.0163)	**0.972** (0.0156)
	po	0.714 (0.0781)	0.746 (0.0817)	**0.908** (0.0435)	0.869 (0.0642)
7/8	il	0.971 (0.0178)	0.975 (0.0161)	**0.981** (0.0169)	0.968 (0.0212)
	po	0.743 (0.1172)	0.781 (0.1246)	0.892 (0.0778)	**0.902** (0.0769)

Table 5.1: (m_{ir}) and, in parentheses, (σ_{ir}) for all the different face recognition settings with number of output neurons $O = 1$, varied receptive field size S, feature representation (2 and 4-bit coding), gallery/matching set-ups (illumination il; pose po) and portion of images used for the gallery ($F\epsilon\{1/2, 3/4, 7/8\}$). The highest correct matching rate for each gallery/matching setting is highlighted. For $S = 32 \times 32$ and Illumination setting, 580 images were used in total; for $S = 32 \times 32$ and Pose setting, 88 images were used in total; for $S = 64 \times 64$ and Illumination setting, 460 images were used in total; for $S = 64 \times 64$ and Pose setting, 86 images were used in total; These numbers reflect the number of faces available for validation, which are only the ones that were successfully detected by the face detector. In general, the correct matching rate for pose variation is much lower than for illumination variation. The possible reason is that the Gabor filters combined with the max-pooling mechanism are more robust to local variations caused by different illumination than the variations resulting from different poses which are less local.

		$S = 32 \times 32$		$S = 64 \times 64$	
		2-bit	4-bit	2-bit	4-bit
1/2	il	**0.974** (0.0098)	0.956 (0.012)	0.974 (0.0077)	0.931 (0.0151)
	po	0.771 (0.0647)	0.783 (0.0703)	**0.947** (0.0548)	0.927 (0.0495)
3/4	il	**0.982** (0.0108)	0.969 (0.0147)	0.978 (0.0105)	0.946 (0.0174)
	po	0.856 (0.0683)	0.836 (0.0729)	**0.983** (0.0317)	0.97 (0.0339)
7/8	il	**0.983** (0.0128)	0.974 (0.0191)	0.978 (0.0161)	0.951 (0.183)
	po	0.878 (0.0947)	0.866 (0.0993)	**0.987** (0.0319)	0.986 (0.0337)

Table 5.2: (m_{ir}) and, in parentheses, (σ_{ir}) for all the different face recognition settings with number of output neurons $O = 3$, varied receptive field size S, feature representation (2 and 4-bit coding), gallery/matching set-ups (illumination il; pose po) and portion of images used for the gallery ($F\epsilon\{1/2, 3/4, 7/8\}$). The highest correct matching rate for each gallery/matching setting is highlighted. For $S = 32 \times 32$ and Illumination setting, 580 images were used in total; for $S = 32 \times 32$ and Pose setting, 88 images were used in total; for $S = 64 \times 64$ and Illumination setting, 460 images were used in total; for $S = 64 \times 64$ and Pose setting, 86 images were used in total; This numbers reflect the number of faces available for validation, which are only the ones that were successfully detected by the face detector. With this setting ($O = 3$), the pose correct matching rates improved considerably when compared to $O = 3$. The different levels of detail and slightly different face regions introduced by having three output neurons are likely to be the reason for this improvement. This is because of the redundancy introduced by having more output neurons and because each neuron corresponds to a different part of the face, and some of those parts might vary less with pose than others.

identification of the correct face. The 2-bit representation of individual features also leads to slightly better results than the 4-bit representation. This is expected because the 2-bit representation minimises the similarity between further apart orientations, and maximises the distance between closer orientations, as opposed to the 4-bit representation where the distance between any two orientations is always constant. For the pose test cases, it is better to use a 64×64 receptive field, while for the illumination cases, a 32×32 receptive field leads to better results.

5.2 Conclusions

We present a face features coding scheme, which achieved identification rates at the same level as some well-known face recognition algorithms. This outcome is achieved with our binary coding of features that enables an efficient representation of faces and a quick simple comparison with other vectors, which achieved correct identification rates higher than 0.97 in most of the cases.

This results are comparable with other algorithms that have been tested in the same dataset, namely Jaiswal et al. (2011) and Vu and Caplier (2011). They have also tested their systems across different illuminations variations, but not across different poses.

The state of the art algorithms such as Taigman et al. (2014) and Sun et al. (2014) achieve recognition rates in the same region as our approach, but in a much more difficult dataset, the Labelled Faces in the Wild (Huang et al., 2007). This dataset contains a very large variation of poses, lighting, age, hair and beard styles, among other variations that occur in images taken in "the wild", i.e., in an totally uncon-

trolled environment. For this reason, the state of the art approaches should easily outperform the ours in the smaller, easier, test set used in this chapter. Nevertheless the approach here presented, is much simpler and computationally much more efficient than the state of the art, therefore it could be used as a pre-processing stage to filter out most of the false matches, thus reducing the search space for a more complex algorithm.

Chapter 6

Memory model approach for multi-pose face recognition

One of the most challenging problems in face recognition is the large variability of the appearance of faces from the same subject, due to changes of lighting conditions, pose, etc. As shown in the previous chapter, the face detection and feature extraction model presented in this thesis, combined with the features coding scheme and comparison method from Chapters 4 and 5, copes better with lighting conditions variations than with pose variations. Therefore we have decided to propose an method for assisting with memorization and recall of different poses, which could help improving the recognition performance. This problem has been broadly studied in the literature (Chai et al., 2003; Perrett et al., 1998; Shepard and Metzler, 1971; Sinha and Poggio, 1996; Xie and Lam, 2006; Yamaguchi et al., 1998).

In computer vision, several algorithms have been proposed to tackle this problem, which can be divided in two main categories:

- Normalisation-based algorithms
- Sequence-based recognition algorithms

On one hand, the normalisation-based algorithms try to normalise the input image in order to bring it as close as possible to a standard view. This can be done, for instance, by using a transformation to project a side view of a face into a frontal view (Chai et al., 2003), or to normalise the image in order to smooth out the effects of the lighting (Xie and Lam, 2006).

On the other hand, the sequence-based recognition algorithms rely on a series of images for recognition, instead of using a single image (Yamaguchi et al., 1998), which increases the data available for comparing two faces, making the recognition more robust.

Similarly, in neuroscience, there are two alternative explanations for the ability of the brain to recognise faces from different views. The first alternative relies on a mental transformation of the face before the recognition process (Shepard and Metzler, 1971), which would transform the observed face into a standard view. The second alternative is that we learn the 3D appearance of the objects by looking at different views (Sinha and Poggio, 1996) and then we recognise new faces according to our previous experience, which includes several views (Perrett et al., 1998). The later hypothesis is more in line with the hierarchical processing model we use, in particular, with the multi-view face detector proposed in this thesis, therefore it was the chosen alternative to tackle the problem of the input variability in our proposed face recognition system.

In order to test this hypothesis we have assumed that there is some sort of memory

organisation where the different views are stored in an ordered manner, and not in a completely random way. Therefore we propose an approach based on the model of Borisyuk et al. (2013), which is a completely new and complementary model to what have been discussed in the previous chapters, for a multi-view memory organisation of the faces in the brain to assist the face recognition process. The same model could also be used for other memory organisation structures based in a transition between views, such has transitions between light sources.

In previous chapters of the thesis we consider how to extract faces from images, how to extract features from faces, how to code faces in an optimal way. In this chapter we concentrate on face recognition. Our approach is based on a new idea to use a neural network model for memorising sequences. This model consists of two layers, the first layer deals with representation of objects of sequences to be memorise and the upper layer keeps labels of sequences. During the learning process the connections between neurons of the first layer are adjusted to realise a chain rule of consequent memories. We use both Hebbian and anti-Hebbian learning rules (for two separate networks) to be able to recall in forward and backward directions. Also, all to all connections from the upper layer to the first layer are adjusted to prescribe a tag to the objects of some particular sequence. To recall a sequence we need to stimulate any object (or even a part of the object) from the sequence as well as the neurons of upper layer related to the sequence tag. Stimulation of one object of the sequence is not enough to start re-play because connections between neurons in the first layer which have been adjusted according the chain rule are not sufficiently strong to start recall of the sequence and additional input from the

upper layer corresponding to this particular sequence should arrive to start replay. After that the sequence will be played in forward and backward directions (by two different networks) starting from the initiating object.

We use this new model of memory to recognise faces. For that we assume that each subject is represented by some number of faces in different poses and these faces are objects of the sequence to memorise with a tag which is prescribed to the sequence in the upper layer. Thus we have a sequence of P poses (generally speaking it is possible to have different number of poses for different subjects) for each subject. To select faces, select features and code them we use methods and techniques which are described in previous chapters of the thesis. There are S subjects and we train the neural network to memorise S sequences with the tag prescribed to each sequence. To recognise some face we use the face's code as an input to the first layer and in the upper layer we initiate a tag for sequence 1. If this particular face belongs to the sequence 1 then the whole sequence will be re-played starting from a giving face (one network will show a part of sequence in forward time and another network will play another part of sequence in backward time). If the face is not from sequence 1, then the memory model will not play a sequence because there are no proper connections in the first layer and there are no input from the upper layer to support recall. After that we repeat this procedure: we initiate neurons corresponding to the same face in the first layer and the tag in the upper layer corresponding to sequence 2. The sequence 2 will be recalled if the face belongs to the sequence 2, otherwise it will be no replay. Thus, the maximum number of repetition of this procedure to recognise a face is S, and this number is relatively small in comparison with a total number

of faces which is P*S. For example if we have 100 subjects and for each subject we have 200 poses then the total amount of faces is 20,000. For recognition of the face we do not need to do 10,999 comparisons but 100 repetitions of the recall procedure will be enough to recognise the face.

In this chapter we start by presenting briefly the original model for sequence memorisation (Borisyuk et al., 2013; more details about this model can be found in the bounded copy of the paper attached to the end of this thesis). Then we present our theoretical approach for the application of this memory for sequences model to the problem of multi-view face recognition.

6.1 Model for memorisation of sequences

The system is made of several functional units which are neural groups of inhibitory and excitatory spiking neurons. These units represent different elements of a sequence of labels, which could be assigned to events, images of other categories. The training phase consists in presenting one element at a time by stimulating the corresponding groups. The synaptic weights are adjusted during this phase according to a rule similar to STDP. Finally the recall is performed stimulating only one of the groups of the sequence.

The system consists of two layers, as shown in Figure 6.1. The top layer has several groups of excitatory neurons. Each group has 60 neurons and represents some high level processing task, for instance a label for a person. The neural groups in the top layer project modifiable connections to all the neural groups in the bottom layer that belong to a given sequence, therefore each group in the top layer is able to

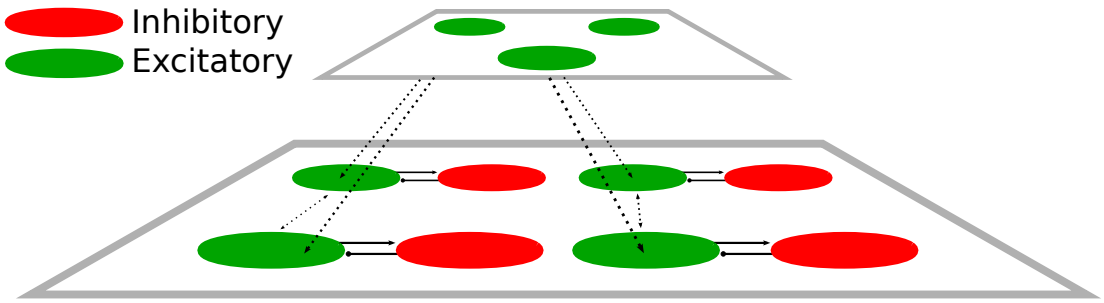


Figure 6.1: Spiking neural network model. The top layer contains groups of excitatory neurons. The bottom layer contains groups of coupled excitatory and inhibitory neurons. Each neural group contains 80 excitatory neurons (in green) and 20 inhibitory neurons (in red). A neuron in a neural group delivers connections to all other (both excitatory and inhibitory) neurons within the group (abbreviated as arrows and lines with a circle end). Between excitatory neurons of different groups or layers, there are plastic connections (shown as dotted lines).

label any sequence by helping to activate the bottom groups on the recall period. Therefore these top-to-bottom connections have an important role during recall, but they don't take part in the learning phase. The bottom layer contains many groups of low-level neurons.

Each neural group from the bottom layer has 80 excitatory and 20 inhibitory spiking neurons. The neurons of each group are connected with all other neurons inside the same group. These neural groups, when stimulated, produce rhythmic activity in the gamma range. Gamma range oscillations, alongside with theta range oscillations, have been shown to have an important role in the memorization of sequences of events among other cognitive functions (Burgess and O'Keefe, 2011) This layer implements the chain mechanism which will be able to record the sequence of face

views. The recording of the sequence is made adjusting the modifiable connections among groups. These connections are always all-to-all links between excitatory neurons of different neural groups.

The ionic dynamics of these individual neurons are described using Hodgkin-Huxley equations (Hodgkin and Huxley, 1952),

$$\frac{dV_i}{dt} = -I_{ion,i} + I_{syn,i}^{lower} + I_{ext,i} + I_{rest}, \tag{6.1}$$

$$\frac{dX_i}{dt} = A_X(V_i)(1 - X_i) - B_X(V_i)X_i, \quad X_i \in \{m_i, h_i, n_i\}, \quad i = 1, 2, ..., N, \tag{6.2}$$

where N is the number of neurons in the layer; $V_i(t)$ is the membrane potential of a neuron; X is a notation for any of the variables $m_i(t)$, $h_i(t)$, $n_i(t)$, where $m_i(t)$ is the activation variable of the sodium conductance channel, $h_i(t)$ is the inactivation variable of the sodium conductance channel, and $n_i(t)$ is the activation variable of the potassium conductance channel; $I_{syn,i}^{lower}(t)$ is the synaptic current received by a neuron from other neurons in the lower layer; $I_{syn,i}^{upper}(t)$ is the synaptic current received by a lower layer excitatory neuron from upper layer neurons; $I_{ext,i}(t)$ is the external current induced by the external input (40 mA); I_{rest} is a universal constant current that controls the activities of the neurons (equal to -25 mA). The total ionic current $I_{ion}, i(t)$ and the gating functions A_X and B_X are described using the following equations:

$$A_m(V) = \frac{2.5 - 0.1(V - V_{rest})}{\exp(2.5 - 0.1(V - V_{rest})) - 1}, \tag{6.3}$$

$$A_h(V) = 0.07 \exp\left(\frac{-(V - V_{rest})}{20}\right), \tag{6.4}$$

$$A_n(V) = \frac{0.1 - 0.01(V - V_{rest})}{\exp(1 - 0.1(V - V_{rest})) - 1}, \tag{6.5}$$

$$B_m(V) = 4 \exp\left(\frac{-(V - V_{rest})}{18}\right), \tag{6.6}$$

$$B_h(V) = \frac{1}{\exp(3 - 0.1(V - V_{rest})) + 1}, \tag{6.7}$$

$$B_n(V) = 0.124 \exp\left(\frac{-(V - V_{rest})}{80}\right), \tag{6.8}$$

where V is the membrane potential of the neuron. The sum of ionic currents I_{ion} of a neuron is

$$I_{ion} = g_{Na} m^3 h(V - V_{Na}) + G_K n^4 (V - V_K) + g_L(V - V_L), \tag{6.9}$$

where V_{Na} is the reversal potential for the sodium current (equal to 50 mV), V_K is the reversal potential for the potassium current (equal to -77 mV), V_L is the reversal potential for the leak current (equal to -54.4 mV), $g_N a$ is the maximum conductance for the sodium current ($g_N a = 120(1 + 0.02\eta) mS/cm^2$, η is uniformly distributed in [-1,1]), g_K is the maximum conductance for the potassium current ($g_K = 36(1 + 0.02\eta) mS/cm^2$, η is uniformly distributed in [-1,1]), g_L is the maximum conductance for the leak current ($g_L = 0.3(1 + 0.02\eta) mS/cm^2$, η is the symmetry breaking uniformly distributed random variable in [-1,1]).

A positive external signal is received by a neuron in the bottom layer at each moment when a member of the stimulation sequence is presented. Otherwise the external signal is equal to zero. The external current is strong enough to transfer a neuron into the firing state. Without external current a neuron can fire if it receives both upper the synaptic current from the upper layer $I^{upper}_{syn,i}(t)$ and synaptic currents from other modules. Synaptic conductance is described using a standard alpha-function (see, e.g., Gerstner and Kistler, 2002). The total synaptic current of the ith neuron in the lower layer received from the neurons of the lower layer is described by the following equation:

$$I_{syn,i}^{lower} = \sum_{j \in N_i^{inh}} I_{i,j}^{inh} + \sum_{j \in N_i^{exc}} I_{i,j}^{exc} + \sum_{j \in N_{i,external}^{exc}} I_{i,j}^{exc,external}, \quad i = 1, 2, ..., N. \quad (6.10)$$

Here, the two first terms describe the sum of inhibitory and excitatory influences, N_i^{inh} (N_i^{exc}) is a set of indexes of incoming inhibitory (excitatory) connections from neurons of the same module; the third term describes the sum of external excitatory influences from neurons of other modules at the lower layer, $N_{i,external}^{exc}$ is a set of indexes of incoming excitatory connections from the neurons of other modules.

$$I_{i,j}^{inh} W_{inh}^{A} (V_i - V_{syn}^{inh}) \sum_{k=1}^{M^j} \alpha_j (t - T_k), \quad (6.11)$$

$$I_{i,j}^{exc} W_{exc}^{A} (V_i - V_{syn}^{exc}) \sum_{k=1}^{M^j} \alpha_j (t - T_k), \quad (6.12)$$

$$I_{i,j}^{exc,external} W_{ij,exc}^{B}(t) (V_i - V_{syn}^{exc}) \sum_{k=1}^{M^j} \alpha_j (t - T_k), \quad (6.13)$$

Here, $W_{inh}^{A} = W_{exc}^{A} = 0.1$ are constant connection strengths for inhibitory and excitatory connections inside the module; the alpha function is defined in the following way: $\alpha_j(t) = at \exp(-bt)$, for $t \geq 0$ and the alpha function equals to zero for $t < 0$; the parameters of the alpha function are: $a = 0.6$ m per second and $b = 0.03$ m per second; M^j is the total number of spikes from the jth neuron to the ith neuron; T_k is the time of the kth spike generated by the jth neuron. $V^{inh}syn$ is the synaptic reversal potential of inhibitory coupling (V_{syn}^{inh} = -80 mV), $W_{ij,exc}^{B}(t)$ is a modifiable excitatory connection strength from the jth neuron to the ith neuron of different modules.

