Masterarbeit

Slow Feature Analysis On Visual Data

Rene Sendlhofer

Institut für Grundlagen der Informatik
Technische Universität Graz
Vorstand: O. Univ.-Prof. Dipl.-Ing. Dr. rer. nat. Wolfgang Maass

Betreuer und Begutachter: Ass.Prof. Dipl.-Ing. Dr.techn. Robert Legenstein

Graz, im Dezember 2009

Kurzfassung

Zahlreiche neurophysiologische Experimente zeigen, dass Objekte von Neuronen in höheren visuellen Arealen in einer translationsinvarianten Weise kodiert werden. Neueste Experimente von Li und DiCarlo (2008) deuten darauf hin, dass diese invarianten Darstellungen durch das sogenannte Prinzip der Langsamkeit erlernt werden können.

Das Lernen nach dem Prinzip der Langsamkeit beruht auf der Tatsache, dass sich wichtige Umgebungsvariablen wie etwa Objektidentitäten oder deren Positionen auf einer langsameren Zeitskala ändern als die sensorisch aufgenommenen Signale. Schafft man es, diese sich langsam ändernden Variablen aus den Sensordaten herauszufiltern, kann man damit invariante Kodierungen, etwa von visuellen Inputs bezüglich Transformationen wie Rotation, Translation oder Skalierung, erreichen.

Ein effizienter Algorithmus, basierend auf dem Prinzip der Langsamkeit, ist Slow Feature Analysis (SFA) von Wiskott und Sejnowski, welcher langsam ändernde Eigenschaften eines Signals extrahiert.

Diese Masterarbeit zeigt, dass es mit Hilfe von SFA und einem detaillierten Modell des Primren Visuellen Kortex (V1) möglich ist, translationsinvariante Objekterkennung durchzuführen.

Simulierte visuelle Inputs dienen als Input für ein Modell der Retina und des Lateral Geniculate Nucleus (LGN). Der Output dieses Modells wird vom V1-Modell weiterverarbeitet. Die simulierten visuellen Inputs sind Videos von zwei verschiedenen Objekten. Dabei werden Sakkaden, das sind schnelle Augenbewegungen, simuliert. Sakkaden finden zwischen drei bis fünf mal pro Sekunde statt und tasten das gesamte visuelle Feld ab. Die Neuronenaktivität des V1 Modells wird dem SFA Algorithmus zugeführt, welcher durch unüberwachtes Lernen translationsinvariante Objekterkennung durchführt.

Die Ergebnisse zeigen, dass der SFA Algorithmus, angewandt auf die Neuronenaktivität eines V1-Modells, in der Lage ist, translationsinvariante Objektrepräsentationen zu lernen.

Abstract

There exists abundant evidence from psychophysiological and neurophysiological experiments that visual objects are represented in higher visual areas in a translation invariant manner. Recent experimental data suggests that such invariant representations are learned based on the slowness principle. This principle is based on the observations that environmental signals and raw sensory signals from the retina vary on different time-scales. Causes for these variations usually vary on a slow time-scale, for instance the identity or position of an object that changes over time. By extracting these slowly varying features from a time varying signal, the properties of the surrounding environment can be reflected. An efficient algorithm that is based on the slowness principle is Slow Feature Analysis (SFA) developed by Wiskott and Sejnowski in 2002. This thesis investigates whether SFA is able to perform translation invariant object recognition on the output of a detailed simulated patch of V1 neurons. The visual input is composed of two different objects and sequences are generated in a biologically plausible way by using saccades and fixation periods. These saccadic eye movements occur fast with three to five movements per second and scan the visual field. The sequences serve as an input to a model of the retina and the lateral geniculate nucleus (LGN) before being processed by the V1 circuit. The SFA algorithm is then applied to the output of the circuit in an unsupervised manner for translation invariant classification of visual objects. The results clearly show that SFA is able to distinguish between the objects in a translation-invariant way with very high performance.

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....
date (signature)

Acknowledgement

It is a pleasure to thank those people who made this thesis possible. First of all my supervisor Robert Legenstein. He supported me with a lot of important facts and organisational advices during the whole process of writing. It is also an honour for me to thank Prof. Wolfgang Maass, the head of the department for Theoretical Computer Science, for his support and ideas on improving my simulations.

I am furthermore indebted to my colleagues at the department, Stefan Klampfl and Klaus Schuch, who helped me getting started with all the different development tools. They also continually brought new input for testing my work and improving the results.

I also owe my deepest gratitude to my parents, who actually facilitated the opportunity to access the education I wanted. Without their supporting background I would have never been able to graduate.

This appreciation also goes to my girlfriend Maria, who tolerated my behaviour during busy periods, when there was no time for anything else than focussing on my thesis.

Graz, 18th december 2009 Rene Sendlhofer

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Everyday we make decisions based on things we see and need to classify. We do not even think about the process of discriminating different objects and it seems as it is the most simple thing on earth. Even for dramatic changes in object scale, rotation or position, we manage to recognize everything perfectly, and if things are partially hidden by other objects we are still able to build up a representation of the whole object to recognize it. Abundant evidence exist from psychophysiological and neurophysiological experiments (Biederman and Cooper, 1992, Oram and Perrett, 1994, Tovee et al., 1994, Ito et al., 1995, Furmanski and Engel, 2000, Li and DiCarlo, 2008) that visual objects are represented in higher visual areas in a translation invariant manner, but how are these invariances build in our visual system? How is it possible to form these clear representations despite the existence of large clutter? In a more general sense one can say that object recognition in the visual system is not a single task but rather a composition of multiple processing stages (Logothetis and Steinberg, 1996). Many approaches exist that try to model the way visual information is processed by the visual pathway (Fukushima, 1980, Riesenhuber and Poggio, 1999, Franzius et al., 2008). They use the concept of simple and complex cells by Hubel and Wiesel (1962) or use temporal slowness as the main principle to uncover invariant representations.

The surrounding real-world environment and internal representations of this environment vary on different time scales. As objects pass us by, e.g. their identities remain and do not vary fast. The representation of these objects and scenes on photoreceptors in the retina vary on a much faster time scale, because a small movement of the eye or the head may result in a completely different lightning condition and a fast variation of intensity received by the cells in the retina. These representations should vary on a similar time scale as the environment, and based on the principle of slowness, it is possible to extract slowly varying features out of the input signal. These features tend to be invariant with regard to fast changes in position or rotation. Due to their robustness to such transformations

they are able to reflect the properties of the environment.

In Wiskott and Berkes (2003) it was shown that learning based on slowness can extract features that have properties similar to complex cells. These properties are used for invariant object recognition based on slowness (Franzius et al., 2008) by building a hierarchical model with non-linear (quadratic) expansion (Franzius et al., 2007) and a linear readout on top of it. This concept is similar to the idea of the liquid-state-machine from Maass et al. (2002) that a neural microcircuit provides a non-linear expansion of the inputs, followed by a linear read-out map which is provided by the linear Slow Feature Analysis algorithm (SFA) from (Wiskott and Sejnowski, 2002).

In Li and DiCarlo (2008) temporal slowness is the main principle to achieve position invariance of IT neurons. In experiments with monkeys they showed that invariances are learned by temporal contiguity of object features during natural visual experience. Due to the tendency of contiguous retinal images to belong to the same object it was possible to uncover a neuronal signature by targeted alteration. They consistently swapped a specific pre-defined object if it appeared on a specific retinal position. After just a few training sessions, the visual system incorrectly associated the objects at different positions to the same object and drastic changes of position tolerance of IT neurons occured.

In contrast to Li and DiCarlo (2008), the focus of this thesis is not on read-outs of single IT neurons, but on a patch of V1 neurons. This patch of neurons is represented by a model from Schuch et al. (2009) and forms, together with detailled models of the retina and lateral geniculate nucleus (LGN) from Schuch et al. (2009), a non-linear preprocessing step for the SFA algorithm. The models of the retina and LGN are used to create biologically plausible inputs for the V1 circuit. The V1 model itself is based on the cortical microcircuit from Häusler and Maass (2007) and implements experimental data from Thomson et al. (2002). In contrast to Franzius et al. (2007) who used a quadratic expansion, the V1 model is used as an expansion method for pre-processing. The output of the circuit serves as an input to the linear SFA algorithm and no further processing stage is needed for a classification task. Thus, the extraction of object identity in our model is learned in a completely unsupervised manner and no teacher signal is needed for learning. This is an important aspect of this work, because it is questionable whether supervised learning occurs in a real biological circuit. The input that is used in Franzius et al. (2007) consists of different objects that carry out smooth movements. To generate the input for the model of the retina, the concept of saccades is used in this work. These saccadic eye movements are rapid shifts of the eye to quickly scan the environment. Between three and five of these movements are typically made per second. I show that position invariance can be achieved with the models of Schuch et al. (2009) and the method of SFA. Additionally I demonstrate that the experimental results from Li and DiCarlo (2008) can be reproduced

using SFA and detailled models for the retina, LGN and V1.

The thesis is structured in the following way. First of all I give a short overview of how visual information is processed through different stages of the visual system in Chapter 2. 3 shows some alternative approaches of translation invariant object recognition. In Chapter 4 the computational models will be described that are used to simulate the retina, LGN and V1. Chapter 5 gives a detailed mathematical description of the SFA algorithm. The simulation results will be presented in Chapter 6 and a discussion and outlook can be found in Chapter 7.

Chapter 2

The visual system

The following text is based on information from Bear et al. (1996), Purves et al. (2008).

The visual system and the processing of visual information is based on the ability to extract information out of light waves that are reflected from objects surrounding us everyday. Electromagnetic energy is absorbed, scattered, reflected and bent by all objects from the real world. It is a hard job to filter out useful information and therefore a lot of neural computation power is needed. This can be best attested by the fact that nearly half of the cerebral cortex is devoted to the analysis of the visual world.

The visual system begins with the eye. At the back of each eye is the retina with photoreceptors that converts light energy into neural activity. The eye has the same task as a camera and projects clear images of the world onto the retina. The output of the retina isn't an exact reproduction of the intensity of light, but it is specialized to detect differences in the intensity of light, almost regardless of the absolute intensity.

Axons of the retina are bundled into optic nerves and distribute information to different brain structures that are responsible for different tasks. The next processing step in this pathway is the lateral geniculate nucleus (LGN) that is situated in the thalamus, the largest part of the interbrain (diencephalon). From there visual information is sent to the cerebral cortex where it is interpreted and further processed.

A short introduction to all parts of the visual pathway will be given on the next pages, starting with the eye.

2.1 The eye

The eye is an organ that is responsible for the detection and analysis of light which is reflected vom objects. Light enters the eye through the pupil, which appears dark because of the light-absorbing pigments at the retina. To adjust to different light environments,

the pupil's size is controlled by a circular muscle, the iris. At the back of the eye is the retina, that is attached to the optic nerve via ganglion cells. The point where the optic nerve fibers exit the retina is called optic disk, where also retinal blood vessels originate. This small region is a dark spot and isn't responsible for sensing light because there are no photoreceptors and shadows on the retina are produced by these blood vessels. The macula can be found at the middle of the retina and does the job of central vision. Due to the absence of large blood vessels the quality of central vision is highly improved at this point. In the center of the macula lies a small dark spot, the fovea. It is responsible for sharp central vision.

2.2 The retina

After an image has been formed on the retina the light intensities are transformed into neural activity. The architecture of the retina is composed of photoreceptors, bipolar cells and ganglion cells. Only the ganglion cells are connected to the optic nerve and send action potentials towards the brain. Photoreceptors are the only light sensitive cells in the retina, all other cells are only influenced indirectly by synaptic connections. The cells in the retina are organized in layers that can be seen in Figure 2.1. These layers have to be seen inside-out. Light needs to pass the ganglion and bipolar cells before it is absorbed by the photoreceptors that sends back the neural information to the ganglion cells which are connected further to the optic nerve. Ganglion and bipolar cells are quite transparent, so image distortion is low.

There exist two types of photoreceptors in the retina, rods and cones. The former have a long and cylindrical outer segment and contain many membranous disks with light-sensitive photopigments. These pigments absorb light and trigger changes in the photoreceptor membrane potential. Cones on the other hand have a shorter outer segment containing fewer membranous disks. Due to the higher number of light-sensitive photopigments, rods are much more sensitive to light and are mainly responsible for vision during nighttime light. Cones are most active under daytime light. These different sensitivities are the reason why the retina is called duplex, one part only using rods and another part only using cones.

The structure of the retina changes gradually from the fovea to the peripheral areas. The central retina is dominated by cones and the peripheral retina by rods. Due to the fact that rods are most sensitive to nighttime light they do not have much influence at resolving fine details during daylight - this job is done by cones. Getting back to the fovea, it is the part of the retina with the sharpest and highest resolution, therefore it only consists of cone photoreceptors with high sensitivity to light.

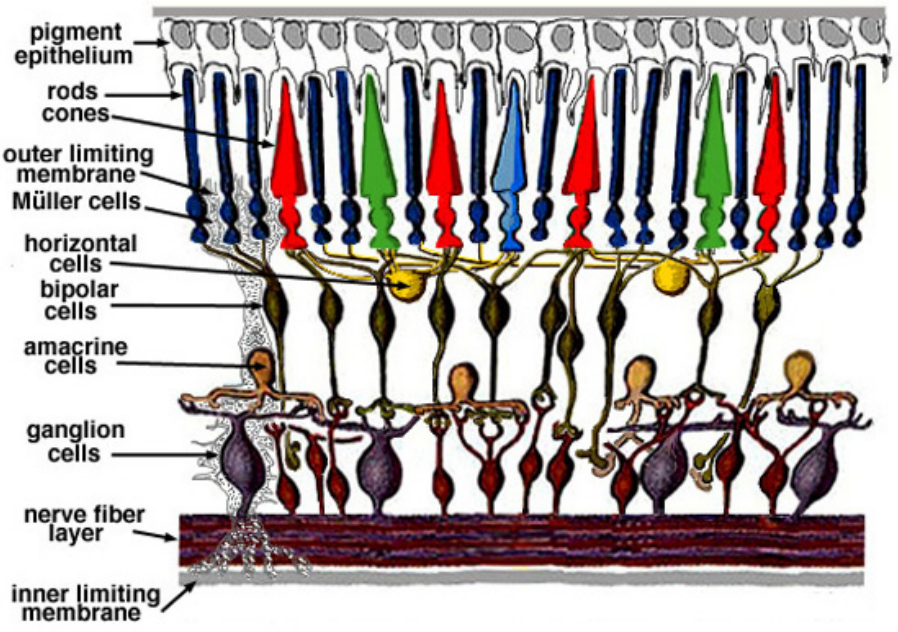


Figure 2.1: Overview of the retina. Image extracted from Kolb et al. (1996)

The retinal output are the action potentials arising from millions of ganglion cells connected to the optic nerve. Most of these cells have a concentric center-surround architecture, see Figure 2.2. In ON-center cells, action potentials are blocked when a dark spot enters the middle of the receptive field. An OFF-center cell in contrast will respond to a dark spot presented at the middle of the receptive field. In both types of cells the stimulation of the center is cancelled out by the stimulation of the surround. Due to the organisation of ganglion cells it can be seen why the visual system is specialized to detect local spatial variations mainly independent of the absolute light intensity.

2.3 The optic nerve, optic chiasm and optic tract

After the retina processed and transformed light waves into neural activity, these action potentials are transmitted to the optic nerve, optic chiasm and optic tract. The optic nerves of the left and the right eye cross each other at the optic chiasm. The optic nerves carry information from the nasal and temporal retina, but only information from the nasal retina crosses at the optic chiasm, so all information from the left eye is passed to the right half of the brain and all information from the right eye to the left half of the brain. This is done due to the fact that the visual fields from the left and the right eye partially overlap. The last stage before the information reaches the diencephalon is the optic tract. Most of

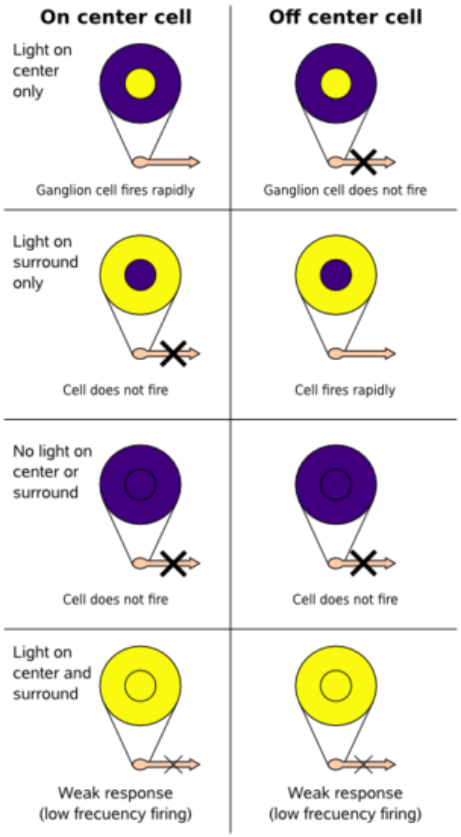


Figure 2.2: ON and OFF cells and their responds to different lightning conditions. ON center cells are stimulated by providing light to their center and are inhibited when their surround is exposed to light. OFF center cells have the opposite behaviour. By stimulating the center and surround of both types of cells, they only show a mild firing activity.

the axons of the optic tract are part of the so called lateral geniculate nucleus (LGN) of the dorsal thalamus.

2.4 LGN

The LGN is arranged in six layers, it serves as a gateway to the visual cortex (V1) and therefore to conscious visual perception. The information from the left and the right eye is kept separated and synapse on different layers from the LGN. The LGN receives information from retinal ganglion cells via the optic tract. As in the retina, the LGN also consists of ON and OFF ganglion cells and they are only connected to their namesakes of the retina. Most of the input to the LGN does not come from the retina, but from the visual cortex, forming a feedback loop. The exact role of this massive input (around 80%) is not clear yet. Each hemisphere of the brain has a LGN. In humans and primates each LGN has six distinct layers, where the first two layers are called magnocellular layers and the others are called parvocellular layers. Magnocellular cells receive their input from rods and are responsible for the perception of form, movement and difference in lumination. Parvocellular cells on the other hand, receive input from cones and are responsible for the perception of color and fine details. Each LGN receives input from both eyes, but due to the optic chiasm the left hemisphere receives input from the right visual field and the right hemisphere receives input from the left visual field. Information from the eyes is sent to different layers in the LGN. The ipsilateral eye (eye that is on the same side as the considered LGN structure) sends information to layer 2, 3 and 5 whereas the eye on the contralateral side (opposite side) sends information to layer 1, 4 and 6. The output of the LGN is projected further to the primary visual cortex (V1) through the optic radiations.

2.5 Primary visual cortex

The primary visual cortex (V1) consists of six layers named layer I to layer VI. The numbering is ordered from the outer part to the inner part of the brain. Layer I, just under the pia mater, nearly exclusively consists of axons and dendrites of cells in other layers. The V1, or striate cortex, has a thickness (from pia mater to white matter after layer VI) of about $2mm$, see Figure 2.4. There are actually nine different layers, but following Brodmann's convention (Brodmann, 1909) that neocortex has six layers, three sublayers in layer IV are combined. The LGN input is mainly fed into layer IV. This projection maintains the retinotopy of the LGN. This means that neighbouring cells in layer IV receive input via the LGN from neighbouring cells of the retina to preserve their positional influence of other neurons. About 20% of the neurons in layer II-VI are inhibitory and do not project axons to the outside of the V1 area. Layers II/III mainly consist of exhibitory

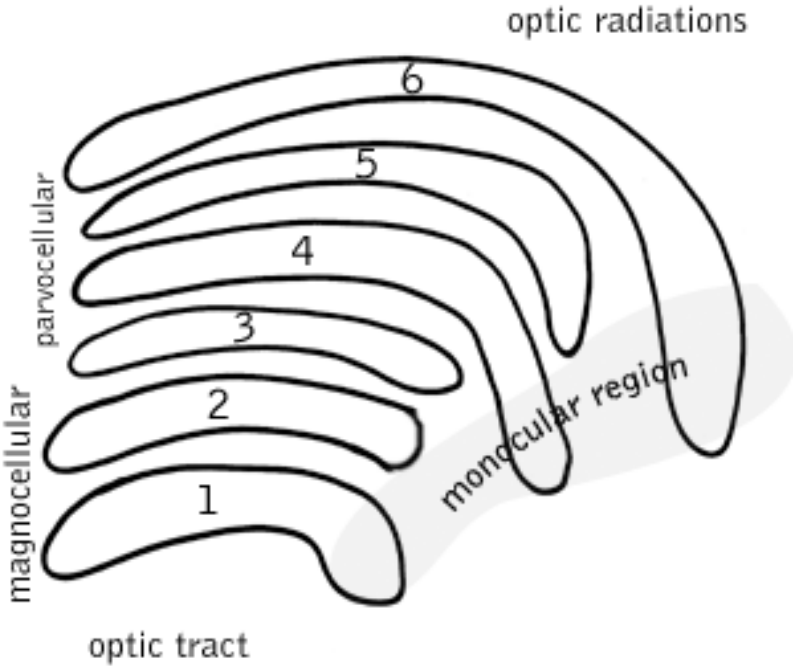


Figure 2.3: Layered structure of the primate LGN. The name arises from it's shape, that reminds of a bent knee (*genu* is latin for knee).

neurons and project further to extrastriate cortical neurons like V2, V3, V4, MT, etc. As already mentioned above, Layer IV receives input from the LGN and can be divided into 4 horizontal sublayers IVA, IVB, IVCa and IVCb. Layers V and VI contain many excitatory neurons that project back to the LGN to form a feedback path. In contrast to the upper 5 layers, layer I is nearly aneuronal, predominantly composed of dendritic and axonal connections. The segregated signals from the left and the right eye are still segregated in V1 via cortical columns. So special columns in V1 are dominated by either the input of the left or the right eye. The connections within cortical columns are called *radial* connections, running from the white matter to layer I, perpendicular to the cortical surface. In layer III, some of the connections of pyramidal cells extend collateral branches that make *horizontal* connections within layer III. Radial and horizontal connections play different roles in the analysis of visual information processing.