The total synaptic current of the ith neuron in the lower layer received from neurons

of the upper layer is described by the following equation:

$$I_{syn,i}^{upper} = \sum_{j \in N_{i,upper}^{exc}} W_{ij,exc}^{C}(t)(V_i - V^{exc}syn) \sum_{k=1}^{M^j} \alpha_j(t - T_k). \tag{6.14}$$

where $W_{ij,exc}^{C}(t)$ is a modifiable excitatory connection strength from the jth neuron of the upper layer to the ith excitatory neuron of the lower layer; $N_{i,upper}$ is a set of indexes of incoming excitatory connections from neurons of the upper layer to the ith excitatory neuron of the lower layer; the alpha function: $\alpha_j(t) = at\exp(-bt)$, for $t \geq 0$ and the alpha function equals to zero for $t < 0$; the parameters of the alpha function are: $a = 0.6$ m per second and $b = 0.03$ m per second; M^j is the total number of spikes from the jth neuron at the upper layer to the ith neuron; T_k is the time of the kth spike generated by the jth neuron at the upper layer; $V^{exc}syn$ is the synaptic reversal potential of excitatory coupling ($V^{exc}syn = 0$ mV).

Before the memorisation or training period, all the connections between different groups are set to zero. The memorisation of a sequence is done by making each neural group belonging to the sequence oscillate sequentially by applying and external current ($I_{ext,i} = 40$). Each group is stimulated during 200 ms, and after this period the external stimulus is withdrawn.

In order to effectively record the sequence, the connection strengths have to be adjusted. The proposed method for changing the value is a temporally asymmetric learning rule which is similar to Spike-Timing-Dependent Plasticity (STDP, see e.g. Markram et al., 1997). The activity level of pre- and post-synaptic neural groups is monitored in two subsequent time windows of 200 ms each. This time window is in accordance with the theta rhythm observed in the hippocampus according to Colgin and Moser (2006). When a neuron k in one group fires a spike within the

previous time window before another neuron i in another group fires a spike within the current time window, then the connection strength value will increase by 3 and will decrease by 0.0001 otherwise. This small decrease was introduced to implement a basic forgetting mechanism. The maximum value of the connection strength is 3 mS/cm^2, and the minimum 0 mS/cm^2.

Two types of the learning directions are proposed: one for forward recall and another for backward recall. For forward recall the STDP type rule is applied (the direction of coupling is from neuron j to neuron i) and for the backward recall, the anti-STDP type rule is applied (the direction is from neuron i to neuron j). Anti-STDP has been suggested by previous papers (Han et al., 2000; Rumsey and Abbott, 2004). For simplicity, it is assumed that neural groups can be distinguished as either STDP or anti-STDP type. All neurons in one neural group have the same type of connection direction and they only connect to groups of the same type (i.e. a STDP neural group only connects to another STDP group).

In parallel with the modification of the connection strengths between neurons of the group which "tags" the sequence being memorised to the neurons of active modules is also modified using the same learning rule. An external current ($I_{ext,i} = 40$) is applied to the top neuron group that is selected to label the sequence. After a sequence is memorised, each neural group in the top layer will project connections to all the neural groups in the bottom layer that belong to the corresponding sequence. In this way the sequence is labelled by that top layer excitatory neural group. This top layer is implemented delivering a special current to the connected bottom layer neural groups.

At the beginning of the recall period the neural group representing the first element of the sequence is stimulated briefly (50 ms) by applying and external current (equal to 40). Also the top layer neural group delivers constant current (equal to 25) to all the labelled neural groups in the bottom layer which encode elements of the same sequence, until the end of the recall process. Note that during recall the periods of activation of the neural groups overlap in time, but the moments when the activity in each group starts are ordered in the same way as during memorisation of the sequence.

Since there are many neural groups in the bottom layer, we can assume that there always exists a "representative" neural group for each element of the sequence, such that that element is sufficiently identified by this neural group. There may be other neural groups which also share the representation of that element, but they are not required to participate in the system because the sequence can be accurately memorised by connecting only those representative neural groups.

6.2 Proposed memory model for multi-pose face recognition

The model described above is a very good candidate for a memory for faces based in a sequence of poses for each subject.

A proof of concept was developed in C as part of this thesis, in which the model presented above was implemented and simulations to memorise and recall sequences

were performed.

Because of the limitations of the model, i.e., there are no real images or features vectors stored in memory, we propose a theoretical approach to how this model could help with recognition, rather then results from further experiments combining this model with the model and coding schemes presented earlier in this thesis.

We propose that each neural group in the bottom layer of the network represents a specific pose, and each of the top layer groups represents a subject, i.e., each top neural group would label a particular subject. Figure 6.2 illustrates the architecture of the proposed network structure. During the training phase a sequence of poses are presented to the network. This sequence of poses corresponds to different views of the same subject's face. Such sequence of views results from the observation of a face in movement. This is how we learn new faces in the real world, i.e., we don't usually look at a single image of a face, instead we look at a moving face, which can be represented by a number of different poses.

The top layer group that labels the current subject, projects connections to all the elements of the sequence of poses. There are two separated networks, one trained with a STDP rule, for forward recall, and another with an anti-STDP rule, for backward recall. Both are trained simultaneously in a similar manner, apart from the rule for adjusting the weights, which would differ.

The training process is repeated for every new subject, therefore the resulting networks would have S top groups, one for each subject, and N bottom groups, one for each pose and subject, i.e., $N = S * P$, where P is the number of different poses in each sequence. The memorisation process of a sequence of four poses for one subject

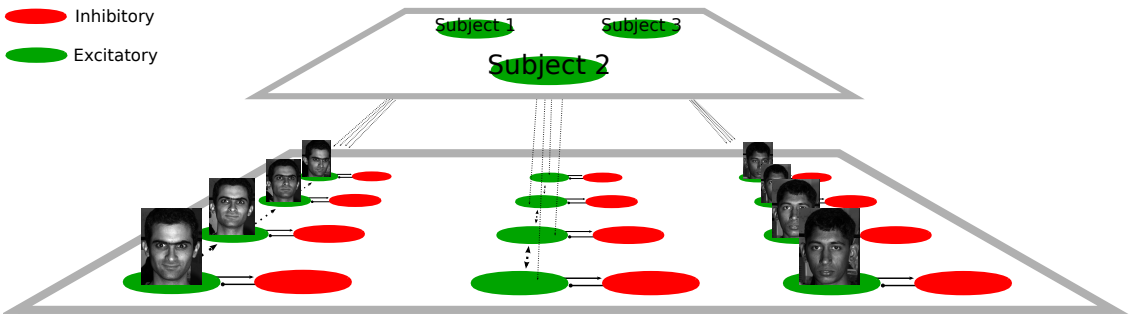


Figure 6.2: Spiking neural network model. The top layer contains groups of excitatory neurons. The bottom layer contains groups of coupled excitatory and inhibitory neurons. Each neural group contains 80 excitatory neurons (in green) and 20 inhibitory neurons (in red). A neuron in a neural group delivers connections to all other (both excitatory and inhibitory) neurons within the group (abbreviated as arrows and lines with a circle end). Between excitatory neurons of different groups or layers, there are plastic connections (shown as dotted lines). Each set of connected groups in the bottom layer corresponds to a sequence of poses, which is memorised in a particular order during the memorisation process, and can be recalled in the same or inverse order.

is illustrated in Figure 6.3. This network can be used to recall a sequence of poses starting by activating one neural group corresponding to a pose in the sequence. This pose would be detected by with our proposed improved model for multi-pose face detection. Then, we stimulate the first top neural group corresponding to an individual. This recall mechanism is illustrated in Figure 6.4 at two different levels of detail. If, in one hand, when we stimulate the top neural group (while the bottom layer group corresponding is still active) , the sequence of poses is successfully recalled, that means that the top group which is currently stimulated, labels the

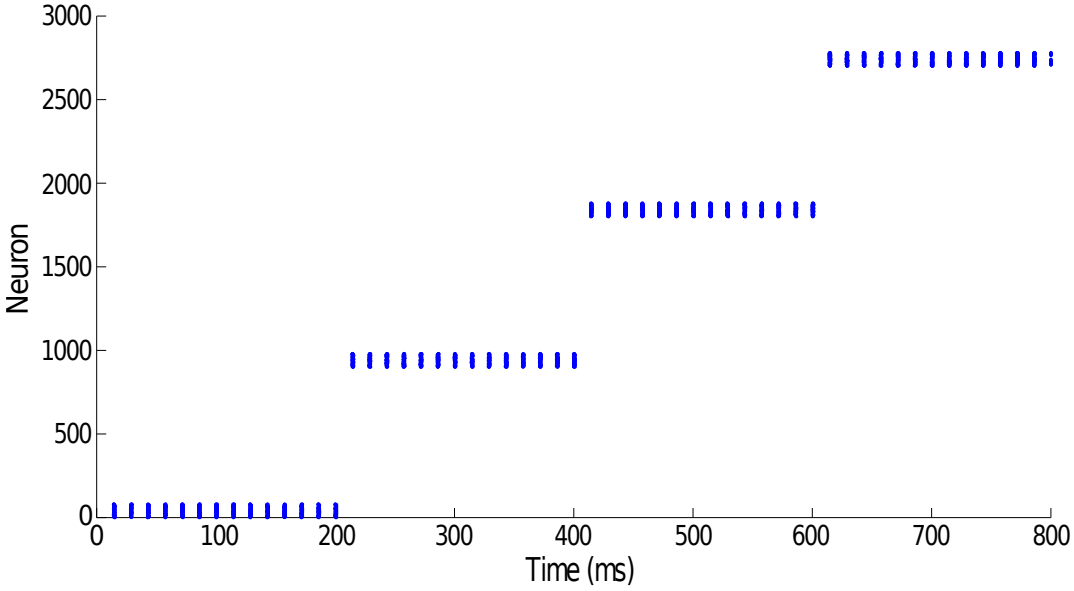
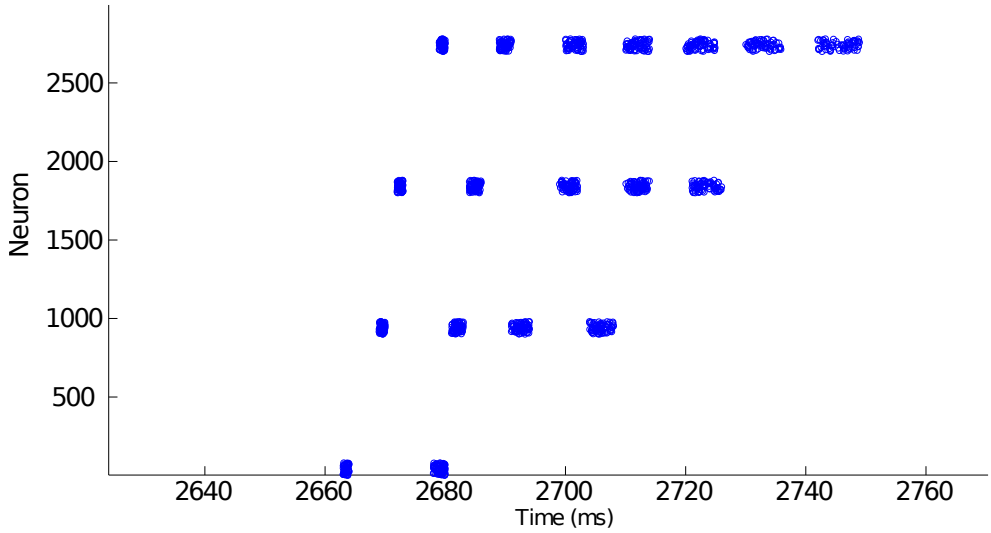
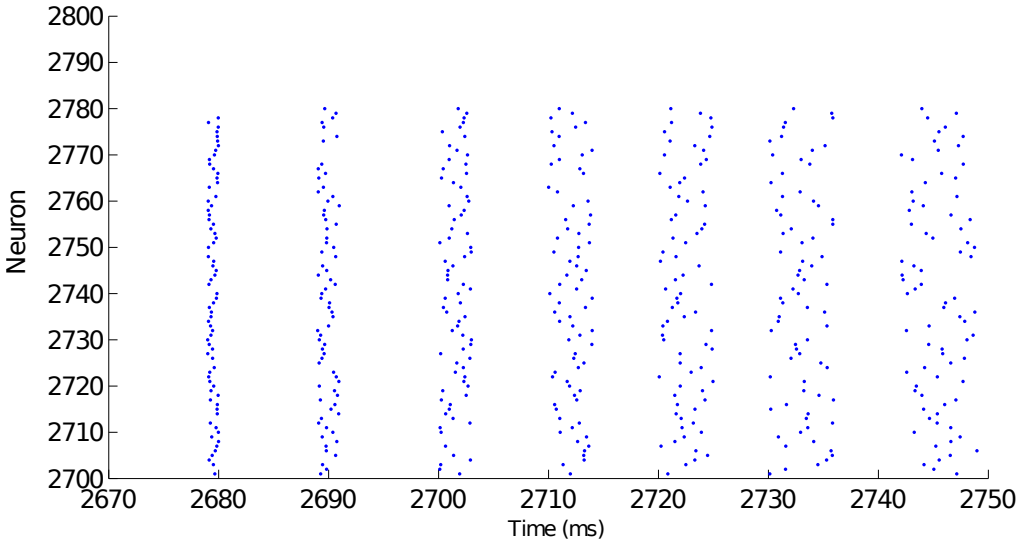


Figure 6.3: Memorisation of a sequence of four poses for a single subject. The memorization for the forwards and backwards recall is done in the same way, apart from the connections adjustment algorithms, which is not shown in this figure. Each small dot in the graph represents a neural spike. Each cluster of spikes corresponds to one bottom layer neural group which is representative of a pose for a particular subject.

probe face, i.e., the sequence recalled identifies the most likely subject. This recall occurs in both directions, because we have both forward and backward recall networks, therefore all the available poses would be recalled independently of the start point (Figure 6.5 illustrates the backward recall process). If, in another hand, the top layer being stimulated does't correspond to the subject active in the bottom layer then the network wouldn't recall the whole sequence. Therefore the current top group does not label the face detected, and the recall process would restart by stimulating the next group from the top layer, while the bottom probe group would



(a) Recall.



(b) Recall zoom.

Figure 6.4: Each small dot in the graph represents a neural spike. (a) Forwards recall of a sequence of four poses for a single subject, showing the subsequent firing of neural groups with a short delay. (b) A magnified picture in the recall period of a single group.

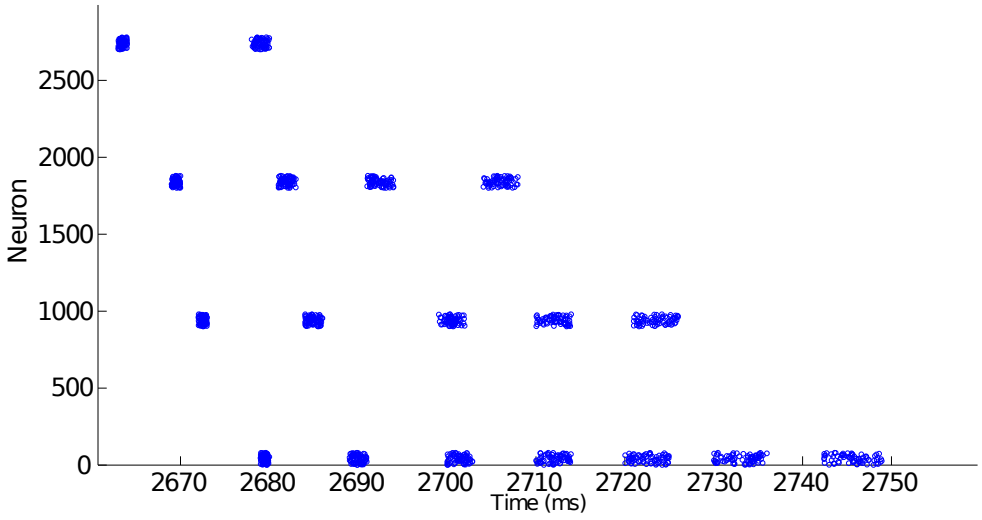
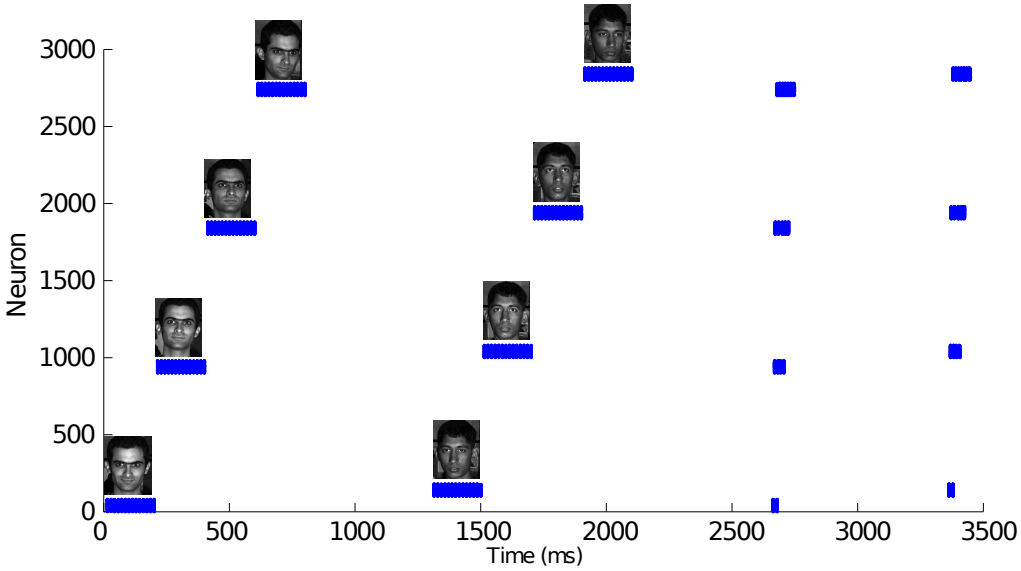


Figure 6.5: Each small circle in the graph represents a neural spike. Backward recall of a sequence of four poses for a single subject, showing the subsequent firing of neural groups with a short delay.

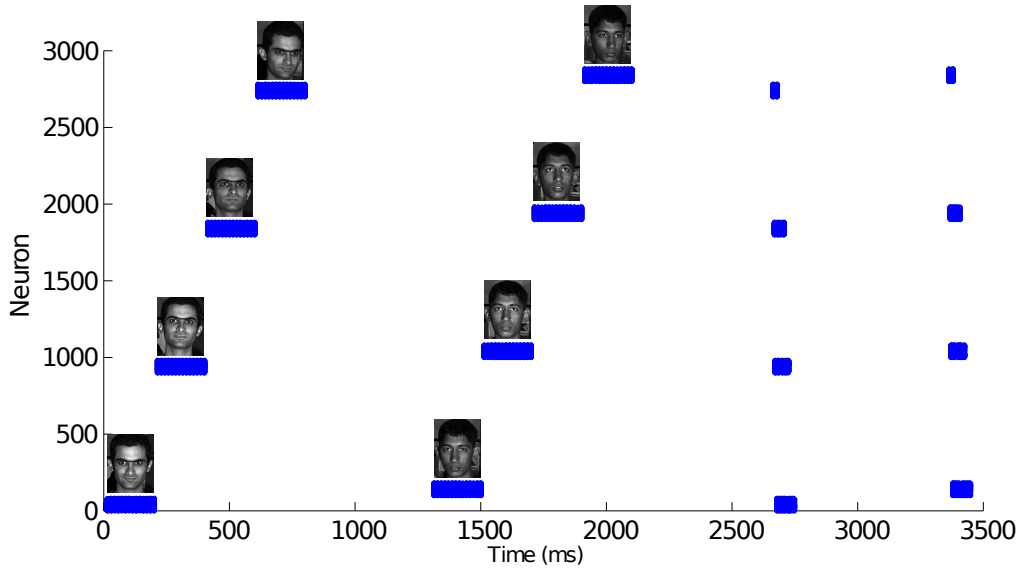
still be active, and check if the recall of the sequence is performed successfully.