2.6 Simple and complex cells

Two important cell types that exist in the primary visual cortex are simple and complex cells. These cells have been analysed in detail by Hubel and Wiesel (1962). They found huge differences in orientation-specific cells, not only in terms of stimulus orientation or

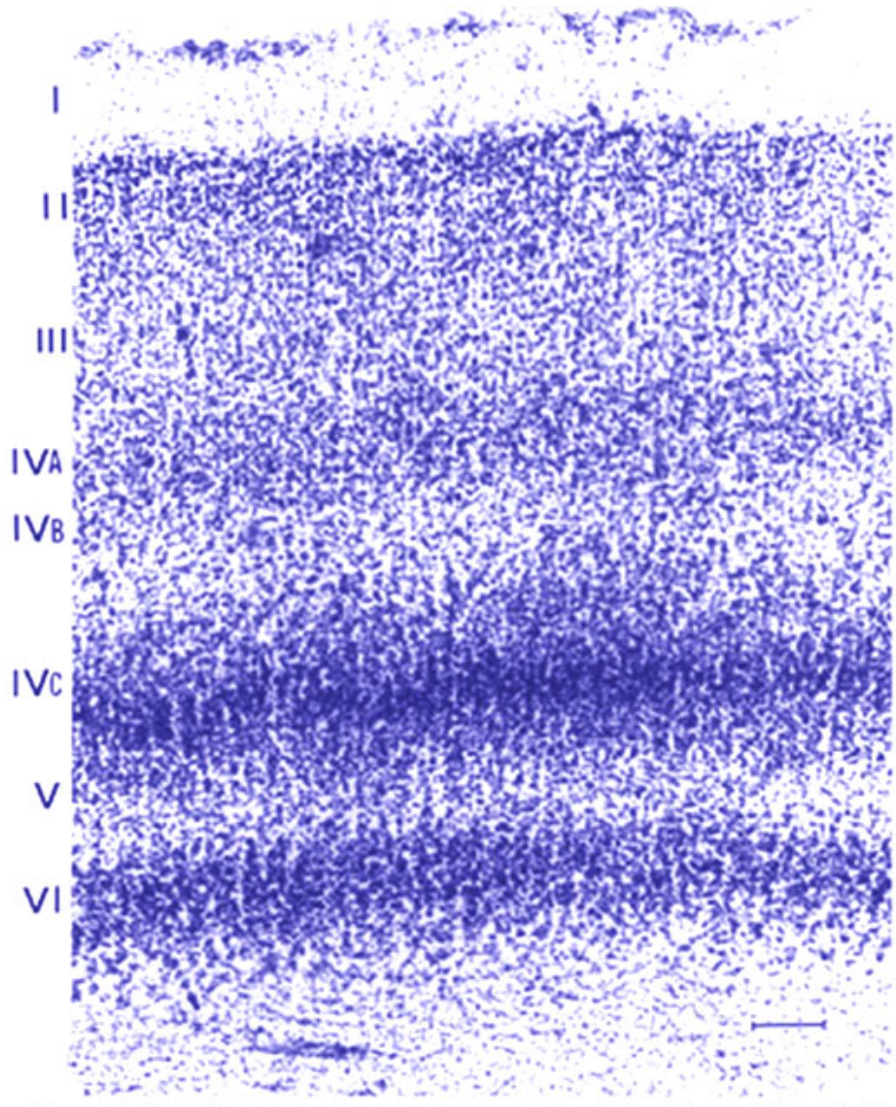


Figure 2.4: Radial cut through the cortical surface from the pia mater to the white matter. The different layers are numbered by roman numerals.

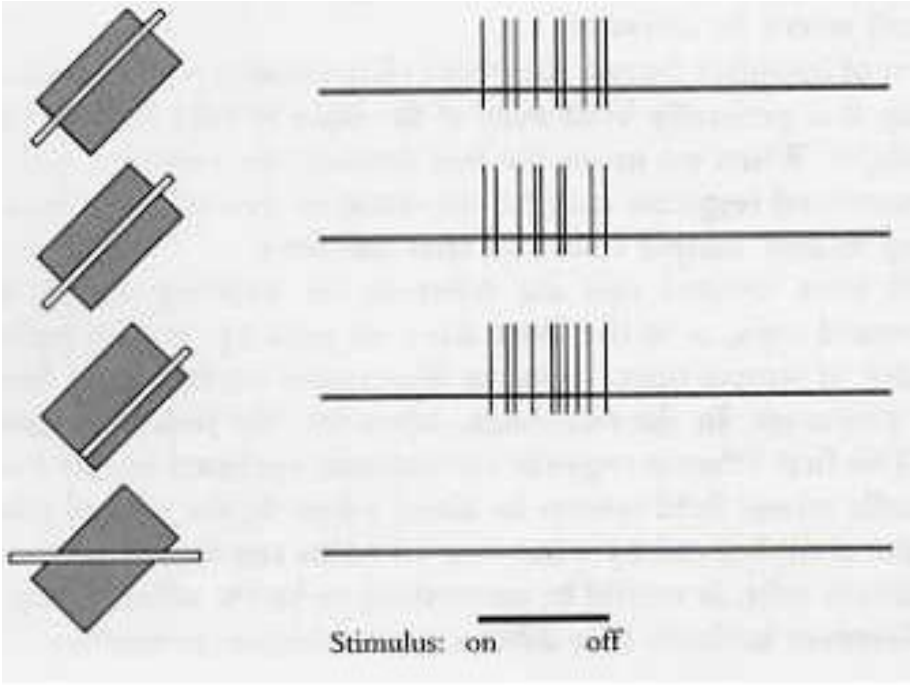


Figure 2.5: The dark shaded area represents the receptive field of a complex cell. The line represents a narrow slit of light and evokes a response independent of it's position in the receptive field as long as the orientation is correct. If the orientation is wrong, it does not respond to the stimulus. Image extracted from Hubel (1987).

position in the receptive field, but also in their behaviour. Simple cells can be found at earlier processing stages of the visual system and project further to complex cells. Orientation-specific cells respond only if the stimulus' orientation in their receptive field is appropriate, see Figure 2.5.

2.6.1 Simple cells

This type of cell responds primarily to oriented bars or edges and it's response is not invariant to changes in position or scale. The receptive field of such a cell can be divided into distinct antagonistic inhibitory and excitatory regions, see Figure 2.6 for a model of a simple cell by a Gabor filter as it is also used in the model of Schuch et al. (2009) and described in Chapter 4. Hubel and Wiesel (1962) suggested that the response of such a cell can be predicted by knowing these distinct regions.

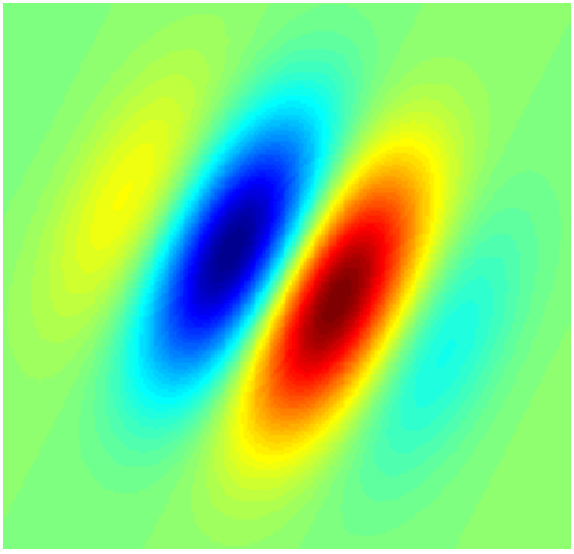


Figure 2.6: Modelling of a simple cell with a Gabor filter. Blue regions represent inhibitory regions, red ones excitatory regions.

2.6.2 Complex cells

Complex cells also respond to oriented edges or bars, but are more invariant to spatial changes like scale, rotation or position. Their receptive field can therefore not be divided into distinct inhibitory and excitatory regions. Complex cells receive input from groups of simple cells and their receptive field is therefore a summation of many receptive fields of simple cells. Due to the summation and the emerging large receptive field sizes, they respond to stimuli regardless of their exact position.

With a layered structure of alternating simple and complex cells it is possible to build a more and more complex representation of the visual information we perceive (Fukushima, 1980). This structure is used by many approaches to extract invariances of neurons or populations of neurons and will be described in Chapter 3.

2.7 Saccadic eye movements

Humans and many other animals do not stare at one specific point of an object for a long time, instead they make fast eye movements, called saccades. They only fixate on a specific location for a short period of time, then saccading to another location within the visual field and fixating again. One reason for this is, that the fovea at the center of the retina is responsible for sharp and clear images and therefore for high resolution. By moving the center of gaze over a scene or image, the whole field will be sampled by the fovea and therefore sharp images at the retina are provided. Another reason is

that these movements refresh the image for the cone and rod photoreceptors. Because they are sensitive to intensity changes the images would fade out without this refreshing mechanism.

A usual duration of a saccade varies between 20ms and 200 ms, depending on it's amplitude value. For an amplitude value of about 3° a saccade lasts for approximately 30ms (Morrone et al., 2005, Harris et al., 1990). The term amplitude refers to the angular distance the eye covers during movement. Between three and five saccades are made per second.

Chapter 3

Previous approaches to translation-invariant object recognition

The following chapter will present previous approaches to achieve translation invariances. The different models will be explained briefly to give an overview of what has been done so far and what are the differences to this work.

3.1 Neocognitron

The neocognitron by Fukushima (1980) is one of the oldest models based on simple and complex cells. It can be trained in a supervised and unsupervised manner. Figure 3.1 shows the architecture of the neocognitron. The input layer models an array of photoreceptors, followed by alternating layers of simple and complex cells. Simple cells serve as feature extractors and complex cells are used to allow positional errors of features of the stimulus. Simple cells have variable input connections that are learned during training. The connections to complex cells however are fixed. Each single complex cell receives input from a group of simple cells that extract the same feature but on different positions. The last stage of complex cells works as a recognition layer and each complex cell only responds to a specific pattern. Due to the fact that small errors are tolerated in the intermediate processing stages the output of the network is very robust to changes in shift and size. The model has also been extended to recognize occluded patterns. This has been done by adding backward connections to the standard model (Fukushima, 1987).

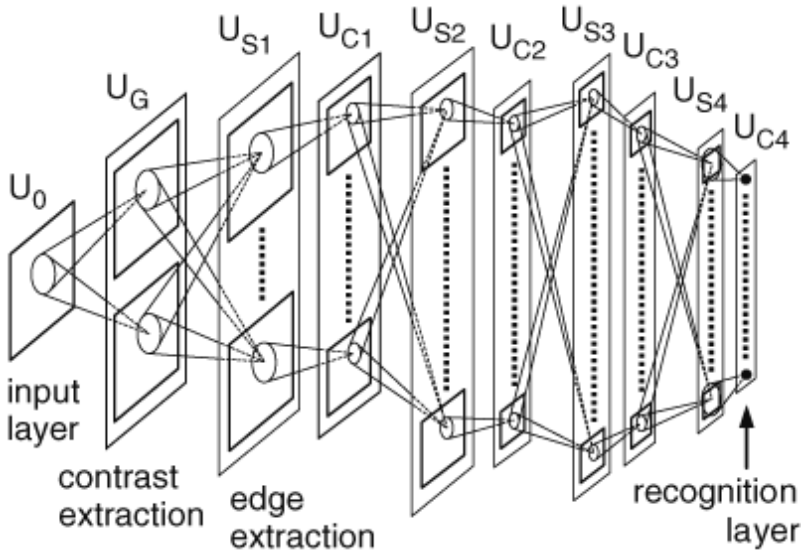


Figure 3.1: Schematic of the neocognitron. It shows how the alternating layers are connected. The neocognitron consists of alternating layers of simple and complex cells followed by a recognition layer. Image extracted from (Fukushima, 1980).

3.2 HMAX

The HMAX model by Riesenhuber and Poggio (1999) is another approach based on simple and complex cells, the schematic can be seen in Figure 3.2. In comparison to the neocognitron it is able to extract more invariances than only image-plane transformations. Another important feature of the HMAX model is, that it is quantitatively defined and comparable to experimental data. The basic idea of HMAX is that pooling over afferents, tuned to various transformed versions of the same stimulus, extracts invariances. These view-tuned units project further to view-invariant units. There are only a few output units that represent the stimulus and the view-tuned units are created during a learning process. By learning only on one single view of an object, the cells show limited invariance to three-dimensional rotation around the training view. This can also be used to extract scale and position invariances by only showing the object at one specific scale or position. The model is based on a simple feed forward architecture consisting of several layers. Riesenhuber and Poggio (1999) showed that feature specificity and invariances need to be achieved by different mechanisms. The model therefore shows two alternative pooling approaches, a linear summation method and a non-linear maximum operation where the strongest input is most responsible for the synaptic response. Summation of afferents that represent simpler features is a mechanism suitable for increasing feature complexity. They additionally showed that the maximum operation is the more robust one in case of recognizing invariances with lot of clutter and is biologically plausible. The latter hypothesis

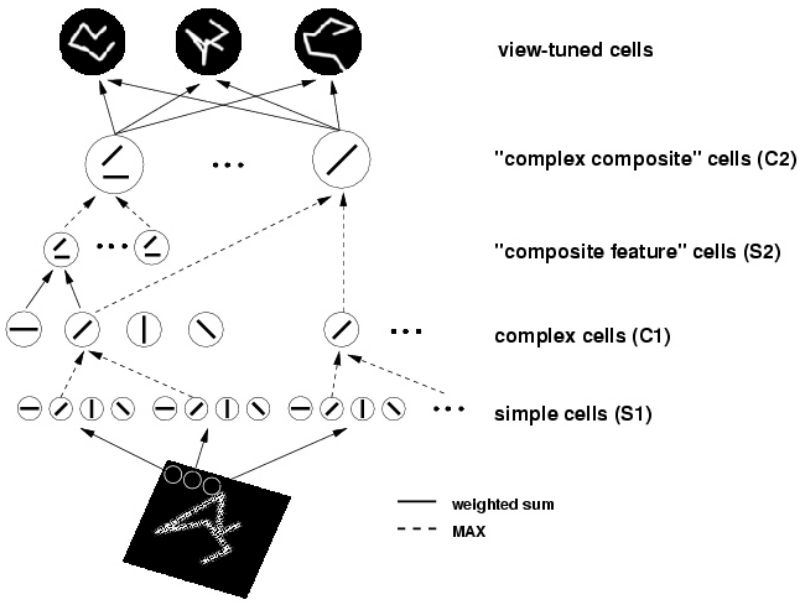


Figure 3.2: Basic HMAX model with several alternating layers of simple and complex cells to build more and more invariant object representations. This hierarchy consists of layers with linear units ('S' units, template matching, solid lines) and non-linear operators ('C' units, performing a MAX operation, dashed lines)

can be supported by a cortical operation that arises from microcircuits of lateral, possibly recurrent, inhibition between neurons in a cortical layer.

3.3 Invariant object recognition with SFA

Franzius et al. (2008) implemented a model based on temporal slowness. It is capable of extracting invariances of any aspect (e.g. size, viewing angle, position, etc.). The model is based on Franzius et al. (2007), where they present a model for the self-organization of place cells, head-direction cells and spatial-view cells based on unsupervised learning. To extract invariances they use a hierarchical structure of layers of SFA nodes, see Figure 3.3. Each node performs the following tasks: linear SFA for feature reduction, quadratic expansion, linear SFA for slow-feature extraction and a clipping procedure for extreme values. All the nodes have to be trained in advance. The model achieves two general aspects of visual processing: increasing receptive field sizes and accumulating power at higher layers. In general one can say that a single node performs a subset of a full quadratic SFA. This is one of the major differences to this work, where only linear SFA on the output of the V1 circuit is applied. Another difference is, that although the slow features can be extracted from raw image data, they used a multivariate linear regression

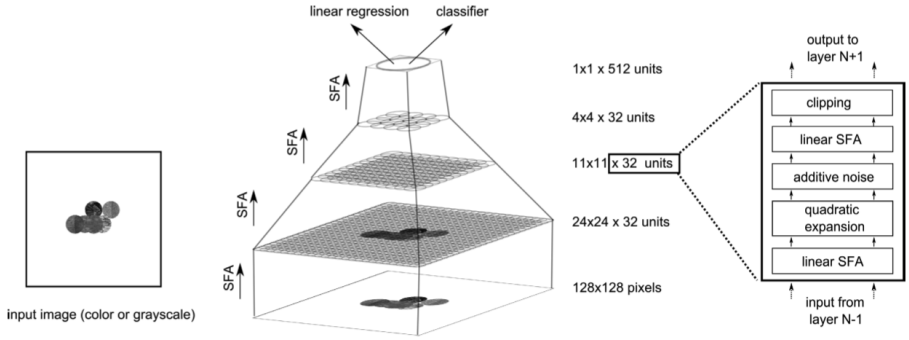


Figure 3.3: Hierarchical structure of SFA modules from (Franzius et al., 2008). The circles represent overlapping receptive fields that increase in every layer. On the lowest layer, the receptive field of a node represents a 10x10 pixel area and nodes partially overlap to cover the whole image. The second layer consists of 11 by 11 nodes and each of them receives input from 4 by 4 layer one nodes. The third layer contains 4 by 4 nodes, receiving input from 5 by 5 nodes from the previous layer. These 4 by 4 layer nodes all converge into a single node at the top. This layer is used as the actual SFA output layer.

(supervised) of the SFA output against the known configurations (parameter set that should be extracted). It is not possible to extract the relevant features linearly from the raw image data only by unsupervised learning because if many transformations occur simultaneously and on similar time-scales, solutions are mixed up. There is also a need to explain where the supervised processing occurs in a real biological circuit. The advantage of my work is, that it does not need any further processing stages after the slow feature extraction. The SFA itself can be used as a classifier on the position invariant features. One result of the work from Franzius et al. (2008) is that the object identity is encoded pretty well. For N objects, $N-1$ step functions are needed to represent the identities which are invariant to all other transformations. In the SFA output, this is represented by separated clusters for every single object. As a last step they applied a simple linear classifier to the output of the regression model, to show that new objects are represented by new clusters of data. The classifier (e.g. k-nearest neighbours) reached an accuracy of about 96%. The big advantage of this model is it's capability to extract all these different features at once. The identities e.g. of N different objects are optimally encoded with $N-1$ step functions and are invariant to all other transformations. But the main problem is the supervised classification procedure which is biologically not plausible.

Chapter 4

Models

In this chapter, the models from Schuch et al. (2009) for the retina, LGN and V1 will be described that are used in this thesis. They consist of a patch of V1 neurons that gets input from a model of the retina and the LGN. The V1 model is one of the largest and most detailed models of V1. It models an area of 25 mm^2 and is based on the cortical microcircuit model from Häusler and Maass (2007) built on experimental data from Thomson et al. (2002) on lamina-specific connection probabilities. The experimental data of the model comes from electrophysiological recordings of 4 macaque monkeys and experiments showed that the model is compatible with this real data (Schuch et al., 2009). The model has been extended laterally and anatomical particularities of macaques have been incorporated. First of all the V1 model will be described and subsequently the models for the retina and the LGN.

4.1 V1 model

The V1 model of Schuch et al. (2009) represents a 5x5mm patch of area in V1. It consists of 34596 neurons with nearly 3.6 million synapses. For the simulations in this work only 3249 neurons and 219135 synapses are used. Because there are much more neurons in a real 5x5mm patch of V1 the synaptic weights were modified to achieve biologically plausible results. Also the strength of the synaptic connections from the LGN input to V1 was scaled to obtain realistic results.

The circuit is similar to the cortical microcircuit model of Häusler and Maass (2007) and consists of three layers, corresponding to the cortical layers 2/3, 4 and 5. Every layer has a population of excitatory and inhibitory neurons equally distributed over all layers with a ratio of 4:1 and the layers are connected like in Figure 4.1. The excitatory pools consist of regular spiking, intrinsically bursting, and chattering cells. The inhibitory pools are comprised of fast spiking and low-threshold spiking neurons. Although the data from

Thomson et al. (2002) originates from rat and cat, the connectivity structure of macaque is quite similar to that of cat. In contrast to the cortical circuit of Häusler and Maass (2007) a simpler neuron model of Izhikevich (2003) was used to speed up simulation time. This model can easily be adjusted to achieve different firing dynamics. Additionally to the neuronal input from other neurons in the model, each neuron receives synaptic background input to model the absence of neurons that are not implemented in the model. This input is responsible for the depolarization of the membrane potential and a lower membrane resistance. This state of neurons with depolarized membrane potential and low membrane resistance is commonly referred as 'high conductance state' (Destexhe et al., 2001).

The model of Häusler and Maass (2007) consists of one column with about $100\mu m$ of diameter. The model of Schuch et al. (2009) is extended to several millimeters laterally so that connection probabilities depend on the lateral distance. For intra-cortical connections, a bell-shaped Gaussian distribution is used to determine the connection probabilities laterally. The standard deviation of these Gaussians is set to $200\mu m$ for excitatory neurons and about $150\mu m$ for inhibitory neurons because of the occurrence of very narrow inhibitory dendritic spreads observed (Lund et al., 2003).