The main advantage of using this approach for recognition is that for a memory with S subjects and P poses we just need to try, at most, S times to recall the sequence of poses to find the identity of the subject. With a nearest neighbour approach we would always have to compare a probe face with $S \times P$ faces in memory.

The second advantage is that the novelty detection is performed automatically, because if after trying starting the recall with all the top groups, the network is not capable of recalling any sequence, this means that the probe face is a novel subject of a novel view of an existing subject. Figure 6.6 illustrates the whole process of memorisation, forward and backward recall of sequences of four poses for two different subjects. In order to achieve a more robust recognition, we propose to compare several views of the probe with the views stored in the memory. Therefore, after



(a) Memorisation (0-2100ms) and recall (2500-3500ms).



(b) Memorisation and (0-2100ms) and backward recall (2500-3500ms).

Figure 6.6: Each small dot in the graph represents a neural spike. (a) Memorisation of two sequences of four poses. The left part shows the activations of neurons (memorisation period). The right part shows the first sequence recall which is initiated by the activity pattern corresponding to the first event (recall period). (b) Memorisation and backward recall of two sequences.

finding the group s from the top layer that successfully labelled the sequence containing the first probe view, we can, for the following probe views, start the search from the same s, knowing that, with high probability, we would be able to recall the sequence without having to try additional top layer groups. The reason for this is that the neighbouring probe views are probably of the same subject, with a slightly different pose.

6.3 Conclusions

In this chapter we present an approach for the problem of memorisation of different poses, based on the work of Borisyuk et al. (2013). The proposed memory for sequences of poses is in line with psychological and biological studies, which indicate that the recognition is a dynamical process, in which several views of the same object take part in the matching and recognition. Therefore the proposed memory model can assist in the recognition of an individual by recalling the sequence of poses, starting from stimulating one neural group corresponding to the observed pose, and also the top neural groups, one by one, which correspond to the identity labels, until the sequence is successfully recalled. If the sequence is successfully recalled, there is a strong indication that the observed pose is labelled by the top layer group being stimulated, otherwise the network will perform novelty detection.

The main contributions of this chapter are:

- A theoretical approach to the problem of multi-view memorisation of faces
- A methodology to take advantage of the multi-view memorisation of faces for

improved recognition, in terms of speed and accuracy

- A methodology to take advantage of the multi-view memorisation of faces for novelty detection

Chapter 7

Summary

In this thesis we present a brain inspired approach to computational face recognition. We start by making a review of the state of the art in face recognition, which covers the different processing stages related to this task. In particular, this review covers aspects of face detection and face recognition, which have been looked at from three different perspectives: psychology, neuroscience and computer vision.

From the psychology point of view there are several conclusions that can be drawn. From birth, humans have a particular interest in looking at faces, and familiar faces are recognised from early days. There is some evidence that there are differences in the way we recognise familiar and unfamiliar faces, in particular, by using internal or external features for recognition. Furthermore, there is evidence that face recognition is performed separately from other objects' recognition, and that the human performance on this task has a good degree of invariance to changes in pose and expression, achieving a very high degree of accuracy. This indicates the special nature of this task as a cognitive function, and suggests that it has an important

role in the child's development.

From a neuroscience point of view, some of the mechanisms behind such an efficient system have been studied. Most notably, it has been experimentally observed that an hierarchical system of increasing feature complexity and decreasing dimensionality of the data places a central role in the face recognition, and visual recognition in general. This hierarchical structure plays a central role in the main model presented in this thesis.

Finally, from a computer vision perspective, many algorithms have been developed in pursue of optimal verification rates. Despite a large number of different methods having been proposed, they share mostly the same structure. Firstly a feature extraction method is applied, followed by a dimensionality reduction algorithm. These two steps produce a representation of the face, commonly denominated feature vector. Then some sort of storage method is used to create a gallery of faces, and finally a classifier is applied in order to recognise a new face. Several methods from all these stages have been briefly presented in our literature review. After analysing the state of the art we propose four components that are essential in a biologically inspired face recognition system: 1) a face detection model, 2) a mechanism to extract the features relevant to the face, 3) a coding scheme for the face features, and 4) a theoretical memory model for faces.

An existing biologically inspired face detection detection model developed by Masquelier and Thorpe (2007) has been studied in order to find strong and weak features in regards to its usage in a face recognition framework context. This model is in line with the hierarchical structure of the visual system and is capable of learning face

features. Furthermore, several face recognition and detection algorithms which have been recently published and became very popular due to their performance, follow a similar structure by using deep convolutional networks for solving this problem (Sun et al., 2014; Taigman et al., 2014). We have improved the original model by introducing a feedback mechanism and changing some parameters in order to be capable of detecting faces with a single output neuron, determining the region of interest corresponding to the face, to get a face features representation better suitable for recognition, and also we have added the capability to detect faces from different poses, and label the pose detected. The features chosen to represent the face are the intermediate features in this model, located in the V1 complex cells layer. We propose two alternative binary coding schemes for the local orientations represented by these features. The features are biologically plausible. The binary coding schemes also implement biologically plausible mechanisms: a winner-take-all mechanism and a representation in which similar orientations are closer to each other in the feature space and farther away orientations are also further away in the feature space. Furthermore, the coding scheme has very interesting properties that are desirable when building a complex recognition system, since in one of the variants they automatically reduce the dimensionality by half and separate the basic features nicely in the feature space. We also have developed a demo application to demonstrate the improved model working and detecting and recognising faces directly from a webcam.

Then we analysed the quality of the proposed coding schemes regarding the identification rates achieved in a well-known face database (Georghiades et al., 2001),

according to variability in pose and lighting conditions. This analysis shows that an identification rate of 0.97 can be achieved in some of the test conditions, and this rate is at the same level of the algorithms used for comparison tested in the same data such as the algorithms tested by Jaiswal et al. (2011) and Vu and Caplier (2011).

Finally we propose an approach for the face recognition problem with the memorisation and recall of different poses. The proposed method uses a biologically plausible oscillatory neural network of spiking neurons for memorisation and recall of sequences of poses (Borisyuk et al., 2013). An STDP rule is used to adjust the synaptic weights during the learning process. This goal approach was to improve classification by recalling an entire memory of different views of the same subject, from a single initial view. This approach reduces the computational cost of searching the best match, and is in line with the biological and psychological findings that indicates that the recognition is a dynamic process, that in most cases involves matching not only a snapshot of the individual face, but rather a range of views that changes smoothly during the time of observation.

In conclusion, we presented a complete biologically plausible approach for face recognition, which covers several mechanisms, from the face detection and feature extraction and representation to the memorisation and classification of faces.

7.1 Contributions

The main contributions and achievements of this thesis are:

- A literature review that describes the state of the art in face recognition from

three different perspectives: Neuroscience, Psychology and Computer Vision..

- An improved version of a biologically inspired model for visual features extraction was developed. The improvements significantly increase the performance of the model by adding the possibility of detecting faces from multiple poses, and knowing which pose has been detected. Furthermore, the region of interest corresponding to the face has been made available through a feedback mechanism which has been added to the original model.
- A binary scheme of face features coding inspired by the brain has been proposed and tested. Using this scheme, high recognition rates are achieved (Silva Gomes and Borisyuk, 2012).
- A new biologically realistic model for memorisation and recall of sequences of images was developed (Borisyuk et al., 2013) and a theoretical framework was proposed for using this model to memorise sequences of poses and use the recall capabilities to improve the face recognition process.
- A demo application for face detection, memorisation and recognition using a webcam was developed. The goal of this application is to demonstrate how the proposed improved model and binary coding work in a real world scenario.

7.2 Future work

The following improvements and extensions to the two models presented in this thesis are proposed:

- Creation of a very efficient library implementing the face coding schemes and comparison method presented in this thesis. This implementation could

perform very fast comparisons between two faces because the binary coding schemes proposed in this thesis lead to a very small face feature vector, which could have as little as 2048 bits, depending on the network configuration and coding scheme chosen. This library could be used to very quickly narrow down the search space of a very large scale face database that posts a problem to many state of the art very accurate algorithms, which have a much larger template size.

- From my experience in the development of face recognition systems in the real world, it is clear that despite the level of sophistication of the current artificial algorithms particularly under a controlled environment, which can outperform humans in a one to one verification task, there is still a big room for improvement for the performance under uncontrolled environments. This improvement could come from bringing the ideas from biological systems, in particular, the memory model proposed in this thesis could be used to organise the gallery of faces and boost the performance of a one to many verification task.

Bibliography

Ahonen, Timo, Abdenour Hadid and Matti Pietikäinen (2004). 'Face recognition with local binary patterns'. In: *Computer vision-eccv 2004*. Springer, pp. 469–481.

Ahonen, Timo, Abdenour Hadid and Matti Pietikainen (2006). 'Face description with local binary patterns: Application to face recognition'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28.12, pp. 2037–2041.

Baron, Robert J (1981). 'Mechanisms of human facial recognition'. *International Journal of Man-Machine Studies* 15.2, pp. 137–178.

Bartlett, Marian Stewart and Terrence J Sejnowski (1997). 'Independent components of face images: A representation for face recognition'. In: *Procs. of the Annual Joint Symposium on Neural Computation, Pasadena, CA*.

Bartlett, Marian Stewart, Javier R Movellan and Terrence J Sejnowski (2002). 'Face recognition by independent component analysis'. *Neural Networks, IEEE Transactions on* 13.6, pp. 1450–1464.

Barwinski, Marek (2008). 'A neurocomputational model of memory acquisition for novel faces'. PhD thesis. Ruhr-University Bochum.

Bengio, Yoshua, Aaron Courville and Pierre Vincent (2013). 'Representation learning: A review and new perspectives'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.8, pp. 1798–1828.

Beymer, David J (1994). 'Face recognition under varying pose'. In: *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, pp. 756–761.

Biederman, Irving and Peter Kalocsais (1997). 'Neurocomputational bases of object and face recognition'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 352.1358, pp. 1203–1219.

Borisyuk, Roman, David Chik, Yakov Kazanovich and João da Silva Gomes (2013). 'Spiking neural network model for memorizing sequences with forward and backward recall'. *BioSystems* 112.3, pp. 214–223.

Bronstein, Alexander M, Michael M Bronstein, Eyal Gordon and Ron Kimmel (2004). 'Fusion of 2D and 3D data in three-dimensional face recognition.' In: *ICIP*, pp. 87–90.

Bronstein, Alexander M, Michael M Bronstein and Ron Kimmel (2005). 'Three-dimensional face recognition'. *International Journal of Computer Vision* 64.1, pp. 5–30.

Bruce, Charles, Robert Desimone and Charles G. Gross (1981). 'Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque.' *Journal of Neurophysiology* 46.2, pp. 369–384.

Bruce, Vicki and Andy Young (1986). 'Understanding face recognition'. *British Journal of Psychology* 77, pp. 305–327.

Brunelli, Roberto and Tomaso Poggio (1993). 'Face recognition: features versus templates'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.10, pp. 1042 –1052.

Buhmann, Joachim, Martin Lades and Christoph von der Malsburg (1990). 'Size and distortion invariant object recognition by hierarchical graph matching'. In: *IJCNN International Joint Conference on Neural Networks, 1990*, 411 –416 vol.2.

Burgess, Neil and John O'Keefe (2011). 'Models of place and grid cell firing and theta rhythmicity'. *Current Opinion in Neurobiology* 21.5. Networks, circuits and computation, pp. 734 –744.

Carey, Susan, Rhea Diamond and Bryan Woods (1980). 'Development of face recognition: A maturational component?' *Developmental Psychology* 16.4, pp. 257 –269.

Cevikalp, Hakan, Marian Neamtu, Mitch Wilkes and Atalay Barkana (2005). 'Discriminative common vectors for face recognition'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.1, pp. 4 –13.

Chai, Xiujuan, Shiguang Shan and Wen Gao (2003). 'Pose normalization for robust face recognition based on statistical affine transformation'. In: *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia*. Vol. 3, 1413–1417 vol.3.

Chen, Li-Fen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin and Gwo-Jong Yu (2000). 'A new LDA-based face recognition system which can solve the small sample size problem'. *Pattern Recognition* 33.10, pp. 1713 –1726.

Chua, Chin Seng and Ray Jarvis (1997). 'Point Signatures: A New Representation for 3D Object Recognition'. *Int. J. Comput. Vision* 25.1, pp. 63–85.

Colgin, L. L. and E. I. Moser (2006). 'Neuroscience: rewinding the memory record'. *Nature* 440.7084, pp. 615–7+.

Cortes, Corinna and Vladimir Vapnik (1995). 'Support-Vector Networks'. *Machine Learning* 20 (3), pp. 273–297.

Cox, David and Nicolas Pinto (2011). 'Beyond simple features: A large-scale feature search approach to unconstrained face recognition'. In: *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), 2011*, pp. 8–15.

Delorme, Arnaud and Simon J. Thorpe (2001). 'Face identification using one spike per neuron: resistance to image degradations'. *Neural Networks* 14.6-7, pp. 795–803.

Delorme, Arnaud and Simon J Thorpe (2003). 'SpikeNET: an event-driven simulation package for modelling large networks of spiking neurons'. *Network: Computation in Neural Systems* 14.4, pp. 613–627.

Delorme, Arnaud, Jacques Gautrais, Rufin van Rullen and Simon Thorpe (1999). 'SpikeNET: A simulator for modeling large networks of integrate and fire neurons'. *Neurocomputing* 26–27.0, pp. 989 –996.

DiCarlo, James J. and David D. Cox (2007). 'Untangling invariant object recognition'. *Trends in Cognitive Sciences* 11, pp. 333–341.

Eimer, Martin (2000). 'Effects of face inversion on the structural encoding and recognition of faces: Evidence from event-related brain potentials'. *Cognitive Brain Research* 10.1-2, pp. 145 –158.

Ellis, Hadyn D., John W. Shepherd and Graham M. Davies (1979). 'Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition'. *Perception* 8, 431–439.

Er, Meng Joo, Shiqian Wu, Juwei Lu and Hock Lye Toh (2002). 'Face recognition with radial basis function (RBF) neural networks'. *IEEE Transactions on Neural Networks* 13.3, pp. 697 –710.

Etemad, Kamran and Rama Chellappa (1997). 'Discriminant Analysis for Recognition of Human Face Images'. *Journal of Optical Society of America A* 14, pp. 1724–1733.

Farah, Martha J. (1996). 'Is face recognition 'special'? Evidence from neuropsychology'. *Behavioural Brain Research* 76.1-2. Advances in Understanding Visual Cortex Function, pp. 181 –189.

Freund, Yoav and Robert E Schapire (1997). 'A decision-theoretic generalization of on-line learning and an application to boosting'. *Journal of computer and system sciences* 55.1, pp. 119–139.

Fukushima, Kunihiko (1988). 'Neocognitron: A hierarchical neural network capable of visual pattern recognition'. *Neural networks* 1.2, pp. 119–130.

Gautrais, Jacques and Simon Thorpe (1998). 'Rate coding versus temporal order coding: a theoretical approach'. *Biosystems* 48.1–3, pp. 57 –65.

George, Nathalie, Raymond J. Dolan, Gereon R. Fink, Gordon C. Baylis, Charlotte Russell and Jon Driver (1999). 'Contrast polarity and face recognition in the human fusiform gyrus'. *Nature neuroscience* 2, pp. 574–580.

Georghiades, Athinodoros S., David J. Kriegman and Peter N. Belhumeur (2001). 'From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose'. *IEEE Trans. Pattern Anal. Mach. Intelligence* 23.6, pp. 643–660.

Gerstner, Wulfram and Werner M Kistler (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press.

Gordon, Gaile G. (1992). 'Face recognition based on depth and curvature features'. In: *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pp. 808 –810.

Guo, Guo-Dong and Hong-Jiang Zhang (2001). 'Boosting for fast face recognition'. In: *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings*, pp. 96–100.

Guo, Guo-Dong, Hong-Jiang Zhang and S.Z. Li (2001). 'Pairwise face recognition'. In: *Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings.* Vol. 2, 282 –287 vol.2.

Guo, Guodong, Stan Z. Li and Kapluk Chan (2000). 'Face Recognition by Support Vector Machines'. *IEEE International Conference on Automatic Face and Gesture Recognition* 0, p. 196.

Günther, Manuel and Rolf P. Würtz (2009). 'Face Detection and Recognition using maximum likelihood classifiers on Gabor graphs'. *International Journal of Pattern Recognition and Artificial Intelligence* 23.1, pp. 433–461.

Haan, Michelle de, Olivier Pascalis and Mark H. Johnson (2002). 'Specialization of Neural Mechanisms Underlying Face Recognition in Human Infants'. *Journal of Cognitive Neuroscience* 14.2, pp. 199–209.

Haig, N. D. (1984). 'The effect of feature displacement on face recognition'. *Perception* 13, pp. 505–512.

Han, Victor Z., Kirsty Grant and Curtis C. Bell (2000). 'Reversible Associative Depression and Nonassociative Potentiation at a Parallel Fiber Synapse'. *Neuron* 27, pp. 611 –622.

Haxby, James V., Leslie G. Ungerleider, Barry Horwitz, Jose M. Maisog, Stanley I. Rapoport and Cheryl L. Grady (1996). 'Face encoding and recognition in the human brain'. *Proceedings of the National Academy of Sciences* 93.2, pp. 922–927.

Haxby, James V., Elizabeth A. Hoffman and M. Ida Gobbini (2002). 'Human neural systems for face recognition and social communication'. *Biological psychiatry* 51, pp. 59–67.

Heisele, Bernd, Purdy Ho and Tomaso Poggio (2001). 'Face Recognition with Support Vector Machines: Global versus Component-based Approach'. *IEEE International Conference on Computer Vision* 2, p. 688.

Heisele, Bernd, Purdy Ho, Jane Wu and Tomaso Poggio (2003). 'Face recognition: component-based versus global approaches'. *Computer Vision and Image Understanding* 91.1-2. Special Issue on Face Recognition, pp. 6 –21.

Hodgkin, A.L. and A.F. Huxley (1952). 'A quantitative description of membrane current and its applications to conduction and excitation in nerve'. *Journal of Physiology* 117.1-2, pp. 500–544.

Huang, Gary B., Manu Ramesh, Tamara Berg and Erik Learned-Miller (2007). *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. 07-49. University of Massachusetts, Amherst.

Huang, Weilin and Hujun Yin (2009). 'Linear and nonlinear dimensionality reduction for face recognition'. In: *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, pp. 3337–3340.

Hubel, D. H. and T. N. Wiesel (1959). 'Receptive fields of single neurones in the cat's striate cortex.' eng. *Journal of Neurophysiology* 148, pp. 574–591.

Hubel, D. H. and T. N. Wiesel (1962). 'Receptive fields, binocular interaction and functional architecture in the cat's visual cortex.' *Journal of Neurophysiology* 160, pp. 106–54.

Hubel, D. H. and T. N. Wiesel (1963a). 'Receptive fields of cells in striate cortex of very young, visually inexperienced kittens.' eng. *Journal of Neurophysiology* 26, pp. 994–1002.

Hubel, D. H. and T. N. Wiesel (1963b). 'Shape and arrangement of columns in cat's striate cortex.' eng. *Journal of Neurophysiology* 165, pp. 559–568.

Hung, Chou P., Et Al, W. Li, J. T. Ohlmeyer, M. E. Lane, D. Kalderon, G. W. Davis, C. M. Schuster and C. S. Goodman (2005). 'Fast Readout of Object Identity from Macaque Inferior Temporal Cortex'. *Science* 310.5749, pp. 863–866.

Jafri, Rabia and Hamid R Arabnia (2009). 'A Survey of Face Recognition Techniques.' In: vol. 5. 2, pp. 41–68.

Jaiswal, Ajay, Ramesh K. Agrawal and Nitin Kumar (2011). 'Performance evaluation of linear subspace methods for face recognition under illumination variation'. In: *Proceedings of The Fourth International C* Conference on Computer Science and Software Engineering*. C3S2E '11. Montreal, Quebec, Canada: ACM, pp. 103–110.

Jeng, Shi-Hong, Hong Yuan Mark Liao, Chin Chuan Han, Ming Yang Chern and Yao Tsorng Liu (1998). 'Facial feature detection using geometrical face model: An efficient approach'. *Pattern Recognition* 31.3, pp. 273 –282.

Jitsev, Jenia and Christoph v. der Malsburg (2009). 'Experience-driven formation of parts-based representations in a model of layered visual memory'. *Frontiers in Computational Neuroscience* 4, p. 12.

Jones, Judson P. and Larry A. Palmer (1987). 'An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex'. *Journal of Neurophysiology* 58, pp. 1233–1258.

Jones, M. and P. Viola (2001). 'Rapid Object Detection using a Boosted Cascade of Simple Features'. In: *Computer Vision and Pattern Recognition 2001*. Submited to.

Kanwisher, Nancy and Galit Yovel (2006). 'The fusiform face area: a cortical region specialized for the perception of faces'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 361.1476, pp. 2109–2128.

Kong, Seong G., Jingu Heo, Besma R. Abidi, Joonki Paik and Mongi A. Abidi (2005). 'Recent advances in visual and infrared face recognition—a review'. *Computer Vision and Image Understanding* 97.1, pp. 103 –135.

Lam, Kin-Man and Hong Yan (1998). 'An analytic-to-holistic approach for face recognition based on a single frontal view'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.7, pp. 673 –686.

Lawrence, S., C.L. Giles, Ah Chung Tsoi and A.D. Back (1997). 'Face recognition: a convolutional neural-network approach'. *IEEE Transactions on Neural Networks* 8.1, pp. 98 –113.

Li, Stan Z. and Anil K. Jain, eds. (2005). *Handbook of Face Recognition*. 1st ed. Springer, Berlin.

Lin, Shang-Hung, Sun-Yuan Kung and Long-Ji Lin (1997). 'Face recognition/detection by probabilistic decision-based neural network'. *IEEE Transactions on Neural Networks* 8.1, pp. 114 –132.

Liu, Chengjun and H. Wechsler (2002). 'Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition'. *IEEE Transactions on Image Processing* 11.4, pp. 467 –476.

Liu, Chengjun and H. Wechsler (2003). 'Independent component analysis of Gabor features for face recognition'. *IEEE Transactions on Neural Networks* 14.4, pp. 919 –928.

Lu, Xiaoguang, Dirk Colbry and Anil K. Jain (2004). 'Three-Dimensional Model Based Face Recognition'. *International Conference on Pattern Recognition* 1, pp. 362–366.

Markram, H., J. Lübke, M. Frotscher and B. Sakmann (1997). 'Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs.' *Science* 275.5297, pp. 213–215.

Marr, David (1982). *Vision. A computational investigation into the human representation and processing of visual information.* New York: W.H. Freeman.

Marr, David and Tomaso Poggio (1976). 'Cooperative computation of stereo disparity'. *Science* 194.4262, pp. 283–287.

Masquelier, Timothée and Simon J. Thorpe (2007). 'Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity'. *PLoS Computational Biology* 3.2, e31.

Meyers, Ethan and Lior Wolf (2008). 'Using Biologically Inspired Features for Face Processing'. *International Journal of Computer Vision* 76 (1), pp. 93–104.

Moghaddam, B. and A. Pentland (1997). 'Probabilistic visual learning for object representation'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19.7, pp. 696–710.

Moghaddam, Baback, C. Nastar and A. Pentland (1996). 'Bayesian face recognition using deformable intensity surfaces'. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 0, p. 638.

Moghaddam, Baback, W. Wahid and A. Pentland (1998). 'Beyond eigenfaces: probabilistic matching for face recognition'. In: *Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings.* Pp. 30 –35.

Moghaddam, Baback, Tony Jebara and Alex Pentland (2000). 'Bayesian face recognition'. *Pattern Recognition* 33.11, pp. 1771 –1782.

Morton, John and Mark H. Johnson (1991). 'CONSPEC and CONLERN: A Two-Process Theory of Infant Face Recognition'. *Psychological Review* 98.2, pp. 164 –181.

Moscovitch, Morris, Gordon Winocur and Marlene Behrmann (1997). 'What Is Special about Face Recognition?: Nineteen Experiments on a Person with Visual Object Agnosia and Dyslexia but Normal Face Recognition'. *Journal of Cognitive Neuroscience* 9.5, pp. 555–604.

Müller, Marco K. and Rolf P. Würtz (2009). 'Learning from Examples to Generalize over Pose and Illumination'. In: *Proceedings of the 19th International Conference on Artificial Neural Networks: Part II.* ICANN '09. Limassol, Cyprus: Springer-Verlag, pp. 643–652.

Nefian, A.V. and III Hayes M.H. (1998). 'Hidden Markov models for face recognition'. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing.* Vol. 5, 2721 –2724 vol.5.

Nelson, Charles A. (2001). 'The development and neural bases of face recognition'. *Infant and Child Development* 10, pp. 3–18.

Neumann, John Von (1958). *The Computer and the Brain.* New Haven: Yale University Press.

Okada, Kazunori and Christoph von der Malsburg (2002). 'Pose-invariant face recognition with parametric linear subspaces'. In: *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on.* IEEE, pp. 64–69.

Palmeri, Thomas J. and Isabel Gauthier (2004). 'Visual object understanding'. *Nature Reviews Neuroscience* 5, pp. 291–303.

Papageorgiou, Constantine P, Michael Oren and Tomaso Poggio (1998). 'A general framework for object detection'. In: *Computer vision, 1998. sixth international conference on.* IEEE, pp. 555–562.

Patterson, K. E. and A. D. Baddeley (1977). 'When Face Recognition Fails'. *Journal of Experimental Psychology: Human Learning and Memory* 3, pp. 406–417.

Pentland, A., B. Moghaddam and T. Starner (1994). 'View-based and modular eigenspaces for face recognition'. In: *1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Proceedings CVPR '94.* Pp. 84 –91.

Perlibakas, Vytautas (2004). 'Distance measures for PCA-based face recognition'. *Pattern Recognition Letters* 25.6, pp. 711 –724.

Perrett, D. I., E. T. Rolls and W. Caan (1982). 'Visual neurones responsive to faces in the monkey temporal cortex.' eng. *Experimental Brain Research* 47.3, pp. 329–342.

Perrett, D. I., M. W. Oram and E. Ashbridge (1998). 'Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations.' eng. *Cognition* 67.1-2, pp. 111–145.

Phillips, Jonathon P. (1999). 'Support Vector Machines Applied to Face Recognition.' In: *Advances in Neural Information Processing Systems 11.* Vol. 11, pp. 803–809.

Quiroga, R. Quian, L. Reddy, G. Kreiman, C. Koch and I. Fried (2005). 'Invariant visual representation by single neurons in the human brain.' *Nature* 435.7045, pp. 1102–1107.

Riesenhuber, Maximilian and Tomaso Poggio (1999). 'Hierarchical models of object recognition in cortex.' *Nature Neuroscience* 2, pp. 1019–1025.

Rolls, E. T. and G. C. Baylis (1986). 'Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey.' *Experimental Brain Research* 65, pp. 38–48.

Rolls, Edmund T. (2000). 'Functions of the Primate Temporal Lobe Cortical Visual Areas in Invariant Visual Object and Face Recognition.' *Neuron* 27, pp. 205–218.

Rolls, Edmund T. (2008). 'Face processing in different brain areas, and critical band masking.' *Journal of Neuropsychology* 2, pp. 325–360.

Rumsey, Clifton C. and L. F. Abbott (2004). 'Synaptic equalization by anti-STDP.' *Neurocomputing* 58-60. Computational Neuroscience: Trends in Research 2004, pp. 359 –364.

Schwarz, Michael W., William B. Cowan and John C. Beatty (1987). 'An Experimental Comparison of RGB, YIQ, LAB, HSV, and Opponent Color Models'. *ACM Trans. Graph.* 6.2, pp. 123–158.

Senior, Andrew W. and Ruud M. Bolle (2002). 'Face Recognition and its Application'. English. In: *Biometric Solutions.* Ed. by David Zhang. Vol. 697. The

Springer International Series in Engineering and Computer Science. Springer US, pp. 83–97.

Shepard, Roger N. and Jacqueline Metzler (1971). 'Mental Rotation of Three-Dimensional Objects.' *Science* 171.3972, pp. 701–703.

Silva Gomes, João da and Roman Borisyuk (2012). 'Biological brain and binary code: quality of coding for face recognition.' In: *Artificial Neural Networks and Machine Learning. ICANN 2012.* Springer, pp. 427–434.

Sinha, P. and Tomaso Poggio (1996). 'Role of learning in three-dimensional form perception'. *Nature* 384.6608, pp. 460–463.

Soltani, Alireza and Christof Koch (2010). 'Visual saliency computations: mechanisms, constraints, and the effect of feedback'. *The Journal of Neuroscience* 30.38, pp. 12831–12843.

Sun, Yi, Xiaogang Wang and Xiaoou Tang (2014). 'Deep Learning Face Representation from Predicting 10,000 Classes'. In: *CVPR.*

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato and Lior Wolf (2014). 'DeepFace: Closing the Gap to Human-Level Performance in Face Verification'. In: *Conference on Computer Vision and Pattern Recognition (CVPR).*

Tanaka, H.T., M. Ikeda and H. Chiaki (1998). 'Curvature-based face surface recognition using spherical correlation. Principal directions for curved object recognition'. In: *Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings.* Pp. 372 –377.

Tanaka, J. W. and J. A. Sengco (1997). 'Features and their configuration in face recognition.' *Memory & Cognition* 25.5, pp. 583–592.

Tanaka, James W. and Martha J. Farah (1993). 'Parts and wholes in face recognition'. *The Quarterly Journal of Experimental Psychology* 46.2, pp. 225–245.

Tanaka, James W. and Iris Gordon (2011). 'Oxford Handbook of Face Perception'. In: ed. by Andrew J. Calder, Gillian Rhodes, Mark H. Johnson and James V. Haxby. Oxford University Press. Chap. Features, Configuration, and Holistic Face Processing, pp. 177–194.

Thorpe, Simon, Denis Fize, Catherine Marlot et al. (1996). 'Speed of processing in the human visual system'. *nature* 381.6582, pp. 520–522.

Torres, L., J.Y. Reutter and L. Lorente (1999). 'The importance of the color information in face recognition'. In: *1999 International Conference on Image Processing, 1999. ICIP 99. Proceedings.* Vol. 3, 627 –631 vol.3.

Tsao, Doris Y. and Margaret S. Livingstone (2009). 'Mechanisms of Face Perception'. *Annual Review of Neuroscience* 31, pp. 411–437.

Turk, Matthew and Alex Pentland (1991). 'Eigenfaces for recognition'. *Journal of Cognitive Neuroscience* 3 (1), pp. 71–86.

Vu, Ngoc-Son and Alice Caplier (2011). 'State of the art in Biometrics'. In: ed. by Jucheng Yang. InTech. Chap. Biologically Inspired Processing for Lighting Robust Face Recognition, pp. 123–142.

Wang, Yingjie, Chin-Seng Chua and Yeong-Khing Ho (2002). 'Facial feature detection and face recognition from 2D and 3D images'. *Pattern Recognition Letters* 23.10, pp. 1191 –1202.

Wiskott, L., J.-M. Fellous, N. Kuiger and C. von der Malsburg (1997). 'Face recognition by elastic bunch graph matching'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19.7, pp. 775 –779.

Wolf, Lior, Tal Hassner and Itay Maoz (2011). 'Face recognition in unconstrained videos with matched background similarity'. In: *Conference on Computer Vision and Pattern Recognition (CVPR).*

Xie, Xudong and Kin-Man Lam (2006). 'An efficient illumination normalization method for face recognition'. *Pattern Recognition Letters* 27.6, pp. 609 –617.

Yamaguchi, O., K.. Fukui and K.-i. Maeda (1998). 'Face recognition using temporal image sequence'. In: *Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings.* Pp. 318 –323.

Yang, Jian, David Zhang, Alejandro F. Frangi and Jing yu Yang (2004). 'Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, pp. 131–137.

Yarbus, Alfred Lukyanovich (1967). *Eye Movements and Vision.* New York: Plenum Press.

Yin, Robert K. (1970). 'Face recognition by brain-injured patients: A dissociable ability?' *Neuropsychologia* 8.4, pp. 395 –402.

Young, Andrew W. (1998). *Face and Mind.* Oxford University Press, USA.

Zhang, Cha and Zhengyou Zhang (2010). *A Survey of Recent Advances in Face Detection.* Tech. rep. MSR-TR-2010-66.

Zhao, W., R. Chellappa, P. J. Phillips and A. Rosenfeld (2003). 'Face recognition: A literature survey'. *ACM Computing Surveys* 35 (4), pp. 399–458.

Bound copy of published papers

Biological Brain and Binary Code: Quality of Coding for Face Recognition

João da Silva Gomes and Roman Borisyuk

School of Computing and Mathematics, University of Plymouth,
PL4 8AA Plymouth, United Kingdom
{joao.dasilvagomes,R.Borisyuk}@plymouth.ac.uk
http://www.plymouth.ac.uk/schools/compmath

Abstract. A computational model for face feature extraction and recognition capable of achieving a high degree of invariance to illumination and pose is presented. Similar to the complex V1 cells, the model uses a sparse binary code to represent an edge orientation. The binary code represents the face features for recognition. This paper investigates the geometrical structure of the linear space of face representation vectors. For this study the Yale Face Database B is used. It is shown that the biologically inspired procedure provides the face representation of a good quality: vectors representing the faces of the same person under different poses and illumination conditions are grouped together in the vector space. This code enables a very high recognition rate for both the illumination invariance and pose invariance settings.

Keywords: Face Recognition, Face Detection, HMAX, V1 Features, Complex Cells, Simple Cells.

1 Introduction

Invariant face recognition regarding a pose and illumination is a problem solved effortlessly by the human brain, but computational details underlying such an efficient recognition are still far from clear. However, some details on face recognition in primate's brain are known and they provide an inspiration for developing new computational models for face recognition. (1) An hierarchical processing has a central role in face recognition, starting with simple edge responsive cells in the primary visual cortex and, as the information flows through the ventral stream, new cells respond to more and more complex features [7,5] until reaching the IT where there is a single cell responding to a face stimulus [4,1,10]. (2) A highly efficient neural coding, in conjunction with very fast and hierarchical processing through the ventral stream, fulfills an important memory function for faces and other objects [2]. (3) A synaptic plasticity enables the visual system to adapt and learn new object representations, e.g., faces [8]. (4) The brain learns the appearance of a face (or other 3D object) throughout an experience by observing the face for many times [12]. Each time a pose is different as well as an illumination condition, therefore the brain can construct a memory of the face based upon this information.

A.E.P. Villa et al. (Eds.): ICANN 2012, Part I, LNCS 7552, pp. 427–434, 2012.

All mechanisms mentioned above are included in our model, therefore we claim to present a representation and face recognition system which is biologically inspired.

This paper is organized in five sections. In the second section, we describe briefly the face detector. In the third section we explain a process of feature extraction and a coding scheme for the feature representation and memorization. In the fourth section, a study of geometrical properties of multidimensional vector space of face representation is described. The last section is to discuss and conclude the study of the feature vector space.

2 Face Detection

The face detection is performed using a HMAX algorithm [11] and the [8]. This network has four layers: S1 (simple V1 cells), C1 (complex V1 cells), S2 (V4 cells), and C2 (V4/PIT cells). There are four kinds of simple V1 cells, all of which respond to bars of different orientations: $\pi/8$, $3\pi/8$, $5\pi/8$, and $7\pi/8$. Due to a hierarchical structure of the network , there is an alternation between max and sum operations which makes the system robust to scale and shift variation. Illumination invariance is achieved by using an edge information instead of pixel values. Simple V1, complex V1, V4 pathway is replicated at five different scales of the input image which also makes the system to be a scale invariant.
In order to train the system to detect faces, a set of face images is presented and connections between complex V1 cells and V4 cells are adjusted according to a STDP rule.

We have modified the parameters of this face recognition system [8] aiming to define which configuration will lead to best coding of faces. In particular the number of output V4/PIT cells O, and the receptive field size of V4 cells S have been varied. In the first case we used the values $O = 1$ and $O = 3$, which means that the output layer will respond to a single face feature (the whole face), or to three different features (different parts of the face), respectively. For the receptive field size we use either 32×32 or 64×64 grid of complex V1 cells.

We also introduce a mechanism to determine the regions of complex V1 cells corresponding to activated neurons of the output V4/PIT layer, i.e., the Region Of Interest (ROI) corresponding to the face, or face features in a case if there is more than one output neuron. This can be seen as a feedback connection that drives the attention mechanism to the face area. This improvement of the detection algorithm is crucial for the integration of the face detection system with the face recognition mechanism, because it allows us to use only responses of complex V1 cells which relate to the face stimulus, instead of using responses from all neurons.

3 Feature Extraction and Encoding

The features used to represent the faces are the responses from the complex V1 layer. Only the area corresponding to the face stimulus is used for this

purpose. This area is determined by the propagation of the activity from the output V4/PIT layer to the complex V1 layer. These complex V1 cells respond to different bar orientations. The orientation with the strongest response at each position of the grid is then coded by a binary value in order to reduce the dimensionality. Thus, a winner take all approach is used to prescribe the winning orientation at each pixel.