4.2 Input model

Due to the fact that the V1 model processes visual data, a realistic input model is needed that transforms real world visual data into representations of spike trains. The retina model and the LGN model are spatio-temporal filter banks with non-linearities (Gazeres et al., 1998). The time-varying input signal on the retina is converted into firing rates of LGN neurons by these filter banks. All color information is neglected by converting it to gray scale before serving as input to the models. This is done for simplicity due to the fact that it is neglected that ganglion cells typically react to color differences instead of pure luminance differences.

4.2.1 Retina model

The input to the retina (2-dimensional) is filtered by 'Mexican hat' difference of Gaussian spatial filters and the filter sizes (representing the receptive fields of ganglion cells) are adapted to data of macaques. The standard deviations (σ_{center} and $\sigma_{surround}$) of the Gaussians for center and surround of these cells are estimated. After convoluting these kernels with the cells' luminances the response of e.g. a retinal ON-cell at a specific retinal position r can be described by

$$R_{ON}(r) = C(r)\left[S_{center}(r) - wS_{surround}(r)\right]. \tag{4.1}$$

w represents the ratio of center to surround and is defined using data from Croner and Kaplan (1995). By applying these Gaussian bumps to the luminance a quantity called 'contrast gain' results. To compute the actual firing rates, the 'contrast gain' has to be multiplied by a local contrast

$$C(r) = \frac{|S_{center}(r) - S_{surround}(r)|}{S_{center}(r) + S_{surround}(r)}. \tag{4.2}$$

The firing rates are estimated for lagged and non-lagged ON-OFF-cells. Lagged and non-lagged cells can be distinguished by the timing of their responses to stimuli, they have different phases. For an arbitrary stimulus, lagged cells have a time-delayed response w.r.t. non-lagged cells. For simulations in this work all lagged cells were neglected. Otherwise the output of lagged cells would occur exactly at the time when the subsequent input shows up at the receptive field and leads to overlapping inputs.

4.2.2 LGN model

The LGN model filters the retinal output using a temporal kernel with a phasic and tonic component, and combinations of these were assigned to lagged and non-lagged ON-OFF cells. Altogether there are 4 different time-varying output rates for every possible position in the LGN, e.g. any combination of lagged and non-lagged cells in the LGN with either ON- or OFF-cells from the retina. These rates will be converted into spike trains by a 'switching Gamma renewal process' (Gazeres et al., 1998). This process produces a higher spike time regularity for high input rates and switches to a Poisson process for lower input rates. The input connection probabilities from the LGN model to the V1 model can be seen in Figure 4.1. The thalamic input of the LGN (input stream 1) is mainly connected to layer 4. A second input (input stream 2) is modelled as background activity of the LGN and is only fed into layer 2/3 with a connection probability of 20%.

The data from the LGN model is passed to the V1 neurons by a retinotopical mapping. This means that the neurons in V1 are organized topographically to form a 2D representation of the image formed on the retina. Neighbouring regions of the image are represented by neighbouring regions of the visual area. Theses representations are distorted, due to the fact that the fovea is a much more important part of the retina for vision than the peripheral regions and so the fovea area is represented by a much larger area in V1. This mapping in the model is realized with oriented Gabor functions, a 2-dimensional Gaussian

multiplied by a cosine function.

The models are all implemented in Python.

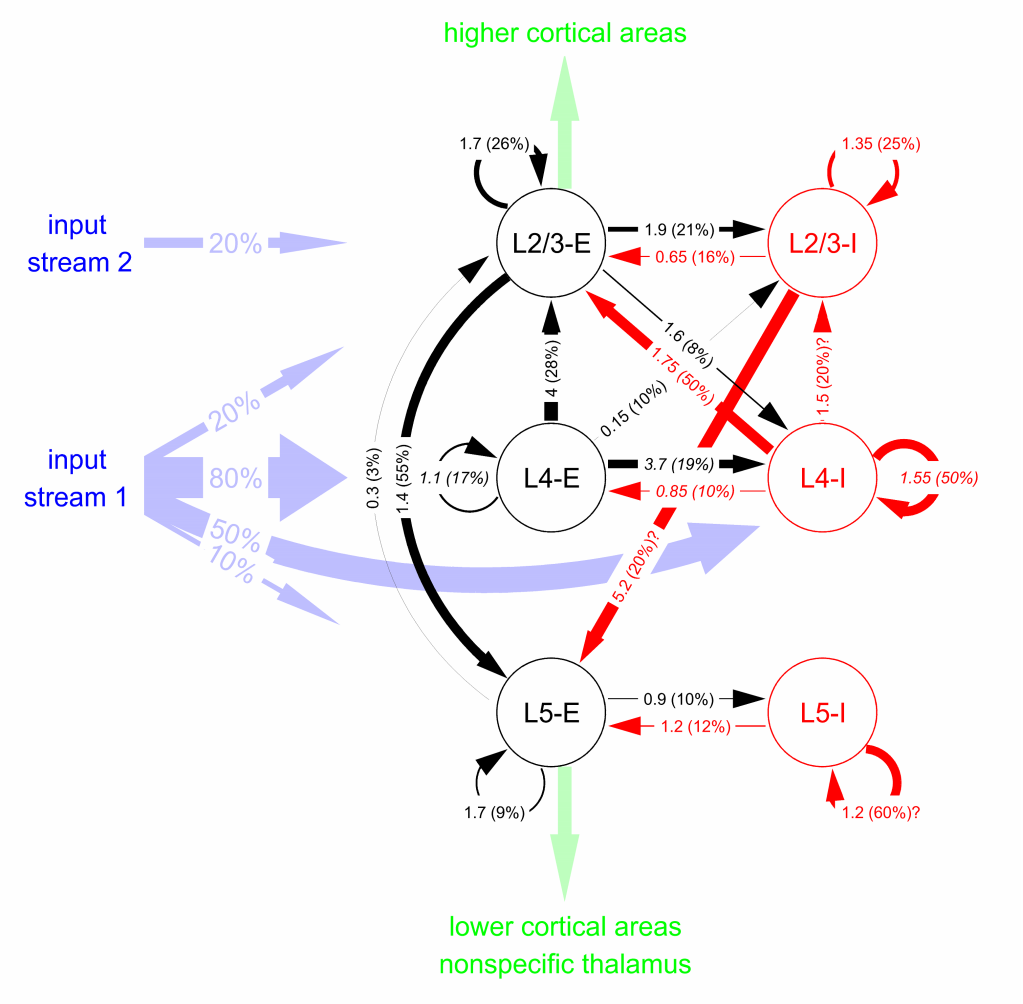


Figure 4.1: Cortical structure with layers 2/3, 4, 5 and their connection probabilities. The numbers on arrows represent the connection strengths and connection probabilities (in parantheses). Every layer consists of excitatory and inhibitory neurons, marked with E or I respectively. Percentages at input streams denote the connection probabilities of inputs to the different layers. Image extracted from Häusler and Maass (2007)

Chapter 5

Methods

5.1 Slow Feature Analysis

Slow Feature Analysis (Wiskott and Sejnowski, 2002) is a learning principle that extracts slow features from a multidimensional input signal $\mathbf{x}(t)$. Given this input signal $\mathbf{x}(t)$ the method tries to find multidimensional input-output functions $\mathbf{g}_i$ that generate as slowly varying output signals $\mathbf{y}(t) = \mathbf{g}(\mathbf{x}(t))$, but still including enough significant information. This can be described mathematically by the following equations:

$$min\Delta(y_i) := \langle \dot{y}_i^2 \rangle_t, \tag{5.1}$$

where $\langle \cdot \rangle_t$ indicates the temporal averaging of a signal and $\dot{y}_i$ is the time derivative of y_i. To compute the slow features, the output signal has to fulfill the following constraints:

$$\begin{aligned} \langle y_i \rangle_t &= 0, & (5.2)\\ \langle y_i^2 \rangle_t &= 1, & (5.3)\\ \langle y_i y_j \rangle_t &= 0 \ \forall j < i. & (5.4) \end{aligned}$$

The constraints can easily be described. Equation (5.2) and (5.3) simply avoid the trivial constant solution $y_i(t) = 0$ and impose zero mean and unit variance. Constraint (5.4) makes sure that all features are different and uncorrelated. Furthermore it introduces an ordering between different slow features.

An important aspect is, that the mapping from input to output happens instantaneously. So although the objective is based on slowness, the processing of the input is done fast after learning has completed. The algorithm can only extract slow features if

there is enough slowly varying information in the input signal, so no low-pass filtering is allowed or useful to generate the output signals.

By reducing the set of output functions to the set of linear functions, the optimization problem of (5.1) can be simplified. The procedure is based on an eigenvector approach and guaranteed to find a global minimum. Without the loss of generality we can assume $\mathbf{x}$ to have zero mean ($\langle \mathbf{x} \rangle_t = 0$), automatically fulfilling constraint (5.2):

$$\langle y_i \rangle_t = \mathbf{w}_i^T \langle \mathbf{x} \rangle_t = 0. \tag{5.5}$$

Constraint (5.3) can be integrated into the objective function (5.1):

$$min \Delta(y_i) = \frac{\langle \dot{y}_i^2 \rangle_t}{\langle y_i^2 \rangle_t} = \frac{\mathbf{w}_i^T \langle \dot{\mathbf{x}} \dot{\mathbf{x}}^T \rangle_t \mathbf{w}_i}{\mathbf{w}_i^T \langle \mathbf{x} \mathbf{x}^T \rangle_t \mathbf{w}_i} \tag{5.6}$$

The covariance matrix of the input is denoted by $\langle \mathbf{x} \mathbf{x}^T \rangle$ and needs to be positive definite and invertible. By fulfilling these constraints, all eigenvalues are greater than zero. $\langle \dot{\mathbf{x}} \dot{\mathbf{x}}^T \rangle$ denotes the covariance matrix of the time derivatives of $\mathbf{x}$. To minimize (5.6) the weight vectors $\mathbf{w}_i$ have to be optimized and are solutions of the generalized eigenvalue problem

$$\langle \dot{\mathbf{x}} \dot{\mathbf{x}}^T \rangle_t \mathbf{W} = \langle \mathbf{x} \mathbf{x}^T \rangle_t \mathbf{W} \mathbf{\Lambda}, \tag{5.7}$$

with $\mathbf{W}$ being the matrix of generalized eigenvectors and $\mathbf{\Lambda} = diag(\lambda_1, ..., \lambda_n)$ the diagonal matrix of generalized eigenvalues. By normalizing these eigenvectors such that $\mathbf{w}_i^T \langle \mathbf{x} \mathbf{x}^T \rangle_t \mathbf{w}_j = \delta_{ij}$, constraints (5.3) and (5.4) are fulfilled.