Two coding schemes are considered: coding a single orientation by 2 or 4 binary values (see Figure 1).

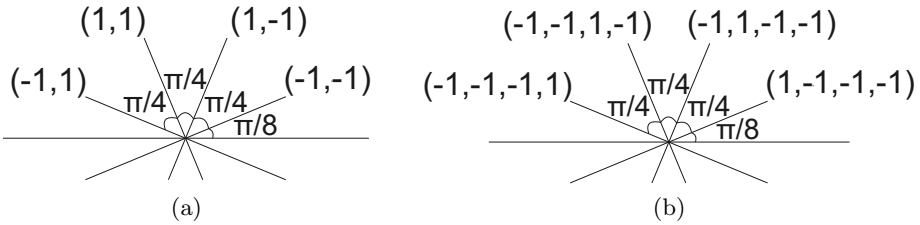


Fig. 1. (a) First coding scheme: two binary values are used to encode a particular orientation. The values chosen are such that immediate neighbour orientations (for instance $\pi/8$ and $3\pi/8$) have only one non-coincident components, while the number of non-coincident components between orientations that are not neighbour is two. (b) Second coding scheme: 4 binary values are used to encode a particular orientation. In this case only one of the 4 binary values can be 1 at a time, defining the orientation being coded. In this case the number of non-coincident components between any two orientations is two.

The binary encoded orientations for each position are concatenated per row in order to form the feature vector representing a face (see Figure 2).

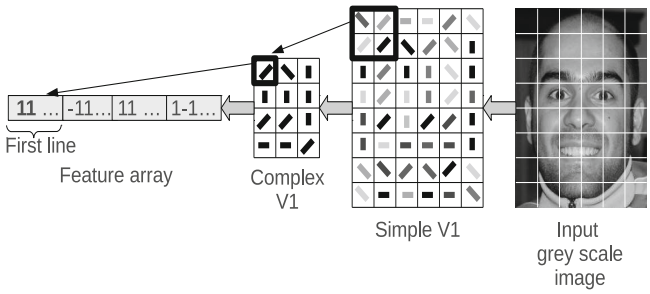


Fig. 2. Feature extraction and coding process. The area shown in all layers corresponds to the region of interest and the scale determined by one V4/PIT cell, through the feedback process. *Simple V1* shows only the direction with highest value for each position (darker directions have higher values). *Complex V1* shows the winner orientations from the *Simple V1 layer*, which are coded and concatenated to form the feature array.

4 Geometry of Multidimensional Coding Space

Here we investigate a geometrical structure of multidimensional space of vectors (V) representing faces. We expect that a set of vectors corresponding to faces of the same person under the variation of pose and illumination is a compact set and has no or small overlap with a set of vectors corresponding to faces of another person. To test this hypothesis we study the geometrical properties of the vector space for different schemes of face representation. The scheme of face representation by two binary values coding the orientation in the grid 32×32 provides coding vectors of length 2,048 (the smallest dimension of the vector space). The highest dimension (49,152) is provided by a scheme with four binary values coding the orientation in the grid 64×64, and there are three such grids corresponding to three output neurons.

4.1 Database of Face Images

The database chosen to investigate the spaces V was the Yale Face Database B [3]. This database contains a large variability of each face captured in different poses and under different illumination conditions. The database has 5,696 images of 10 subjects. For each subject there are 64 images with different illumination settings for every of the 9 different poses (i.e. 576 images per subject, except one of the subjects which only has 512).

4.2 Procedure to Compare Feature Vectors

Given a subset of images from the database which we will call the gallery, the matrix G is constructed. Each row of G contains a binary representation of the winning orientations in the face ROI of the complex V1 cells for a given face. The matrix G is of size $M_G \times N$, where M_G is the number of images in the gallery and N is the length of the binary vector representing a face. In a similar fashion, a matrix R is constructed based on other images which are not included to the gallery.

The number of coinciding components is a similarity measure for comparison of two face representation vectors (respectively, the distance between two vectors is the number of non-coinciding components).

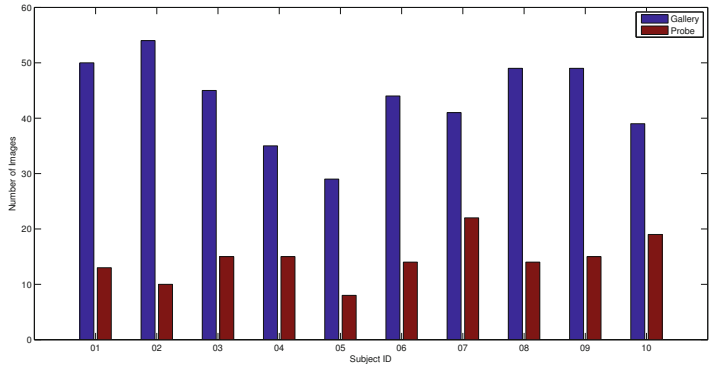
Let us assume that vector y does not belong to the gallery. To compare this vector with gallery vectors, we use the following formula:

$$z = Gy^T, \tag{1}$$

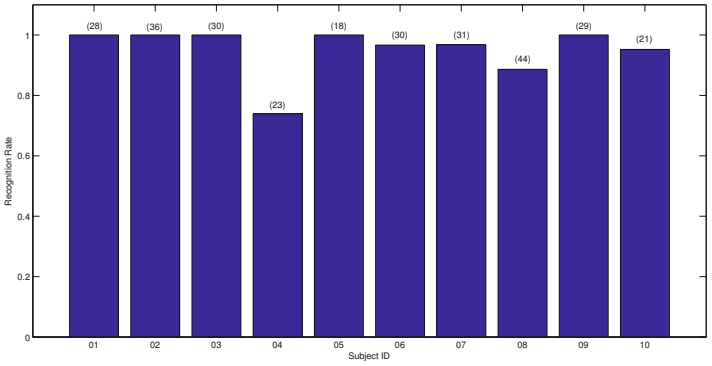
where y^T is a vector-column. The index k of the largest component of the vector z corresponds to the row of the matrix G which is the most similar to y and therefore this index also corresponds to the image in the gallery. Thus, the procedure for comparison of vectors can be expressed in terms of manipulations with matrices (multiplication and finding an index of maximum element) which drastically accelerates computations. It takes only 5.62 seconds to compare 5,696 face vectors of size 2,048 to a gallery of 5,696 faces using Matlab in a desktop computer.

4.3 Results of Feature Vector Comparison

For the scheme with vector length 8,192 (2 bit feature encoding, three output neurons and a receptive field of 64×64), we tested the hypothesis that for each image from the database, the best match is an image of the same person. The following procedure was used: select image from the database; compare this image with all other images, find the most similar image, and verify that IDs of selected image and the best match are the same. Repeating this procedure for all images from the database, it was found that the rate of correct "identification" is 0.9896.



(a)



(b)

Fig. 3. (a) Number of images selected in one of the runs for the gallery and identification sets, for each subject. In this example 3/4 of the images were chosen for the gallery and 1/4 for the identification set. (b) Example of correct identification rate for each of the subjects. The values in parenthesis are the number of identification images for a given subject. In this example 1/2 of the images were chosen for the gallery and 1/2 for the identification set.

Table 1. (m_{ir}) and (σ_{ir}) for all the different face recognition settings (number of output neurons O; receptive field size S), feature representation (2 and 4-bit coding), gallery/matching setups (illumination il; pose po) and portion of images used for the gallery ($F\epsilon\{1/2, 3/4, 7/8\}$). The highest correct matching rate for each gallery/matching setting is highlighted. For O=1, $S = 32 \times 32$ and Illumination setting, 580 images were used in total; for O=1, $S = 32 \times 32$ and Pose setting, 88 images were used in total; for O=1, $S = 64 \times 64$ and Illumination setting, 460 images were used in total; for O=1, $S = 64 \times 64$ and Pose setting, 86 images were used in total; for O=3, $S = 32 \times 32$ and Illumination setting, 630 images were used in total; for O=3, $S = 32 \times 32$ and Pose setting, 88 images were used in total; for O=3, $S = 64 \times 64$ and Illumination setting, 638 images were used in total; for O=3, $S = 64 \times 64$ and Pose setting, 88 images were used in total. This numbers reflect the number of faces available for validation, which are only the ones that were successfully detected by the face detector.

			$O=1$				$O=3$			
			$S=32\times32$		$S=64\times64$		$S=32\times32$		$S=64\times64$	
			2-bit	4-bit	2-bit	4-bit	2-bit	4-bit	2-bit	4-bit
1/2	il	m_{ir}	0.961	0.9623	0.9671	0.9628	**0.9742**	0.9558	0.9735	0.9309
		σ_{ir}	0.0083	0.0125	0.011	0.0116	0.0098	0.012	0.0077	0.0151
	po	m_{ir}	0.6286	0.6273	0.8465	0.8028	0.7714	0.7832	**0.9473**	0.9273
		σ_{ir}	0.0639	0.0832	0.051	0.0803	0.0647	0.0703	0.0548	0.0495
3/4	il	m_{ir}	0.9644	0.9688	0.9701	0.972	**0.9818**	0.9689	0.9775	0.946
		σ_{ir}	0.0124	0.0098	0.0163	0.0156	0.0108	0.0147	0.0105	0.0174
	po	m_{ir}	0.7136	0.7464	0.9076	0.8686	0.8564	0.8364	**0.9827**	0.97
		σ_{ir}	0.0781	0.0817	0.0435	0.0642	0.0683	0.0729	0.0317	0.0339
7/8	il	m_{ir}	0.9714	0.975	0.9811	0.9681	**0.9831**	0.9741	0.9777	0.9511
		σ_{ir}	0.0178	0.0161	0.0169	0.0212	0.0128	0.0191	0.0161	0.183
	po	m_{ir}	0.7436	0.7818	0.892	0.902	0.8782	0.8655	**0.9873**	0.9855
		σ_{ir}	0.1172	0.1246	0.0778	0.0769	0.0947	0.0993	0.0319	0.0337

After this encouraging result, we use more sophisticated procedure for investigating the coding space.

In order to study further the properties of the vector space in relation to images taken under different illumination and pose conditions we used two setups for defining the gallery and matching data:

1. Illumination (il): The frontal pose[1] is fixed and a fraction F of the images with this pose but different illumination was randomly (uniformly) selected

[1] Frontal or pose '00' according to [3] specifications.

for the gallery. The remaining images for the same pose are used for identification.

2. Pose (po): The frontal illumination[2] is fixed and a fraction F of the images with this illumination but different pose was randomly (uniformly) selected for the gallery. The remaining images for the same illumination are used for identification.

We studied a variety of cases corresponding to different values of F (1/2, 3/4, and 7/8), S, O and the number of binary values for each orientation. For each case a gallery was generated and the matching procedure described earlier was repeated for 50 runs.The gallery is generated independently from the previous runs. Figure 3 shows one example how these images are distributed across all subjects for the case where 3/4 of the images chosen for the gallery and 1/4 for the probe set, as well as the correct identification rate per subject for another example, where 1/2 of the images chosen for the gallery and 1/2 for the identification set.

The mean correct identification rate (m_{ir}) over 50 runs and the corresponding standard deviation (σ_{ir}) are shown in Table 1.

The result shown in Table 1 evidences that a simple comparison with the gallery provides a good face "identification" which is comparable with results of [6,9]. The best results are achieved when the number of output features is 3, as opposed to a single feature covering the whole face. The 2-bit representation of individual features also leads to better results than the 4-bit representation. This is expected because the 2-bit representation minimizes the similarity between further apart orientations, and maximizes the distance between closer orientations, as opposed to the 4-bit representation where the distance between any two orientations is always constant. For the pose test cases, it is better to use a 64×64 receptive field, while for the illumination cases, a 32×32 receptive field leads to better results.

5 Conclusion

We present a face features extraction and coding scheme, which achieved identification rates at the same level as some well known face recognition algorithms. This outcome is achieved, due to a combination of methods taken from a neurobiological face recognition system and from efficient computer science algorithms. In particular, a process of face encoding is integrated with the face detection, by introducing a feedback mechanism and using the already extracted features from the complex V1 layer.

Our binary coding of features enables an efficient representation of faces and a quick simple comparison with other vectors, which achieved correct identification rates higher than 0.97 in most of the cases.

[2] Frontal illumination or illumination source direction with respect to the camera axis is at 0 degrees azimuth ('A+000') and 0 degrees elevation ('E+00') according to [3] specifications.

Moreover, we have used clustering algorithm to investigate a set of face vectors. The cluster analysis reveals that the vectors are divided into ten clusters (related to ten faces in the database); each cluster mostly includes (with a small error) the vectors corresponding to faces of the same person. Details of the cluster analysis will be described in a separate publication.

Acknowledgments. We would like to thank to Timothée Masquelier for providing his MATLAB code for learning visual features [8]. We would like to acknowledge also the authors of Yale Face Database B [3] for providing the face database. J.D.S.G. thanks to European Neural Network Society for the travel grant Award.

References

1. Bruce, C., Desimone, R., Gross, C.G.: Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. Journal of Neurophysiology 46(2), 369–384 (1981), http://www.ncbi.nlm.nih.gov/pubmed/6267219
2. Carlson, E.T., Rasquinha, R.J., Zhang, K., Connor, C.E.: A sparse object coding scheme in area v4. Current Biology 21(4), 288–293 (2011), http://linkinghub.elsevier.com/retrieve/pii/S0960982211000364
3. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intelligence 23(6), 643–660 (2001)
4. Gross, C.G., Miranda, R.C.E., Bender, D.B.: Visual properties of neurons in inferotemporal cortex of the Macaque. J. Neurophysiol. 35(1), 96–111 (1972)
5. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. 160, 106–154 (1962)
6. Jaiswal, A., Agrawal, R.K., Kumar, N.: Performance evaluation of linear subspace methods for face recognition under illumination variation. In: Proceedings of The Fourth International C* Conference on Computer Science and Software Engineering, C3S2E 2011, pp. 103–110. ACM, New York (2011), http://doi.acm.org/10.1145/1992896.1992909
7. Jones, J.P., Palmer, L.A.: An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. Journal of Neurophysiology 58, 1233–1258 (1987)
8. Masquelier, T., Thorpe, S.J.: Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Comput. Biol. 3(2), e31 (2007), http://dx.plos.org/10.1371%2Fjournal.pcbi.0030031
9. Vu, N.-S., Caplier, A.: Biologically Inspired Processing for Lighting Robust Face Recognition. State of the art in Biometrics (July 2011), http://intechopen.com/books/state-of-the-art-in-biometrics/biologically-inspired-processing-for-lighting-robust-face-recognition
10. Perrett, D.I., Rolls, E.T., Caan, W.: Visual neurones responsive to faces in the monkey temporal cortex. Exp. Brain Res. 47(3), 329–342 (1982)
11. Riesenhuber, M., Poggio, T., Studies, E.: Hierarchical models of object recognition in cortex (1999)
12. Sinha, P., Poggio, T.: Role of learning in three-dimensional form perception. Nature 384(6608), 460–463 (1996), http://www.ncbi.nlm.nih.gov/pubmed/8945472

BioSystems 112 (2013) 214–223

Contents lists available at SciVerse ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Spiking neural network model for memorizing sequences with forward and backward recall

Roman Borisyuk[a,b,*], David Chik[c], Yakov Kazanovich[b], João da Silva Gomes[a]

[a] School of Computing and Mathematics, University of Plymouth, UK
[b] Institute of Mathematical Problems in Biology, Russian Academy of Sciences, Russia
[c] Department of Brain Science and Engineering, Kyushu Institute of Technology, Japan

ARTICLE INFO

Article history:
Received 12 October 2012
Received in revised form 22 February 2013
Accepted 26 March 2013

Keywords:
Memory of sequences
Spiking neuron model

ABSTRACT

We present an oscillatory network of conductance based spiking neurons of Hodgkin–Huxley type as a model of memory storage and retrieval of sequences of events (or objects). The model is inspired by psychological and neurobiological evidence on sequential memories. The building block of the model is an oscillatory module which contains excitatory and inhibitory neurons with all-to-all connections. The connection architecture comprises two layers. A lower layer represents consecutive events during their storage and recall. This layer is composed of oscillatory modules. Plastic excitatory connections between the modules are implemented using an STDP type learning rule for sequential storage. Excitatory neurons in the upper layer project star-like modifiable connections toward the excitatory lower layer neurons. These neurons in the upper layer are used to tag sequences of events represented in the lower layer. Computer simulations demonstrate good performance of the model including difficult cases when different sequences contain overlapping events. We show that the model with STDP type or anti-STDP type learning rules can be applied for the simulation of forward and backward replay of neural spikes respectively.

In Memoriam of Prof Luigi M. Ricciardi

It was in 1973 when I (RB) first learnt about Prof Luigi Ricciardi. At that time I had graduated from the Moscow State University and started my scientific career at the Biological Centre of the Soviet Union Academy of Science in Pushchino. My supervisor Dr. Vitaly Kryukov had just developed a new theory for the probabilistic modeling of spiking neurons where the first passage problem was approached using a version of the Wald's identity adapted to the non-constant boundary. Of course, I was told that there was a very strong group of mathematicians in Italy, led by Professor Ricciardi, working on the first passage problem with application to the modeling of neuronal activity. Shortly after the publication of the paper (Kryukov, 1976), an invitation from Italy arrived. Luigi wrote to Dr. Kryukov inviting him to Naples and offering to cover his travel and living expenses related to the visit. At that time it was very rare for a scientist in the USSR to be allowed to travel abroad, especially to a non-socialist country. Of course, permission from the Communist party was required. Dr. Kryukov's application for the travel was immediately rejected. The main reason was that Dr. Kryukov was a religious person and sang in a church choir. Luigi was very surprised when he heard that Vitaly Kryukov was not able to come to Italy. No explanation was given at the time. It was only at BIOCOMP 2002 that I told this amazing story to Luigi.

I am grateful to Laura Sacerdote for sending me a copy of the remarkable paper by Ricciardi and Umezawa (1967). The paper is short but it contains many important thoughts and ideas. Although it was written 45 years ago many questions, statements, and approaches which were formulated in the paper are timely, important, and of great interest. In fact, this paper formulates a programme of brain studies for several generations of researchers. For example, in Ricciardi and Umezawa (1967) it is suggested that long-term memory is related to "the ground state" of a large system of many interactive units (probably a quantum system) and "short-term memory can be related to the existence of meta-stable excited states". This idea was developed in detail by our Laboratory of Neural Networks in Pushchino under the leadership of Dr. Kryukov (see Kryukov et al., 1990). Another paper (Borisyuk and Hoppensteadt, 2004) that tried to answer the question: "How can a brain maintain stable memories and behaviors when its underlying electrical and chemical structures are constantly changing?" was discussed by the authors with Luigi Ricciardi at BIOCOMP 2002 at the stage of manuscript preparation. Many very fruitful advises were given by Luigi. The current paper continues this line of research on the neuronal mechanism of memory and demonstrates how oscillatory states and synchronous dynamics can be used for memorizing sequences of events.

* Corresponding author. Tel.: +44 1752584949.
E-mail address: r.borisyuk@plymouth.ac.uk (R. Borisyuk).

0303-2647/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.biosystems.2013.03.018

1. Introduction

Our memory is not a disordered store of incoherent items. In most cases our brain is inclined to organize our experience in sequences of events, actions, images, symbols, thoughts, etc. Recalling or perceiving a single member of a sequence is enough to allow us to quickly restore all the following items, which usually appear in our memory one by one in the same order as they have been learnt. The tasks of sequence storage and recall are not always simple. Errors may appear when recalling a complex sequence, especially if it is correlated with another sequences. The recall can be improved if the sequence is labeled by some tag associated with the context in which the sequence has appeared. We hypothesize that these tags are kept in the higher areas of the cortex and that their interaction with representations of memorized items is implemented through synchronization of activity between the frontal and associative cortices which results in the modification of connections between these regions. This mechanism has been used to design a biologically plausible neural network to improve the recall of complex sequences and to reproduce some known effects of sequence storage.

A complex task usually consists of several consecutive steps that should be fulfilled in a particular time order. How do we learn sequences of memory items? What is the neural mechanism of storage of sequences of events (objects)? In this paper we address these questions and describe a neural network model for the storage of several sequences. The model is constructed from conductance based spiking elements of the Hodgkin–Huxley type which are arranged into two interactive layers. A lower layer represents consecutive events of a sequence in the form of the activity of the modules which are composed of excitatory and inhibitory neurons with all-to-all coupling. Bottom-up convergent connections are directed from the bottom layer to the upper layer which represents the "tags" that are attached to different sequences. Feedback connections to the bottom layer allow the system to distinguish multiple sequences during recall. The rule for synaptic modification is of an STDP (or anti-SDP) type.

The model is biologically inspired and takes into account some well-known facts from neurobiology. However, we do not present this work as a significant contribution to modeling memory and recall processes in the hippocampus or any other brain structure. This would demand consideration of many details of functioning that are specific for these structures. On the contrary, we focus on some general principles which are hypothetically universal for different brain structures and which may be helpful in avoiding some errors in the recall. The reason to pursue this investigation is the hope that it will give pay-offs in neuroscience as other theoretical models do (Borisyuk et al., 1999, 2001; Borisyuk and Hoppensteadt, 2004).

Neuronal mechanisms of memory have recently come under intensive investigation. Recent results clarify some important details of memory processes. In particular, the hippocampal neuronal activity (including place cell firing) has been studied in this context (see, e.g., Foster and Wilson, 2006; Diba and Buzsaki, 2007). It has been reported that when a rat reaches the end of a track, the hippocampal place cells, which fired sequentially during the run, can generate spikes in the ***reverse*** order in a short time window. Lee and Wilson (2002) found that hippocampal CA1 place cells repeatedly fire in the correct sequential order during slow wave sleep immediately following the experience. Davidson et al. (2009) state that firing sequences corresponding to long runs can be robustly replayed with high speed and that this firing is coherent with high frequency ripple events. Thus, the feed-forward or backward replay of behavioral sequences in the hippocampus coherent with sharp wave ripples is considered as a possible mechanism of learning and encoding recent experiences. The paper (Euston et al., 2007) reports that feed-forward and backward replay of recent memory sequences can be seen in the prefrontal cortex during sleep. This spatio-temporal activity is coherent and compressed in time by a factor of seven.

Significant progress has been made in the mathematical and computational modeling of memory formation and recall. For a review on memory encoding based on the dynamics of neural activity, membrane potential oscillations, resonance, and bistable persistent spiking see (Hasselmo et al., 2010).

Several challenging problems arise when creating a model for sequence storage:

1. How to handle sequences with repeating or overlapping elements which may lead to ambiguity during recall?
2. How to adjust a model of sequential memory to experimental evidence, including the data supporting a role for rhythmic activity and synchronization in memory and attention tasks (Singer and Gray, 1995; Malsburg, 2001; Fries et al., 2002; Gregoriou et al., 2009; Melloni et al., 2007)?

None of the known models satisfies both of these demands. Some models can distinguish complex sequences (e.g. Wang and Yuwono, 1996; Scarpetta et al., 2002, 2010) but they are formulated in terms of non-conduction based neuronal networks that have no clear biological interpretation. Other models are in better agreement with neurobiology (e.g. Yamaguchi, 2003; Hopfield and Brody, 2009) but they can only deal with simple sequences. Detailed biophysical models of the CA1 microcircuit are developed in (Cutsuridis et al., 2010; Cutsuridis and Hasselmo, 2012). These models include pyramidal neurons and several types of inhibitory interneurons and can simulate the timing of firing of hippocampal neurons in relation to the theta rhythm as well as the organization of activity in a correct order of sequential memories. The problem of possible ambiguity of the recall for complex sequences is left beyond the scope of the models. Koene and Hasselmo (2008) use different phases in the theta cycle to label different locations of a rat. This approach requires a reliably stable period of theta oscillations while the frequency of the biological theta rhythm can vary widely.

Generally speaking, there are two major methods to specify the order of the events in a sequence which have been used in previous models:

1. Chain method: the events are linked in the order prescribed by the sequence. This approach was put forward by Ebbinghaus (1885/1964) in the pioneer psychological experiments on sequence storage. Its early mathematical implementation in the context of a Central Pattern Generator can be found in Kleinfeld (1986).
2. Labeling method: the order of the events is prescribed by some ordered "tags" which are attached to the events (Grossberg, 1978; Borisyuk and Hoppensteadt, 2004).

Our model combines both chain and labeling mechanisms. The chain method is realized in the bottom layer where ordered sequences of events are represented for storage and recall. The "tags" are prescribed to neurons in the upper layer to associate each neuron in this layer with a particular sequence coded in the bottom layer. Such tags are used to assist in finding a proper order of events in the case when a confusing situation appears with overlapping elements being present in two different sequences.

The following advantages of our model of associative learning and recall of multiple sequences should be emphasized:

1. The system can reliably memorize and recall several sequences of ordered events. To start the recall we have to decide which

sequence should be replayed by presenting to the model the first (or any other) member of the selected sequence as well as the "tag" of the selected sequence.
2. The functioning of the model is based on oscillatory activity in the theta band. This allows us to separate ordered events by presenting each of the events during one theta period.
3. The whole selected sequence (or subsequence starting from the presented event) will be replayed in the time order of the events comprising the sequence. The recall process is reliable and short: the sequence is replayed during one cycle of the theta rhythm.
4. The system can handle multiple sequences including those containing overlapping members. The "tag" signal from the top layer enables the model to select a proper continuation in an ambiguous situation when the same event is present in two different sequences.
5. The system with the anti-STDP learning rule allows the reverse replay of neural activity which reflects some experimental data observed in hippocampal place cells
6. The model is composed of biologically relevant conductance-based spiking elements operating in the theta frequency band and being robust against frequency variation.

The model works in continuous time but for encoding and recall the time is divided into discrete intervals of 200 ms duration. This is used to reflect the experimental evidence on the discrete nature of signal processing in the brain with time windows conditioned by the theta rhythm (Stella and Treves, 2011). Another exciting example of the theta state correlated with episodic memory is reported in (Guderian et al., 2009). MEG recordings from the temporal lobe show that the amplitude of theta oscillations is higher in the case of a memory event than in the case of the preparatory state.

The leaning rule used in the model for updating connection strengths takes into account the activity of pre- and post-synaptic neurons in two sequential time intervals. This rule is in line with recent developments on the STDP type learning algorithm (Dan and Poo, 2004, 2006; Clopath et al., 2010; Cutsuridis, 2013).

The paper has the following structure: Section 2 contains a description of the model; Section 3 explains how the model represents and stores sequences of events and shows the results of computer simulations. These results are discussed in Section 4.

2. Model formulation

2.1. Description of the currents

Our model is based on the oscillatory activity of spiking elements and synchronous dynamics. A recent review on brain oscillatory activity and memory (Burgess and O'Keefe, 2011) states that theta and gamma oscillations play a key role in storing both single events and sequences of events, novelty detection, synaptic plasticity, etc. The oscillatory mechanisms of memory which were first identified in rodents also play a significant role in other species. The storage of sequences of events represents an interesting example of the capabilities and advantages of the oscillatory approach to the study of memory. We show that networks with star-like connectivity (i.e. having a central element) can be a useful instrument for this study.

The model design is grounded on our previous experience with oscillatory network models. It has shown to us that partial synchronization of neural activity in networks with a central element can be used as a key neuronal mechanism for the modeling of various cognitive functions including *novelty detection* (Borisyuk et al., 2001), *selective visual attention* (Chik et al., 2009), *moving object tracking* (Borisyuk et al., 2008, 2009b), and *perception* of ambiguous figures (Borisyuk et al., 2009a). This stimulated us to use the architecture with a central element as a building block of the model that represents a single sequence of events.

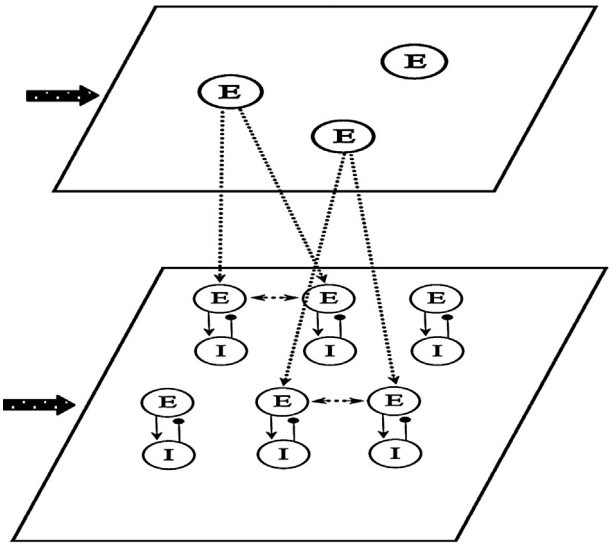


Fig. 1. Model architecture. The top layer contains the groups of excitatory neurons. The bottom layer contains the groups of coupled excitatory and inhibitory neurons. Each neural group contains the populations of excitatory neurons (E) and inhibitory neurons (I). Excitatory (inhibitory) neurons deliver excitatory (inhibitory) connections to all other (both excitatory and inhibitory) neurons within the group (shown as arrows or lines with a circle end). Modifiable connections between excitatory neurons of different groups or layers are shown as dotted lines. External input to each layer is shown by a solid filled arrow.

The model consists of two layers as shown in Fig. 1. The upper layer implements the "tag" mechanism. It contains many non-overlapping groups of neurons representing high level processing associated with the storage of ordered sequences of events (for example, a sequence of visual objects or a sequence of positions of an animal in a track, etc.). Each group is responsible for encoding a single sequence; therefore potentially the number of groups should be large enough to memorize many sequences. The neurons within a group are all-to-all coupled. In simulations the number of neurons in each group is 60. The neural groups of the upper layer project modifiable connections to some neural groups in the lower (bottom) layer.

The bottom layer implements the chain mechanism. It contains many oscillatory modules (i.e. small non-overlapping groups of interactive excitatory and inhibitory neurons). Each module is used for the oscillatory encoding of a single event. Different modules represent different events. In simulations each module in the lower layer contains 100 elements: 80 excitatory and 20 inhibitory neurons which are denoted in Fig. 1 as E and I, respectively. The neurons within a module are all-to-all coupled, generating rhythmic activity in the gamma range. Different modules in the layer interact through modifiable all-to-all connections between their excitatory neurons.

The dynamics of an individual neuron are described by the Hodgkin–Huxley equations (Hodgkin and Huxley, 1952),

$$\frac{dV_i}{dt} = -I_{\mathrm{ion},i} + I_{\mathrm{syn},i}^{\mathrm{lower}} + I_{\mathrm{syn},i}^{\mathrm{upper}} + I_{\mathrm{ext},i} + I_{\mathrm{rest}}, \tag{1}$$

$$\frac{dX_i}{dt} = A_X(V_i)(1 - X_i) - B_X(V_i)X_i, \quad X_i \in \{m_i, h_i, n_i\}, \quad i = 1, 2, \ldots, N, \tag{2}$$

where N is the number of neurons in the layer; $V_i(t)$ is the membrane potential of a neuron; X is a notation for any of the variables $m_i(t)$, $h_i(t)$, $n_i(t)$ (thus Eq. (2) is a concise notation for three equations), $m_i(t)$ is the activation variable of the sodium conductance channel; $h_i(t)$ is the inactivation variable of the

sodium conductance channel; $n_i(t)$ is the activation variable of the potassium conductance channel, $I^{\text{lower}}_{\text{syn},i}(t)$ is the synaptic current received by a neuron from other neurons in the lower layer; $I^{\text{upper}}_{\text{syn},i}(t)$ is the synaptic current received by a lower layer excitatory neuron from an upper layer neurons; $I_{\text{ext},i}(t)$ is the external current induced by the external input (equal to 40 mA); I_{rest} is a universal constant current that controls the activities of neurons (equal to −25 mA). Details about the total ionic current $I_{ion,i}(t)$ and the gating functions A_X and B_X can be found in Eqs. (3)–(9) of Appendix.

A positive external signal is received by a neuron in the bottom layer at each moment when a member of the stimulation sequence is presented. Otherwise the external signal is equal to zero. The external current is strong enough to transfer a neuron into the firing state. Without external current a neuron can fire if it receives both the synaptic current from the upper layer $I^{\text{upper}}_{\text{syn},i}(t)$ and synaptic currents from other modules.

Synaptic conductance is described using a standard alpha-function (see, e.g., Gerstner and Kistler, 2002). The total synaptic current of the ith neuron in the lower layer received from the neurons of the lower layer is described by the following equation:

$$I^{\text{lower}}_{\text{syn},i} = \sum_{j\in N_i^{inh}} I^{\text{inh}}_{i,j} + \sum_{j\in N_i^{\text{exc}}} I^{\text{exc}}_{i,j} + \sum_{j\in N^{\text{exc}}_{i,\text{external}}} I^{\text{exc,external}}_{i,j}, \quad i = 1, 2\ldots, N.$$

Here, the two first terms describe the sum of inhibitory and excitatory influences, N_i^{inh} (N_i^{exc}) is a set of indexes of incoming inhibitory (excitatory) connections from neurons of the same module; the third term describes the sum of external excitatory influences from neurons of other modules at the lower layer, $N^{\text{exc}}_{i,\text{external}}$ is a set of indexes of incoming excitatory connections from the neurons of other modules.

$$I^{\text{inh}}_{i,j} = w^A_{\text{inh}}(V_i - V^{\text{inh}}_{\text{syn}})\sum_{k=1}^{M^j}\alpha_j(t-T_k),$$

$$I^{\text{exc}}_{i,j} = w^A_{\text{exc}}(V_i - V^{\text{exc}}_{\text{syn}})\sum_{k=1}^{M^j}\alpha_j(t-T_k),$$

$$I^{\text{exc,external}}_{i,j} = w^B_{ij,\text{exc}}(t)(V_i - V^{\text{exc}}_{\text{syn}})\sum_{k=1}^{M^j}\alpha_j(t-T_k),$$

Here, $w^A_{\text{inh}} = w^A_{\text{exc}} = 0.1$ are constant connection strengths for inhibitory and excitatory connections inside the module; the alpha function is defined in the following way: $\alpha_j(t) = \text{at}\exp(-bt)$, for $t \geq 0$ and the alpha function equals to zero for $t < 0$; the parameters of the alpha function are: $a = 0.6$ m per second and $b = 0.03$ m per second; M^j is the total number of spikes from the jth neuron to the ith neuron; T_k is the time of the kth spike generated by the jth neuron; $V^{\text{inh}}_{\text{syn}}$ is the synaptic reversal potential of inhibitory coupling ($V^{\text{inh}}_{\text{syn}} = -80$ mV), $V^{\text{exc}}_{\text{syn}}$ is the synaptic reversal potential of excitatory coupling ($V^{\text{exc}}_{\text{syn}} = 0$ mV); $w^B_{ij,\text{exc}}(t)$ is a modifiable excitatory connection strength from the jth neuron to the ith neuron of different modules.

The total synaptic current of the ith neuron in the lower layer received from neurons of the upper layer is described by the following equation:

$$I^{\text{upper}}_{\text{syn},i} = \sum_{j\in N^{\text{exc}}_{i,\text{upper}}} w^C_{ij,\text{exc}}(t)(V_i - V^{\text{exc}}_{\text{syn}})\sum_{k=1}^{M^j}\alpha_j(t-T_k).$$

Here $w^C_{ij,\text{exc}}(t)$ is a modifiable excitatory connection strength from the jth neuron of the upper layer to the ith excitatory neuron of the lower layer; $N^{\text{exc}}_{i,\text{upper}}$ is a set of indexes of incoming excitatory connections from neurons of the upper layer to the ith excitatory neuron of the lower layer; the alpha function: $\alpha_j(t) = \text{at}\exp(-bt)$, for $t \geq 0$ and the alpha function equals to zero for $t < 0$; the parameters of the alpha function are: $a = 0.6$ m per second and $b = 0.03$ m per second; M^j is the total number of spikes from the jth neuron at the upper layer to the ith neuron; T_k is the time of the kth spike generated by the jth neuron at the upper layer; $V^{\text{exc}}_{\text{syn}}$ is the synaptic reversal potential of excitatory coupling ($V^{\text{exc}}_{\text{syn}} = 0$ mV).

2.2. Memory formation

At the initial state all modifiable connections have zero connection strengths and all oscillatory modules of the lower layer work independently. Internal connection strengths of the modules are selected in such a way (see their values above) that each module demonstrates fast oscillations in the gamma range.

The storage of a sequence is paced by the theta rhythm. The storage of one event requires a time window of duration 200 ms. Thus, the storage of a sequence of 5 events requires five sequential time windows with a total duration of 1 s.

Each memory (which we call an event or object) is coded by a prescribed set of oscillatory modules which receive an external current ($I_{\text{ext},i} = 40$) during some particular respective time window. This current excites neurons of the coding modules and these neurons demonstrate fast rhythmic activity during this time window. Thus, during the first time window all modules coding the first event are active, during the second time window all modules coding the second event are active, etc. In fact, in this paper we consider a simple coding scheme where one event is coded by one module.

The learning rule is inspired by Spike-Timing-Dependent Plasticity rule (STDP rule; see, e.g., Markram et al., 1997). This rule is applied to connections between all pairs of excitatory elements where the first element is from module P and another element is from module Q. Our approach is inspired by a temporally asymmetric modification of the STDP learning rule. We take into account the activity of presynaptic neurons of module P and post-synaptic neurons of module Q in two subsequent time windows. The connection strength $w^B_{ij,\text{exc}}(TW_i)$ from jth neuron of module P to ith of module Q is modified at the time corresponding to the current time window TW_i if and only if an excitatory neuron j generates spikes within the previous time window TW_{i-1} and an excitatory neuron i generates spikes within the current time window TW_i. The modified value of $w^B_{ik}(t)$ increases in a stepwise manner by the value:

$$\Delta w^B_{ij,\text{exc}}(TW_i) = \begin{cases} 3, & \text{memorization,} \\ 0, & \text{otherwise.} \end{cases}$$

There is no forgetting, so the modified value is kept indefinitely long.