The objective function (5.1) now takes on the values of the corresponding eigenvalues for these solutions:

$$\Delta(y_i) = \lambda_i. \tag{5.8}$$

The solution vector corresponding to the slowest feature is the generalized eigenvector, $\mathbf{w}_1$, corresponding to the smallest eigenvalue, λ_1; $\mathbf{w}_2$, corresponding to λ_2, representing the second slowest feature, and so on.

For the more general case, where the output functions are not limited to the set of linear functions, an expanded function space F can be generated, e.g. all monomials up to order d which can be considered as a basis $h_1, ..., h_M$ of F which is the set of all polynomials of

degree d. Every single output function $\mathbf{g} \in F$ can now be expressed by a linear combination $g(\mathbf{x}) = \mathbf{w}^T\mathbf{z}$ of these non-linear expansions $\mathbf{z} = \mathbf{h}(\mathbf{x}(t)) = (h_1(\mathbf{x}(t)), ..., h_M(\mathbf{x}(t)))$. The optimization task in this high dimensional space can then be performed as described above.

5.1.1 Importance for visual processing

The signals our eyes receive contain information about the surrounding environment. Even if an object in the real world varies or moves slowly, a drastic change in the eyes' signal can occur. If e.g. an animal slowly moves in front of us, it's position is changed slowly. The visual signal however changes in a much faster way dependent on the position, velocity, angle or distance of the animal. SFA provides a method to filter out these slowly varying signals from a fast varying input signal. It reflects the brain's mechanism to e.g. filter out the identity of the passing animal. This is not the only information we can extract of the visual signal. In addition to it's identity we can also determine the position independently of it's identity or viewing angle (Franzius et al., 2008).

5.2 Mutual Information

Mutual Information (MI) is a quantitative statistical measure from information theory that measures how much information one random variable contains about another variable. The mutual information is strongly linked to the entropy H and conditional entropy (discussed below) and is calculated in the following way:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{5.9}$$

The variable X is the input to a specific system or channel, Y denotes the output. In the case of this thesis, X represents the slow features computed by the SFA algorithm and Y represents the class of the object that produced these slow features. MI can also be defined by the use of probability distributions

$$I(X;Y) = \sum_x \sum_y p(x,y) \cdot log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{5.10}$$

or by the expected value operator:

$$I(X;Y) = E \left\{ log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right) \right\} \tag{5.11}$$

The joint probability distribution is denoted by $p(x, y)$ and the marginal distributions by $p(x)$ and $p(y)$. If the mutual information is zero, the two variables are statistically independent. This means that if full information about one random variable is known, it does not tell us anything about the other variable. The mutual information is maximal, if information about one random variable can fully describe the other variable. If the mutual information increases, the uncertainty about one random variable decreases, assuming that the other variable is known. The uncertainty completely vanishes, if the mutual information reaches it's maximum value. The uncertainty about one random variable can therefore be expressed by the knowledge about the other variable. The quantity is dimensionless and usually measured in bits if the log_2 is used.

$H(X)$ denotes the entropy of the random variable X. It is a measure of uncertainty and was made quantitative by Shannon (Shannon and Weaver, 1949). The entropy satisfies the following conditions:

- It is maximal when $p(x)$ is uniform, it increases with the number of possible values x can take.
- It stays the same, if the probabilities are reordered (assigned to different values of x).
- The principle of additivity is valid. The uncertainty about two random variables is the sum of the single uncertainties.

The entropy can be seen as the optimal number of yes/no question that is needed to guess the value of a random variable, assuming that the underlying probability distribution is known. The conditional entropy $H(X|Y)$ is defined as the average uncertainty of X after a second random variable Y is observed. The mutual information is symmetric $I(X;Y) = I(Y;X)$ and additive for independent random variables. Figure 5.1 clearly shows the relationship of entropy and mutual information.

5.3 Principal component analysis

Principal component analysis is a powerful method to represent high dimensional data in a lower dimensional subspace. For an n-dimensional input vector $\mathbf{x}$, the data can be represented by these components in an m-dimensional subspace by a linear combination of these m principal components, $m < n$. Possibly correlated data will be decorrelated, but only for normally distributed data sets they will be statistically independent.

To calculate the principal components, the covariance matrix Σ of the n-dimensional input data $\mathbf{x}$ (mean value has been subtracted) has to be calculated. The covariance

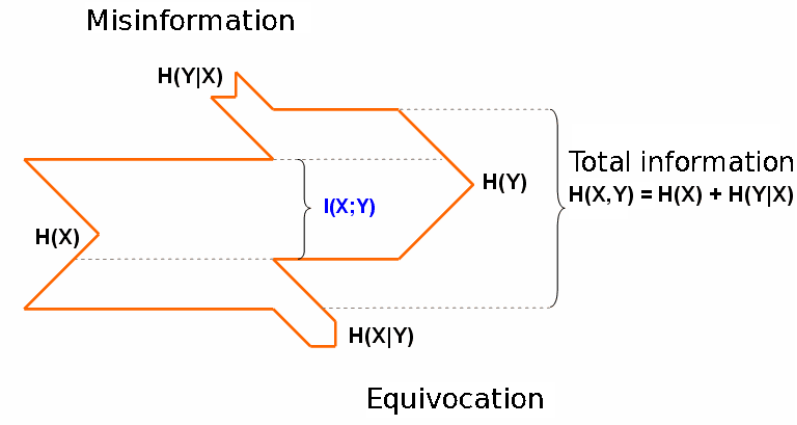


Figure 5.1: The diagram shows the relationship between entropy and mutual information. The entropy on the left represents the input to a channel/system, the entropy on the right side represents the output. The information that comes from the source and runs directly through the channel to the output is called mututal information.

matrix Σ is symmetric and positive definite. The eigenvalues $\lambda_1, ..., \lambda_n$ of Σ are sorted descendently and form a diagonal matrix Λ. The largest eigenvalue corresponds to the largest direction of variability in the data and is therefore the main principle component. The subsequent eigenvalues correspond to decreasing variability in the data. The eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ are column vectors and form the orthogonal matrix $\Gamma_n = [\mathbf{v}_1, ..., \mathbf{v}_n]$ with $\Lambda = \Gamma^T \Sigma \Gamma$. The input vector $\mathbf{x}$ can now be linearly transformed to $\mathbf{y} = \Gamma^T \mathbf{x}$. Commonly the first few principal components are responsible for nearly 100% of the input data variability. But as the covariance matrix is a different one for every dataset, this needs to be checked in advance. By only choosing an adequate subset of m first eigenvectors the data can now be represented by these m orthogonal and uncorrelated principal components. The data is then transformed only using this subset of components $\Gamma_m = [\mathbf{v}_1, ..., \mathbf{v}_m]$.

The SFA and PCA algorithms are implemented in Python using the MDP (Modular Data Processing) Toolbox from (Berkes and Zito, 2005). The Mutual Information is also calculated using Python and the PyEntropy package from (Ince et al., 2009)

Chapter 6

Simulations

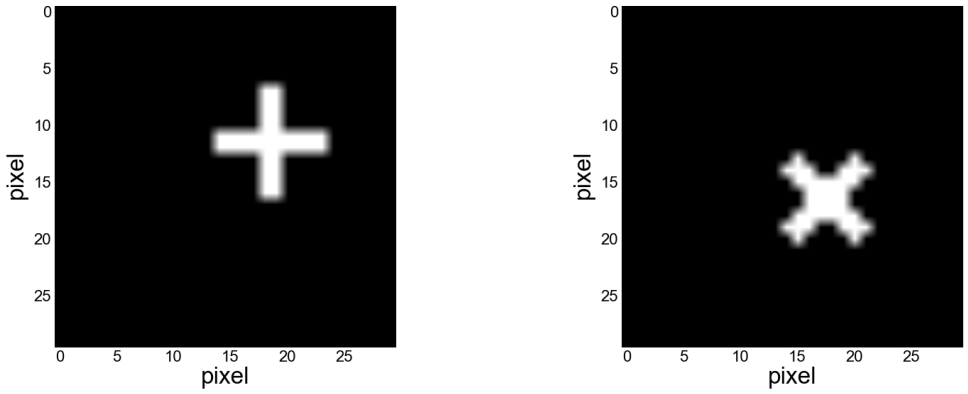
As described in the following, I investigated the performance of the V1 model as a non-linear pre-processing step to the SFA algorithm on visual data. The performance of the circuit is measured by using SFA to extract the identity of the object presented in the visual input. The quality of the predicted object identity is then measured by the mutual information of the prediction with true identity of the presented object. I compare different setups of the neural circuit to show the importance of the V1 model as a non-linear expansion. First of all the generation of the input data will be described, followed by the simulation of the V1 model, including the models for the retina and LGN, with appropriate input sequences. The last part shows that SFA applied to the output of the V1 circuit shows the same behaviour as the expirement of Li and DiCarlo (2008) does, where only the responses of IT neurons have been taken into account.

6.1 Input data generation

The input mainly consists of sequences of two objects, a '+' (plus) and a 'x' (cross), see Figure 6.1. These sequences are put together to create a time-varying input signal, a movie.

To generate biologically plausible data the input consists of a saccade and a fixation period as shown in Figure 6.2. The saccade period consists of a simple gray-value image followed by a fixation period that contains either the object '+' or the object 'x' at different locations in the image. Such a single sequence of saccade and fixation is referred to as a single stimulus in this work. Because the input is modelled with saccades, transition between different objects occur abruptly and not smoothly.

To generate training and test data, sequences of single stimuli have to be generated and concatenated together in time. The initial object is chosen randomly with a probability of 0.5. All subsequent stimuli are drawn according to the Markov model shown in Figure



(a) stimuli object +

(b) stimuli object x

Figure 6.1: Examples of visual stimuli presented to the retina model in the simulations. In a) one can see the stimulus class one, which is represented by a '+' (plus) and in b) the second class, being a 'x' (cross)

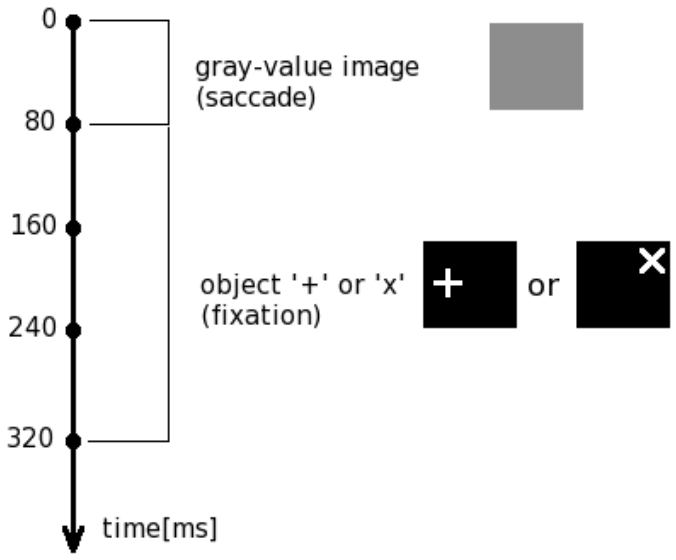


Figure 6.2: Timing of a single stimulus consisting of a saccade and fixation period. During the saccade, a grey-value image is presented. Duration the fixation period one of the two possible objects ('+' or 'x') is presented.