In parallel with the modification of connection strengths between neurons of active modules, the connection strengths from excitatory neurons of the group which "tags" the selected sequence to the neurons of active modules is also modified according to the same learning rule. A group of neurons in the upper layer, which is selected to tag the sequence which is currently being stored, receives an external current ($I_{\text{ext},i} = 40$) during the total time of sequence storage. Due to this external current, the neurons of the "tag" group demonstrate constant oscillatory activity during all time windows corresponding to the stored sequence. Therefore, connection $w^C_{ij,\text{exc}}(WT_i)$ from an excitatory neuron j in the tag group of the upper layer to an excitatory neuron i of the active module in

the time window TW_i is modified. The modified value of $w^C_{ij,\mathrm{exc}}(t)$ increases in a stepwise manner by the value:

$$\Delta w^C_{ij,\mathrm{exc}}(TW_i) = \begin{cases} 3, & \text{memorization,} \\ 0, & \text{otherwise.} \end{cases}$$

After a sequence is stored, a neural group in the upper layer, which tags the sequence, will provide additional excitation to all the participating neural modules in the lower layer. In this way the sequence is "tagged" or "highlighted" by selected neurons of the upper layer.

Besides the STDP type leaning rule with the modification of the connection from a previously active neuron to the currently active one, we use an anti-STDP type leaning rule with the modification of the connection from a currently active neuron to a previously active one. The STDP type learning rule is designed to demonstrate the forward replay of a sequence while the anti-STDP type learning rule provides a possibility of backward recall. In the case of anti-STDP type leaning rule, the sequence is learnt in the increased time order (first event, second event, etc.) but during the recall process this sequence is replayed in the reverse order (from the last to the first event).

An anti-STDP learning rule has been suggested in some papers (e.g. Han et al., 2000; Rumsey and Abbott, 2004). In simulations below we demonstrate how the system works for each of these two types of leaning rule. All neurons in one neural module have the same type of connection direction and they only connect to modules of the same type (i.e. an STDP neural module only connects to other STDP modules and an anti-STDP neural module only connects to other anti-STDP modules).

2.3. The recall

Let us assume that several sequences are stored in the memory. For a recall we should decide which sequence is to be recalled. We do not model the decision making process. In the brain this decision making process and selection of a sequence for a recall is based on a current context and associations. We assume that the information at the input of the model contains: (1) A pointer to the "tag" group in the upper layer associated with the selected sequence. (2) One event from the selected sequence that activates the proper module in the bottom layer to start the recall process (it might be the first event or any other event from the selected sequence). The recall starts from this event and runs through all subsequent events (ignoring all previous events).

Thus, at the beginning of the recall procedure the neural module representing a starting event of the selected sequence is briefly stimulated by the external current ($I_{\mathrm{ext},i} = 40$) which is applied to all the elements of the module for 50 ms. In parallel another external current of the same value is applied to all the elements of the "tag" group in the upper layer, corresponding to the selected sequence. This current is applied for the period of 200 ms (the time of the recall process). In fact, to start the recall it is enough to activate only a subset of neurons in the modules corresponding to the starting event.

During the recall, the module corresponding to the starting event becomes active and its activity (through the connections which have been modified by the memorization process) propagates to the next module corresponding to the second event in the sequence. Excitation from a previous module and activation from the upper layer simultaneously arrive and stimulate the second module. Note that the second module becomes active with a small time delay after the excitation of the first module. The second module excites the following one, etc. The total time of the recall is short and the complete period of the sequence replay is 200 ms. During the recall, the times of activity of different modules overlap, but starting moments of activation appear in the proper order of events in the recalled sequence.

3. Model simulations

To demonstrate the performance of the memory model, we show the results of three simulations. These examples provide a basic idea of model functioning. The results of other simulations as well as detailed description of memory capacity will be given in a separate publication.

3.1. Storage and recall of one sequence using STDP type learning rule

Consider a set of ten modules (M1–M10 containing 1000 neurons) in the lower layer and one group of neurons in the upper layer. A sequence of four events is coded in the following way: Event 1 is coded by module 2 (M2, neurons from 101 to 200). Event 2 is coded by module 8 (M8, neurons from 701 to 800). Event 3 is coded

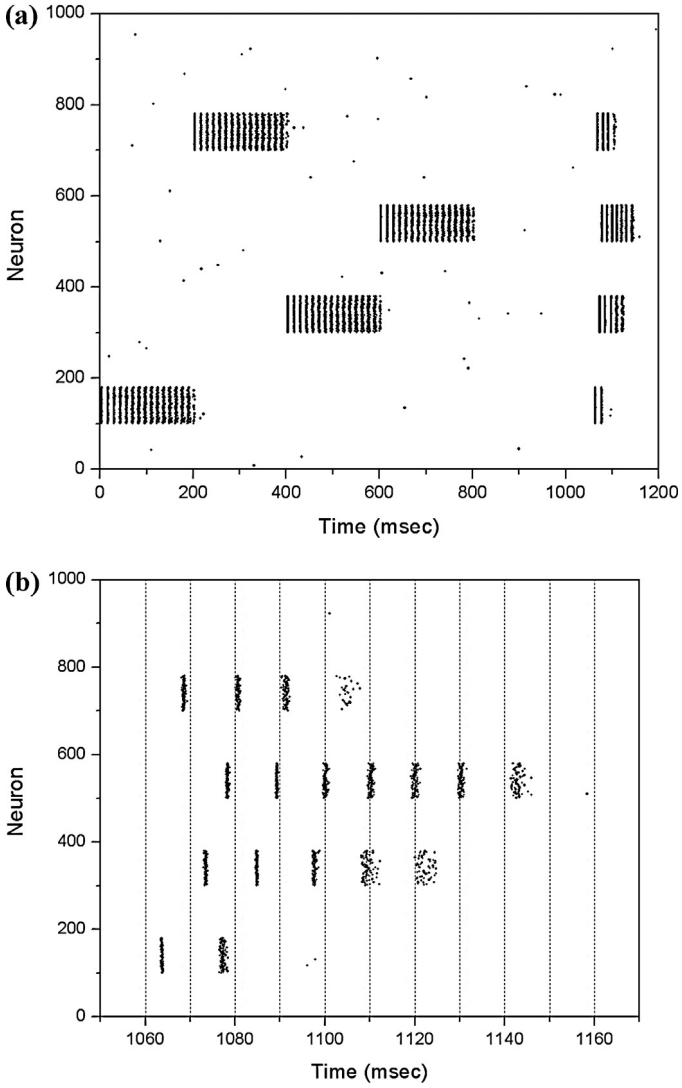


Fig. 2. Storage and recall of a sequence composed of 4 events. Each small dot in the graph represents a spike of an excitatory neuron. (a) Storage of a sequence of four events. The left part shows the activation of neurons during the encoding period. The right part (after 1000 ms) shows the direct recall which is initiated by the activity pattern corresponding to the first event. (b) A magnified picture of the recall period. Neural modules fire with a short delay one after another.

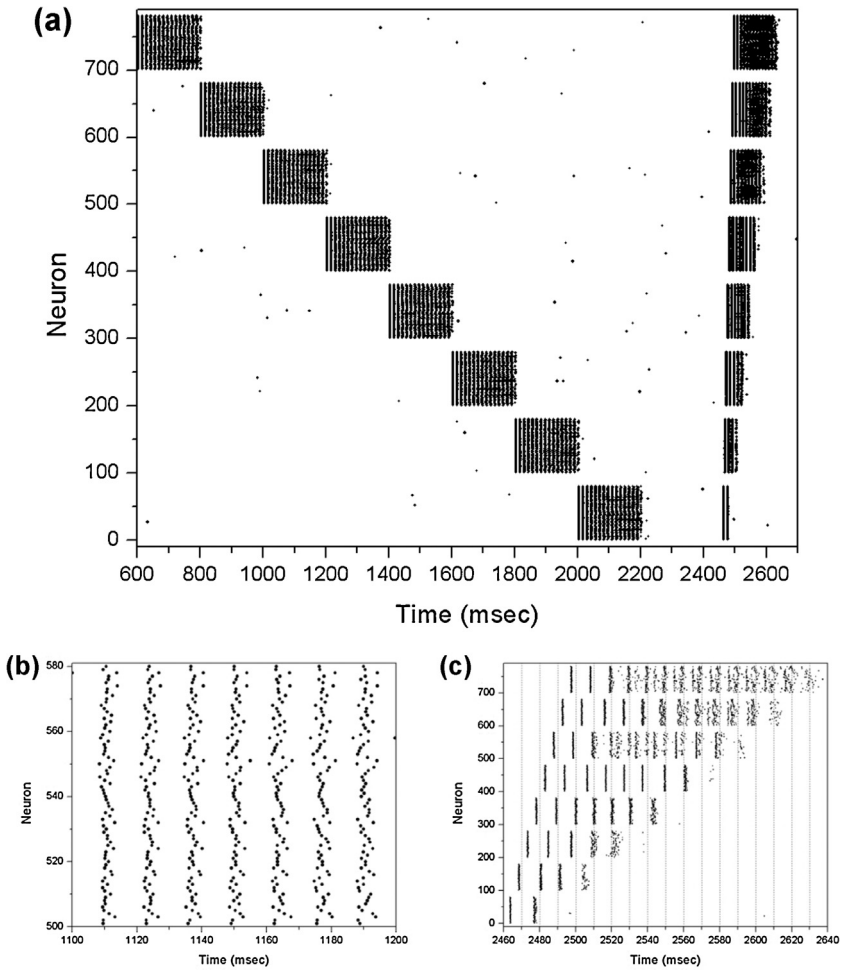


Fig. 3. Each small dot in the graph represents a spike. (a) Storage and recall of eight positions of a rat along the track, represented by eight modules of neurons. The left part shows the activation of neurons (the encoding period). The right part shows the reverse recall which is initiated by the activity pattern corresponding to the end of the track (the recall period). (b) A magnified picture of the encoding period shows noisy but coherent firing of neurons within a module of the bottom layer. (c) A magnified picture in the recall period shows the firing of neural modules one after another (with a short delay between the modules).

by module 4 (M4, neurons from 301 to 400). Event 4 is coded by module 6 (M6, neurons from 501 to 600).

Fig. 2a demonstrates the storage and recall of a sequence of events. According to the encoding scheme, the Events 1–4 are represented by four neural modules in the lower layer. We show the activity of excitatory neurons only, i.e. the activity of neurons 1–80 in each module. The upper layer contains one group of 60 neurons. Their activities are simply periodic (not shown).

During the encoding period of 800 ms (left side of Fig. 2a), the external currents are injected into the corresponding modules in the order of events in the sequence (M2, M8, M3, M6). Each module receives injection for a period of 200 ms: M2: 0–200 ms; M8: 200–400 ms; M3: 400–600 ms; M6: 600–800 ms. The spiking activity of neurons of these modules is shown in Fig. 2a. There is no stimulation in the time interval 800–1000 ms, no oscillatory spiking in this time interval and therefore no modification of synaptic strengths according to the learning rule. It means that the process of sequence encoding has stopped. We assume that the learning rule is of STDP type and forward recall is expected.

The recall procedure is shown in the time interval 1–1.2 s. A brief external current ($I_{ext,i} = 40$ mA) is injected to the neurons of M2 for the period 0–20 ms. In parallel, the same current is injected to all the neurons of the upper layer. The recall procedure is shown in the right hand side of Fig. 2a (a zoom of this part of the figure is shown in Fig. 2b). During the recall, starting times of module activation are arranged in the correct order corresponding to the encoding procedure. Subsequent neural modules are activated one after another by the synaptic currents from the previous module and the additional current from the upper layer. There is an interesting "avalanche effect": the duration of reactivation of subsequent neural modules increases along the sequence.

3.2. Storage and recall of one sequence using anti-STDP type learning rule

This simulation is inspired by exciting recordings from the rat hippocampus that demonstrate the phenomenon of the reverse replay (Colgin and Moser, 2006). During a run along a linear track the hippocampal place cells fire spikes in the theta frequency band (about 5 Hz) sequentially in the order of positions in the track (starting from the beginning of the track). When the rat is rewarded at the end of the track, the hippocampus enters a sharp-wave mode with the firing sequence replaying in the reverse order, from the end to the beginning of the track. This reverse replay is fast and the whole sequence of spikes appears in the time interval of about 200 ms (one period of the theta rhythm).

For this simulation we use 8 modules (M1-M8, 800 neurons, the activity of 80 excitatory neurons is shown for each module). The events from one to eight are coded respectively by the

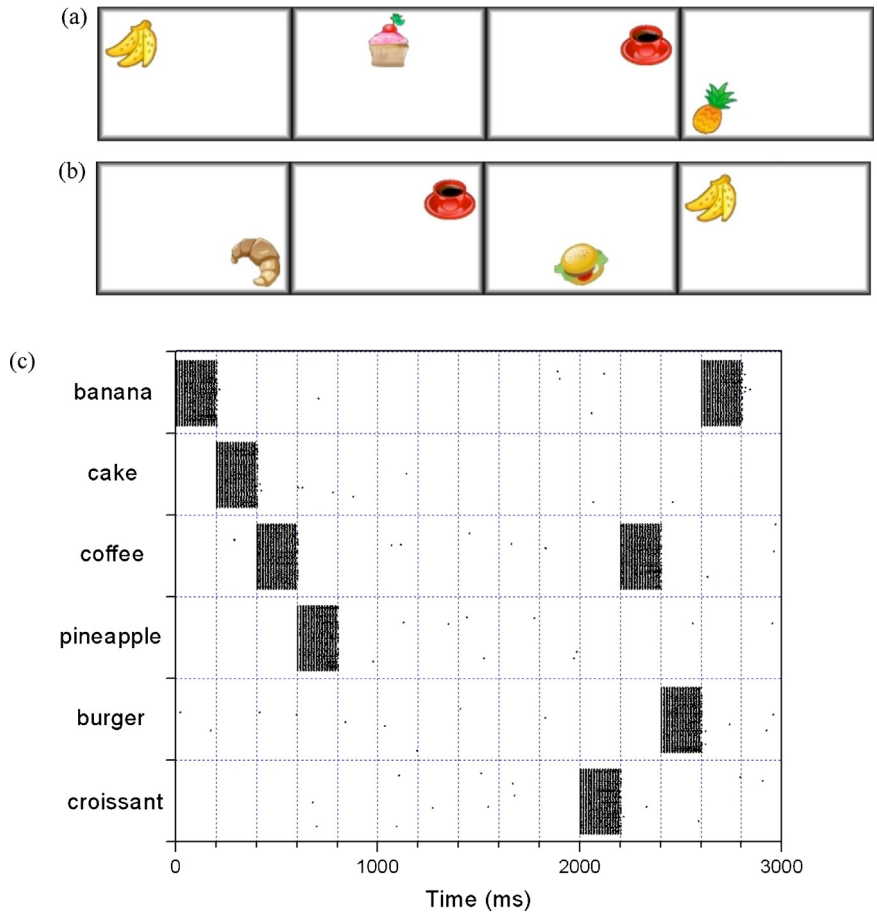


Fig. 4. Storage of two sequences (a and b) with four objects in each sequence. Note that the third object of a and the fourth object of b are the same. Also the first object of a and the last object of b are the same. In the simulation corresponding neural modules are activated one after another, as shown in the rastergram (c). The first sequence was stimulated during the time 0–800 ms (200 ms for each object). The second sequence was stimulated from 2000 ms to 2800 ms.

following modules: M8 is stimulated in the interval 600–800 ms, M7: 800–1000 ms,. . ., M2: 1.8–2 s, M1: 2–2.2 s. The activity of the excitatory neurons of these modules is shown in the left hand side of Fig. 3a during the respective time intervals. Fig. 3b shows a zoom of spiking activity of excitatory neurons of M6. The frequency of spiking is around 70 Hz, though individual neurons may skip one or two cycles. The firing pattern is noisy (see equations in Appendix) but coherent. This is in agreement with the experimental finding on the role of oscillations and synchronization in memory (Duzel et al., 2010). There are no current injections in the interval 2.2–2.4 s which means that encoding is stopped at that period of time. The recall procedure develops in the interval 2.4–2.6 s.

The anti-STDP type learning rule is used in this simulation to modify connection strengths of periodically spiking neurons in the lower layer. As in the previous simulation example, the upper layer contains one group of 60 neurons. The modifiable connections from the neurons of this group to the lower layer are adjusted according to the STDP type learning rule.

The replay procedure in the time interval 2.4–2.6 s is shown in the right hand side of Fig. 3a (a zoom of this part of Fig. 3a is shown in Fig. 3c). The neurons of the module M1 are briefly activated and the current injection is applied to the neurons of the upper layer. After that, subsequent modules are activated one after another by the synaptic current from other modules and additional influence from the upper layer. The replay is in the reverse order. The starting times of module activation appear with a short time shift of 5 ms in the reverse order relative to the ordering of events during storage. The duration of the replay is short and takes one cycle of the theta rhythm, in accordance with experimental results (Diba and Buzsaki, 2007; Foster and Wilson, 2006).

3.3. Storage and recall of two sequences using STDP type learning rule

In this simulation we deal with two sequences of visual objects. There are six objects which are represented by six modules: Croissant: M1 (neurons 1–100); Burger: M2 (neurons 101–200); Pineapple: M3 (neurons 201–300); Coffee cup: M4 (neurons 301–400); Cake: M5 (neurons 401–500); and Banana: M6 (neurons 501–600). These objects are arranged to two sequences of 4 objects in each sequence (see Figs. 4a and b). Also, there are two groups of neurons in the upper layer which represent the tags “1” and “2”, respectively.