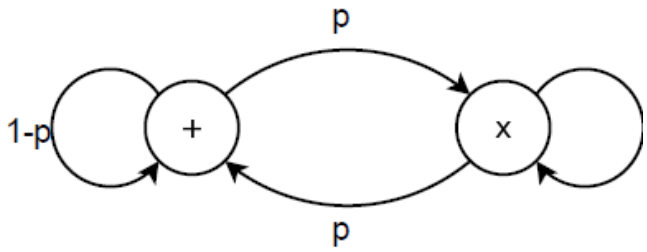


Figure 6.3: Markov model for generating test and training sequences of stimuli. The objects switch with probability $p = 0.001$ to generate sequences with many consecutive objects of the same class. Image extracted from Klampfl and Maass (2009, 2010)

6.3 with a probability of switching the object of $p = 0.001$. This procedure generates sequences with many consecutive objects of the same class and therefore a slowly varying object identity.

The position of the pattern ('+' or 'x') is chosen randomly within the image frame for every single fixation period. This models the situation that the object actually stays at the same position but with a different retinal fixation, so that the object shows up on a different location at the retina.

Due to the spatial filtering of the retina by a Difference of Gaussian filter, edge effects occur at the margins of the image. The object will therefore not be placed close to the image border. The whole image has a size of 30x30 pixel, the object patterns themselves are smaller with only 8x8 pixel. The two different patterns have the same luminance or in other words, they consist of the same amount of white pixels. This is necessary to envoke the same mean firing rate of the LGN model, to ensure that learning is not simply based on the amount of white pixels. Before the images serve as an input to the retina and LGN model, it is important that no information is lost during this process. The size of the receptive field in degree is 1.8°, the size of the LGN field is smaller and has only 1.5°. Due to the spatial filtering on the edge of the image, edge effects have to be taken into account. To avoid these effects the patterns are placed within a margin of appropriate width. As the whole image has 30 pixel that correspond to 1.8 of the visual field, one pixel equals 0.06° degree. For a difference of $1.8° - 1.5° = 0.3°$ the margin has to have a width of 5 pixel. This constraint assures that no filtering effects occur. Taking this margin into account, 20 horizontal and vertical pixel remain to be chosen as the pattern's position. With a pattern size of 8 pixel we have a total number 144 different possible retinal positions where the pattern can occur.

6.2 Position-invariant object recognition using SFA

In the following I describe the simulation to generate invariant object representations. To differentiate between the various setups, the performance of SFA on the output of the V1 circuit needs to be figured out. The input data described in the previous section is fed into the retina and LGN model to transform the image sequences into spike trains. Sequences consisting of 4500 training patterns and the same amount of test patterns are created. As one single stimulus has a duration of 320ms (saccade and fixation period), the simulation time for the sequence of all 4500 training patterns would be too long. To speed up the simulation, only batches of 50 consecutive sequences are simulated at once and are distributed on different cluster machines. The same network model is used for each batch. After all simulations have finished, the results are concatenated in the correct temporal order to obtain the time-varying circuit response to the whole input sequence. In Figure 6.4 one can see the LGN response of a sequence of four objects appearing subsequently. As described in the previous chapter all lagged cells have been deactivated. If they were activated, the lagged response of the first object would occur exactly at the time the subsequent object shows up at the retina and would influence the neural response. Due to the fact that the objects are placed randomly in the image frame, it will probably occur that two objects seem to form a new one by partially overlapping. This fact would create lots of different objects and SFA would not be able to filter out any slowly varying object identity. In Figure 6.5 one can see a possible overlap and the emerging new pattern consisting of two objects. The spike train output of the LGN model served as the input to the V1 patch. For the connection probabilities of each layer in V1 see Figure 4.1. The output of the V1 model are 3249 spike trains, one for each neuron.

The duration of saccades also has a huge impact on the performance of SFA. Durations lasting from 30ms to 80ms are tested for the presented simulations, see Figure 6.6. The usual time it takes to perform a saccade with an amplitude of 3° is around 30ms (Morrone et al., 2005, Harris et al., 1990). For higher amplitudes, durations increase. The performance measure for different saccades are computed using mutual information between the output of the SFA algorithm for object '+' and object 'x'. With decreasing duration of saccades also the performance decreased. A saccade duration of 80ms and a fixation period of 240ms are chosen for all simulations because good performance was achieved for these values, even though they are not completely consistent with biological data.

The output of the V1 circuit is represented by spike-trains. To apply this output to the SFA algorithm, the spike trains are low-pass filtered with a time constant of 30ms to achieve an analog output signal. A histogram of the firing rates of all neurons can be seen in Figure 6.7. In each fixation period, a single snapshot of the low-pass filtered spike trains at time 200ms from the saccade onset was taken. The mutual information has been

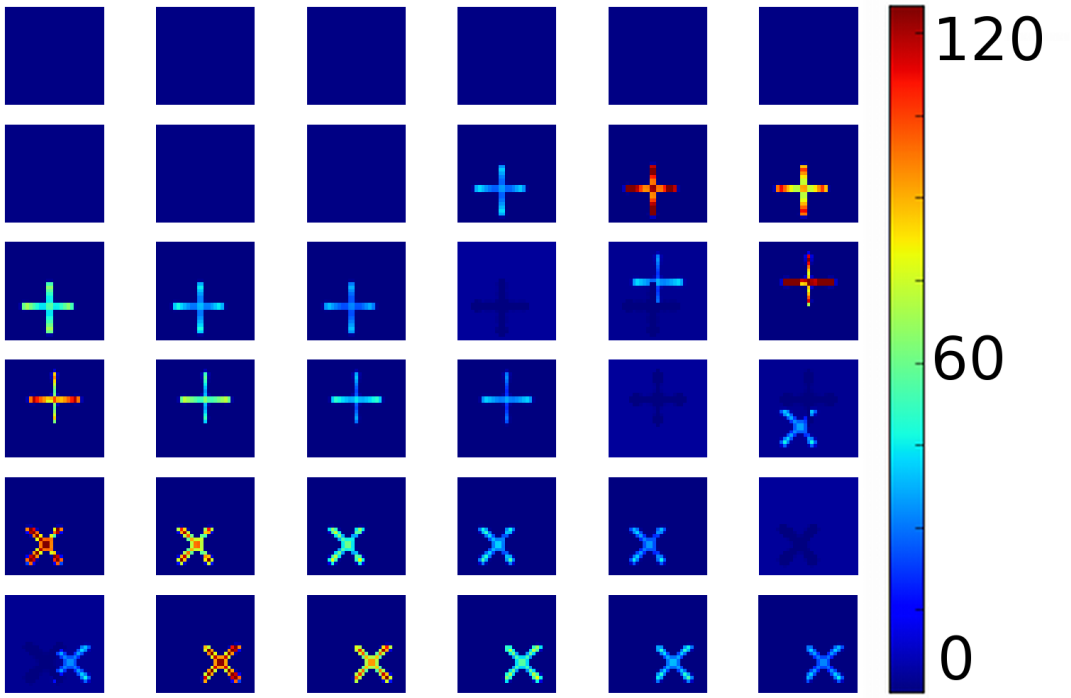


Figure 6.4: A sample response of the LGN model on four objects showing up subsequently with all lagged cells turned off. The frames are ordered row-wise, this means that the first row holds the first six frames, the second row holds frames seven to 12 and so on. The first nine empty frames are just dummy frames. First, there are two '+' shown, divided by a saccade (blue frame). During this saccade, little influence from the previous object can be seen due to the slow fade out of neurons that have be excited by the previous object. With lagged cells turned on, this influence would be much stronger and would last for a longer period of time. Different colors represent different activations of neurons.

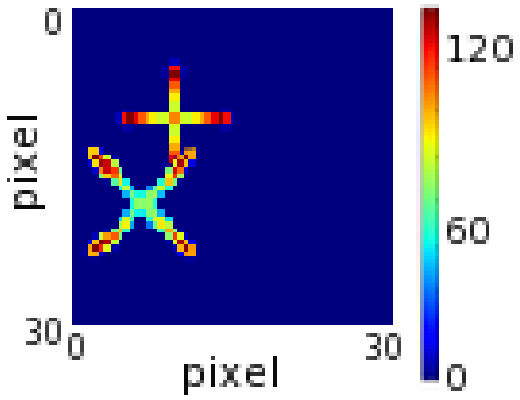


Figure 6.5: If the lagged cells are activated this figure shows a possible overlapping condition consisting of two objects.

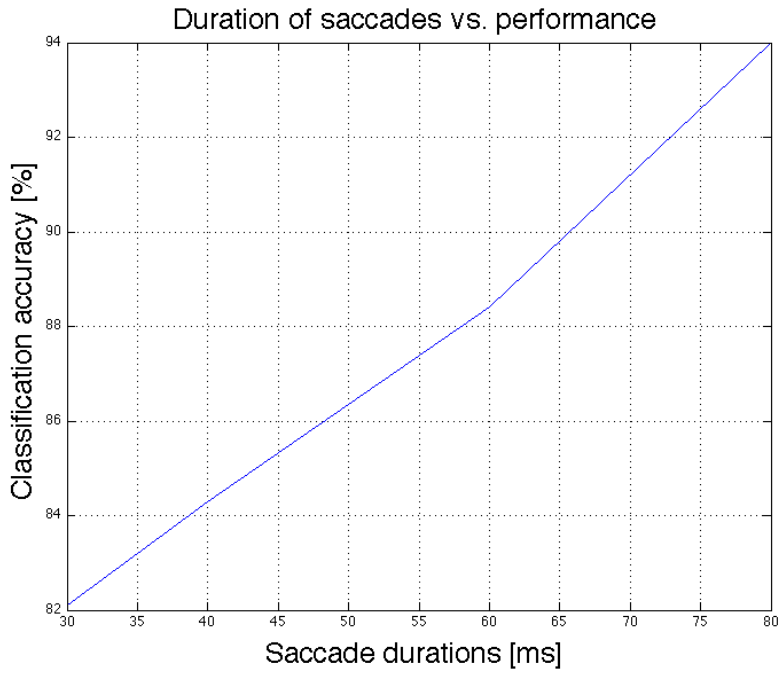
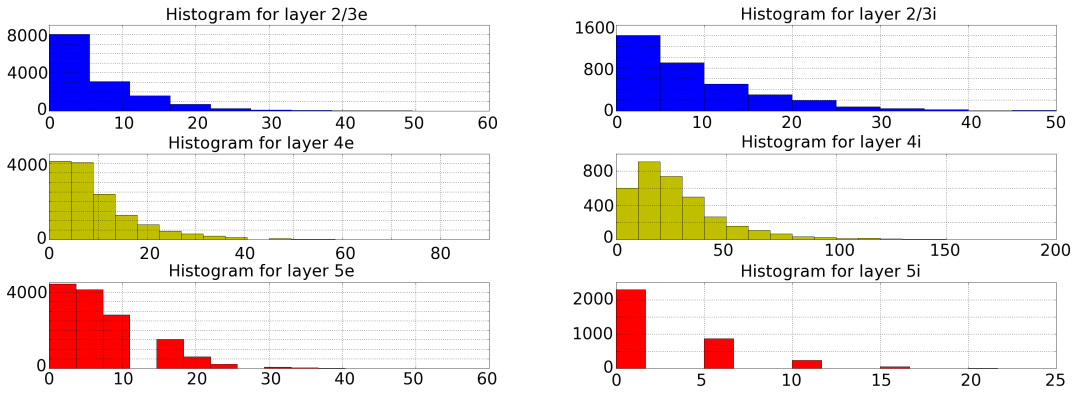


Figure 6.6: Classification performance for different duration of saccades. The performance decreased with smaller durations.