Note that the sequences contain overlapping objects. For example, the object “Coffee cup” is the third object of the first sequence and the second object of the second sequence. Obviously, if only chain connections were used for the memory model then it would be impossible to recall any of these two sequences without an error because after encoding the module M4 (“Coffee cup”) would be connected to two different modules, M2 (“Burger”) and M3 (“Pineapple”). The presence of the upper layer in the model helps to resolve this ambiguity. The input from the upper layer which relates to the sequence “1” will guide the recall process from the object “Coffee cup” to the object “Pineapple” by adding additional excitation to the objects of the sequence “1”. Similarly, the

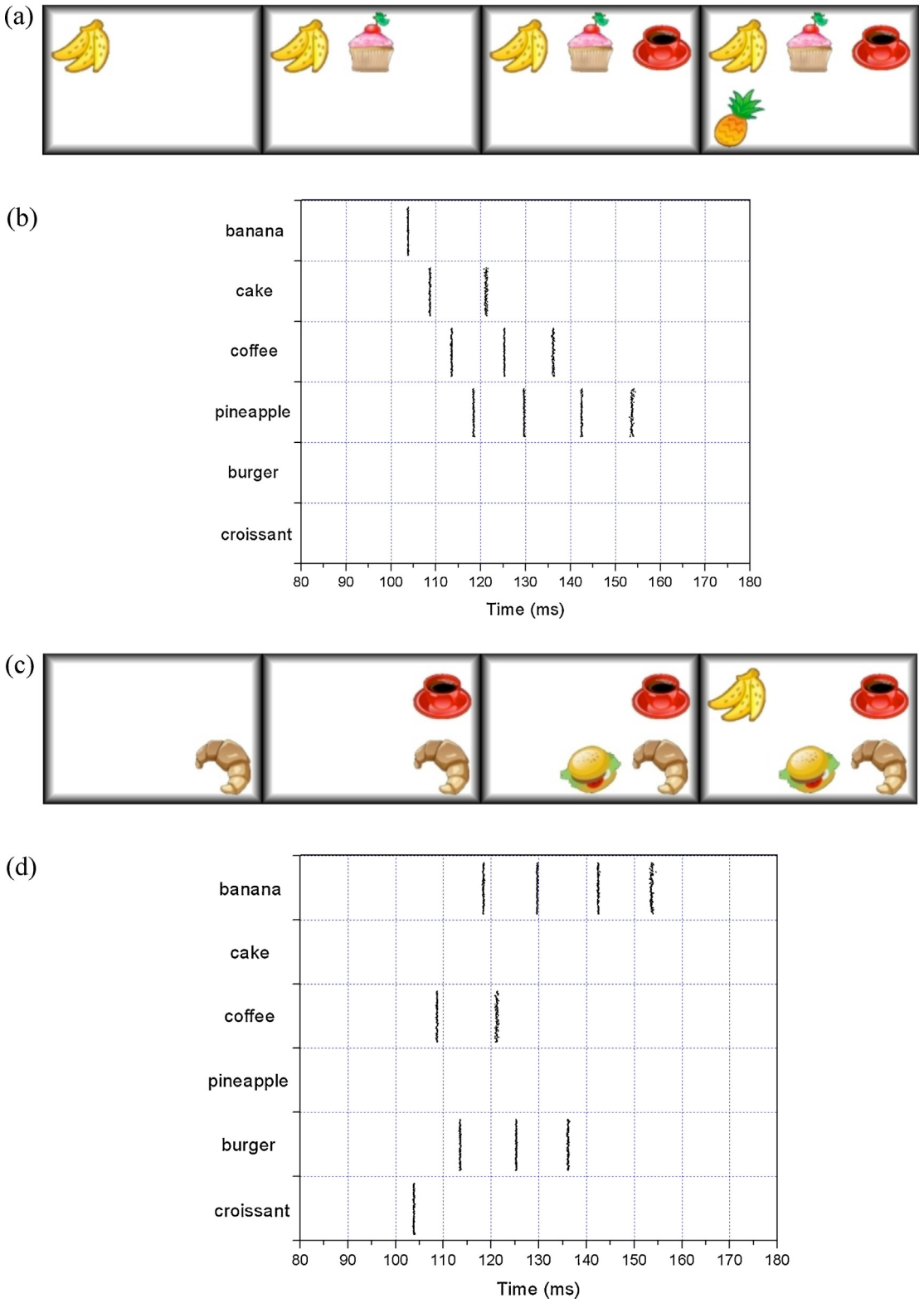


Fig. 5. Recall in an ambiguous situation when two sequences (a and b) contain the same elements. The system is able to correctly recall all objects starting from the first object. (a) and (c) Snapshots of the recall process. (b) and (d) Raster plot of spikes of excitatory neurons in the lower layer during the recall process.

input from the upper layer which relates to the sequence "2" will guide the recall process from the object "Coffee cup" to the object "Burger".

Fig. 4c shows the storage process. The first sequence is stored in the time interval 0–800 ms which includes four time windows corresponding to the following encoding modules: M6, M5, M4, and M3. The second sequence is stored in the time interval 2–2.8 s (also four time windows that correspond to encoding in the modules M1, M4, M2, and M6). During sequence storage, STDP type connections are formed between the modules.

For a recall we briefly activate the module representing the first element of the sequence as well as the tag group of neurons in

the upper layer. Fig. 5 shows the recall procedure in the case of initiation of the sequence "1" or sequence "2".

Figs. 5a and c shows the order in which objects are recalled for the sequences "1" and "2", respectively. Figs. 5b and d shows the corresponding spike raster plot of excitatory neurons of the lower layer. The results of the simulation confirm that there is no confusion between the two sequences even though they have overlapping elements. This is achieved due to the influence of the upper layer which delivers additional current to the lower layer. Since the resting states of neurons are set far from the firing threshold, the activation of a neural module in the lower layer requires both the synaptic currents from the neurons of the previous module in the chain and additional synaptic current from the upper layer. The neurons belonging to another sequence are labeled by another neural group of the upper layer. As a result, only one neural module in the lower layer will be activated at a time, which means that the sequence is recalled unambiguously.

4. Discussion

The model presented in this paper is based on the assumption that there is a general mechanism of storage and recall of sequences of items that is universal for various types of items such as events, spatial positions, or visual objects. The representation of external information is specific at the primary stages of information processing but it becomes abstract at higher stages when the information is prepared for memorization.

The two-layer architecture of the model is intended for reflecting the interaction between the associative regions (the associative cortex or the hippocampus) and the prefrontal cortex. The latter is used to control storage and recall, taking into account the external and internal context. This context provides a modulation of neural activity in particular neural assemblies of the prefrontal cortex which can be helpful for resolving the ambiguity that appears when different sequences have identical items. Thus according to the model the traces of memory are formed both inside associative regions and in connections between the associative regions and the prefrontal cortex. Connections in associative regions are modified for "chaining" representative neural groups. The connections between associative regions and the prefrontal cortex are modified to label all groups representing a particular sequence by the context.

The model reproduces experimental results of the forward and backward replay as it is observed in the hippocampal place cells of rats. Recent biologically plausible models of sequence storage are based on the mechanism of phase coding in the theta frequency band (Yamaguchi, 2003; Koene and Hasselmo, 2008). Unfortunately, the frequency of the theta rhythm in these models varies widely; therefore it is not clear whether phase coding can be reliable under this condition. We provided an alternative approach which is more robust and gives stable results of storage and recall. The results of computer simulations follow the experimental evidence that sequence storage demands a relatively long time (several theta cycles) while the recall can go in a short period of a single theta cycle.

Computer simulations confirm that the system is able to resolve some ambiguities when two sequences contain identical items. Still, the capability of the model to avoid ambiguities is not absolute. It is easy to construct an example when an erroneous recall becomes possible. Consider two sequences $S_1 = (\ldots A, B, \ldots, X \ldots)$, $S_2 = (\ldots A, X, \ldots)$. Suppose that both S_1 and S_2 are stored in memory. Then an attempt to recall S_1 will fail since the activity in the module M_A will simultaneously induce the activity in both modules M_B and M_X. It should be noted that this example represents a difficult case when even people in real life can make a mistake. Still the model has the potential to at least decrease the number of such errors.

The main drawback of the model is that it represents each event (object) by a single module in the lower layer. More efficient construction would be to have different modules of neurons in the lower layer for presentations of the same object under different contexts. This would eliminate this kind of errors but will dramatically increase memory consumption. Therefore in this case the errors may appear due to exhaustion of memory capacity.

Another source of possible errors is the ambiguities that arise from repeated appearances in a sequence of the same events. For example, the model has no information to disambiguate the recall of the sequence $S = (\ldots A, B \ldots A, C \ldots)$. This is a typical problem in sequence storage. It is usually solved by taking into account several consecutive members of the sequence or equivalently to form "chunks" from short sequences of items which become the items of the "higher order". This also radically increases memory consumption. However, we expect that complicated strategies of sequence storage that combine all these mechanisms can be used in reality.

The capacity of sequential memory C is limited by the number of neural groups M in the upper layer because the model requires one neural group in the upper layer to represent a 'tag' of the sequence. Thus, $C \leq M$. The capacity also depends on the length n of a single sequence which should be stored in the memory and the number of modules N in the bottom layer. If all members in the sequences are different, the evident estimation for C is $C = [N/n]$, where $[x]$ is the integer part of x. In the case of sequences with overlapping members the estimation of C depends on the type of overlapping. For example, if there is only one common member in all stored sequences and all other members in the sequences are different the estimation for C is $C = [(N-1)/(n-1)]$. There are some other simple examples when the estimation of C can be obtained theoretically, but a special study is needed to get a general view. We think that purely statistical approach to this problem that is usually applied in the analysis of associative memory is not very useful in the case of sequence storage. In reality, the sequences that should be stored are far from being random. Therefore the estimation of the capacity should be made taking into account the structure of sequences that appear in real experience of animals and people. This will be a subject of our further investigation.

Acknowledgment

We are grateful to Robert Merrison for careful reading of the manuscript and useful comments.

Appendix A.

We use Hodgkin–Huxley model (1)–(2) for each neuron. The functions A_X and B_X $X \in \{m, h, n\}$ in the right part of Eq. (2) are described by the following expressions:

$$A_m(V) = \frac{2.5 - 0.1(V - V_{rest})}{\exp(2.5 - 0.1(V - V_{rest})) - 1}, \quad (3)$$

$$A_h(V) = 0.07 \exp\left(\frac{-(V - V_{\text{rest}})}{20}\right), \quad (4)$$

$$A_n(V) = \frac{0.1 - 0.01(V - V_{\text{rest}})}{\exp(1 - 0.1(V - V_{\text{rest}})) - 1}, \quad (5)$$

$$B_m(V) = 4 \exp\left(\frac{-(V - V_{rest})}{18}\right), \quad (6)$$

$$B_h(V) = \frac{1}{\exp(3 - 0.1(V - V_{\text{rest}})) + 1}, \quad (7)$$

Aurora Face Recognition Technical Report: Evaluation of Algorithm "Aurora-c-2014-1" on Labeled Faces in the Wild

Dr T. Heseltine Dr P. Szeptycki Mr J. Gomes Dr M.C. Ruiz Dr P. Li

t.heseltine@auroracs.co.uk
Aurora Computer Services Ltd.
www.facerec.com

Abstract

We evaluate the performance of a new method of face recognition on the well-known benchmark Labeled Faces in the Wild (LFW) dataset [1]. The method, developed by the Core Technology Research Team at Aurora, achieves a mean classification accuracy of 93.24% on the unrestricted view 2 test set, outperforming all other results on the LFW website.

Brief Method Description

The face recognition technology is comprised of Aurora's proprietary algorithms, machine learning and computer vision techniques. We report results using the unrestricted training protocol, applied to the view 2 ten-fold cross validation test, using images provided by the LFW website, including the aligned and funnelled [6] sets and external data used solely for alignment purposes.

Company Background

Aurora has 15 years' experience in the field of biometric face recognition. For the last few years our team has been working on the challenging problems of uncontrolled visible-spectrum colour images, as detailed in this report. We have previously created infrared face recognition systems, utilising our bespoke hardware, to overcome typical lighting problems. Our systems are widely deployed and we have more installed systems in the UK than any other competitor.

Recognition Pipeline

The complete face recognition pipeline of a full system consists of four primary steps: face detection; feature point localisation; feature extraction and classification. The verification stage (feature extraction and classification) constitutes proprietary descriptor extraction procedures and comparison metrics to produce a similarity score, which in turn is applied to a threshold for the final classification decision.

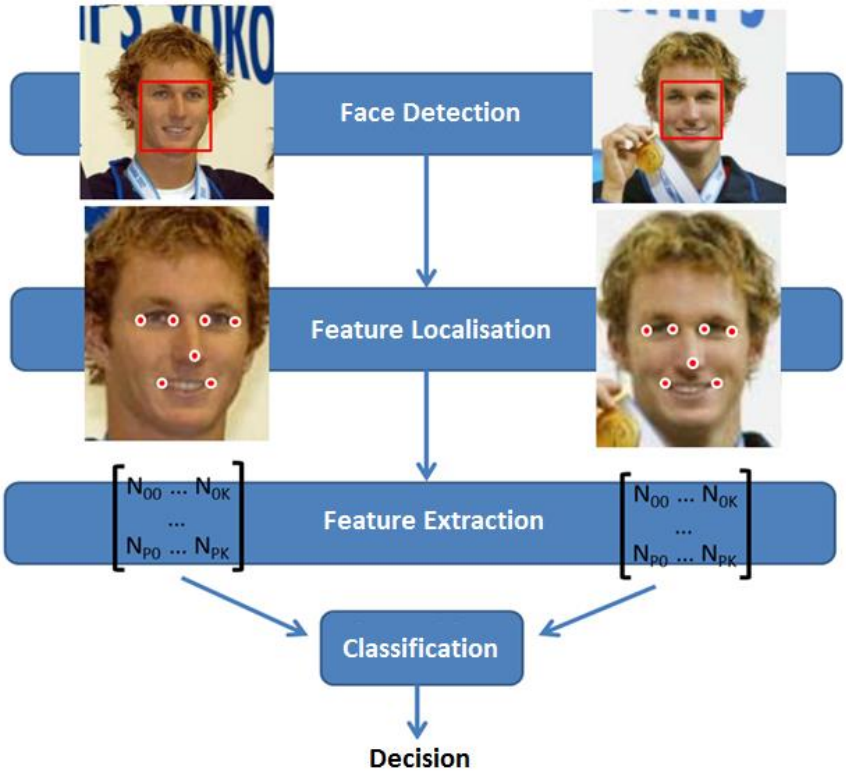


Figure 1. Aurora's recognition pipeline.

Although Aurora has developed a number of highly accurate algorithms for face detection and feature localisation (not described here), the purpose of this investigation is to evaluate the verification stage (feature extraction and classification) of the pipeline as an individual component. Therefore, we assume face detection and feature localisation have been completed successfully, relying on the pre-aligned face images.

The purpose of this investigation is to determine the maximum achievable performance, given accurately located feature points. Other experiments by ourselves investigate the impact of automatically detected feature points.

Evaluation Methodology

The evaluation is carried out according to the 10-fold cross validation test under the unrestricted configuration, strictly following the training and test procedure that was defined in the technical report of LFW database [1].

Ten 'folds' of the view 2 data set are processed. For each fold, 600 image pairs are compared to produce a similarity score, to which a threshold is applied to make the final classification decision. Images from outside of the test set pairs are used to train the face recognition model, comparison

metric and classification threshold. Although we do not restrict training to only those images pairs specified, we do ensure that no training, optimisation or fine tuning is carried out on any of the images present in the 600 image pairs of the test set. Neither are any subjects in the 600 image pairs present in the data used for training, optimisation or fine tuning; hence each fold of the test is conducted blind, as required by the protocol. The training and evaluation process is repeated for each of the ten folds, from which the mean classification accuracy is computed.

Results

Using the LFW unrestricted protocol, the Aurora face recognition engine achieves a mean classification accuracy of 0.9324 ± 0.0044, outperforming all other results published on the LFW website.

Organisation	Algorithm	$\hat{u} \pm S_E$
Aurora	**Aurora-c-2014-1**	**0.9324 ± 0.0044**
UST China, MS Research Asia	High-dim LBP [2]	0.9318 ± 0.0107
Oxford University	Fisher vector faces [3]	0.9303 ± 0.0105
Vision Labs	VisionLabs ver.1.0,aligned	0.9290 ± 0.0031
NEC	CMD+SLBP, aligned [5]	0.9258 ± 0.0136
Face.com	Face.com r2011b [4]	0.9130 ± 0.0030

Table 1. Mean classification accuracy ($\hat{u}$) and standard error (S_E) of the five top performing submissions reported on the LFW website, compared with the Aurora algorithm.

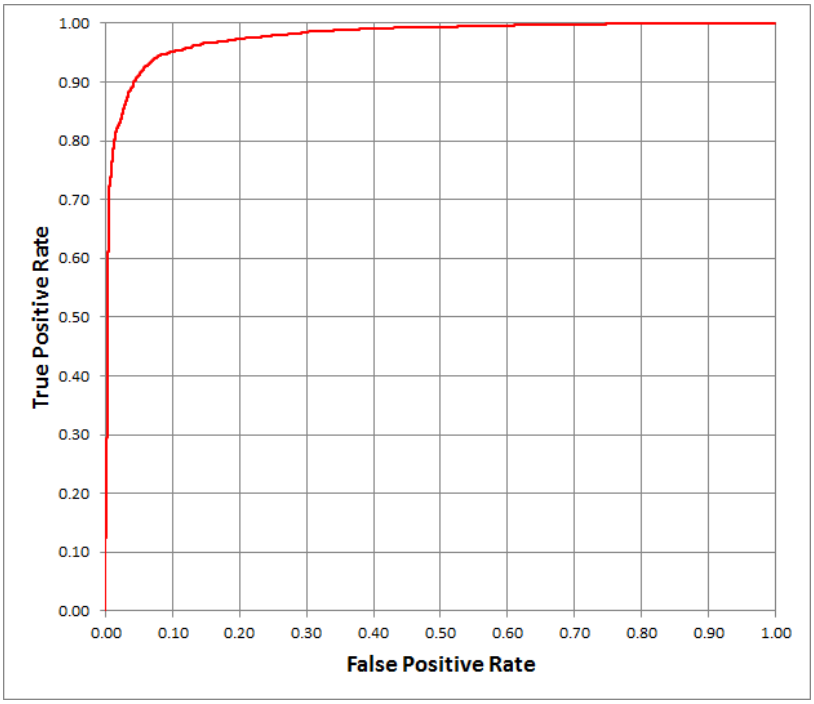


Figure 2. ROC curve of the "Aurora-c-2014-1" algorithm for the full ten-fold cross validation.

Results and Conclusion

The results produced by Aurora are state of the art and able to match faces in surprisingly difficult conditions. Example results from the LFW dataset are shown below, all of which were **correctly** classified by the Aurora system. The matches include extreme appearance changes such as different hair styles, hats and glasses, make-up and ageing, as well as more common difficulties such as pose, expression, poor lighting and partial occlusions.

Figure 3. The algorithm is able to cope with changes in head angle, even matching from a partial left profile to partial right profile.

Figure 4. The algorithm has been developed to account for natural variations in expression, as evident in the above examples.

Figure 5. The above examples demonstrate how the system is able to cope with changes in appearance, such as hair styles (top left), beards (top right), glasses (bottom left) and aging (bottom right).

Figure 6. Examples of correct classification, given some partial occlusion of the facial region.

Figure 7. Examples of extremely difficult cases, correctly classified by the Aurora face recognition algorithm under the LFW protocol, although presenting significant challenges for face and feature localisation algorithms in a complete system.

References

1. Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

2. Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification. Computer Vision and Pattern Recognition (CVPR), 2013.

3. Karen Simonyan, Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Fisher Vector Faces in the Wild. British Machine Vision Conference (BMVC), 2013.

4. Yaniv Taigman and Lior Wolf. Leveraging Billions of Faces to Overcome Performance Barriers in Unconstrained Face Recognition. ArXiv e-prints, 2011.

5. Chang Huang, Shenghuo Zhu, and Kai Yu. Large Scale Strongly Supervised Ensemble Metric Learning, with Applications to Face Verification and Retrieval. NEC Technical Report TR115, 2011.

6. Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. International Conference on Computer Vision (ICCV), 2007.