(a) Firing rates of exc. neurons

(b) Firing rates of inh. neurons

Figure 6.7: Firing rate distributions of neurons in the V1 patch. In a) one can see firing rates for every layer computed over all excitatory neurons. b) shows the firing rates for all inhibitory neurons.

largest at 200ms, measured from 120ms to 300ms. See Figure 6.8 for the read-out time of a single response.

These snapshots of the low-pass filtered spike-trains are concatenated in time. This yields a matrix with 4500 observations each consisting of 3249 channels (neurons). A small amount of Gaussian noise with a variance $\sigma = 0.01$ is added to the data. This is necessary to avoid numerical problems during SFA computations. As many neurons do not or only rarely respond to the stimulus some are highly correlated and cause numerical instabilites. The noisy observation matrices are then used as the input to the SFA algorithm. On the following pages the different performance tests are explained. For every single test, the SFA algorithm is applied to the output of a different processing stage. The main test was on the output of the V1 model. The second test only uses the output of the LGN model. Finally the raw image data was classified using SFA. These additional tests show how much performance can be gained by the V1 circuit as a non-linear pre-processing step. In addition to a single read-out time, the change of performance for ten different read-out times, taken from the output of the V1 circuit, for every stimulus was tested.

Single snapshot at time 200ms of the V1 output

This test is used to see how well the circuit works as a non-linear pre-processing step to SFA and if position-invariant object recognition can be achieved. In Figure 6.9 the results are shown. Already the first slow feature is able to distinguish between the two objects, '+' and 'x', with a mutual information of 0.94 *bit*. It filters out the only slowly varying information, the objects' identities, independent of their positions. All higher features do

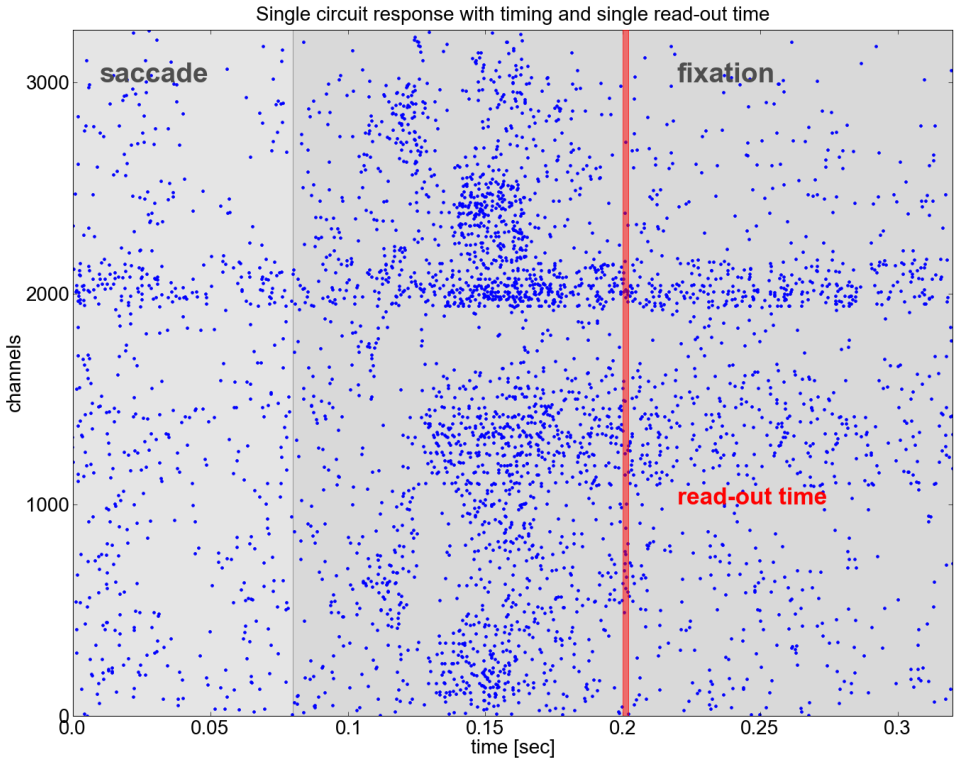


Figure 6.8: Spike response to a single stimulus. The grey shaded areas represent the saccade and fixation periods witch 80ms and 240ms respectively. At a time point of 200ms a snapshot is taken.

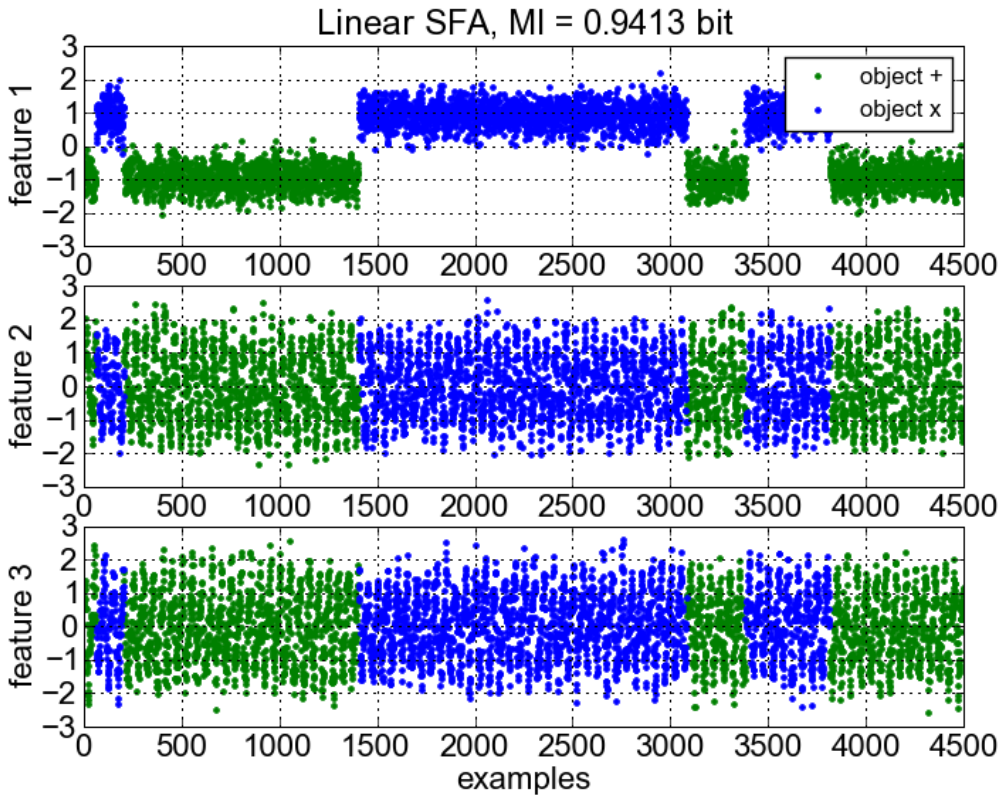
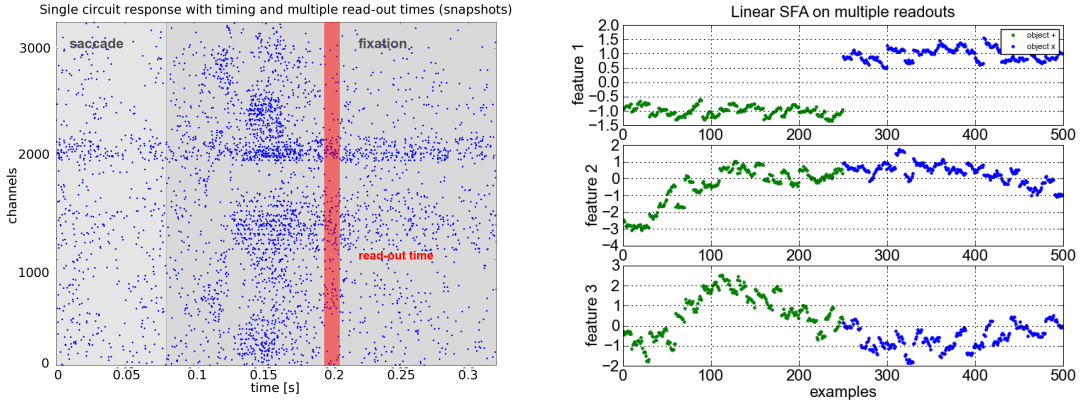


Figure 6.9: SFA results on a single snapshot from the V1 output of every stimulus. The slowest feature is shown in the top panel, the second slowest feature in the middel panel and the third feature in the bottom panel. Green points correspond to the SFA output on test data of object '+' and blue points to object 'x'.

not seem to carry any useful information for this task. The objects' identity is the only slow information the signals contain. It can be seen that SFA applied to a patch of V1 neurons is able to achieve invariant object representations and classifies the objects very well.

Multiple snapshots from time 195ms to 204ms of the V1 output

In addition to a single snapshot the SFA is applied to ten snapshots of every single stimulus output provided by the V1 model. First the snapshots are drawn randomly in an interval where the network has it's highest activity, but the result is pretty poor with only 68% of classification accuracy. Even more training data does not increase the performance. The results are much better when ten consecutive snapshots from 195 to 204ms are taken, see Figure 6.10b. Although the mutual information reaches 0.99 *bit*, the observation matrices now consist of 45000 observations for both training and test



(a) timing with multiple readouts

(b) SFA on multiple readouts

Figure 6.10: SFA on multiple snapshots of the V1 output. In a) one can see the timing of multiple snapshots and the response of a single stimulus. b) shows the result of SFA on multiple snapshots for every single stimulus of V1. Only 500 test samples are plotted.

data which increased the simulation time enormously. By making a trade-off between simulation time and accuracy, the results of a single snapshot taken out of the V1 output have to be favoured.

In Figure 6.11 a 2-dimensional and a 3-dimensional trajectory plot of the slow features one to three can be seen.

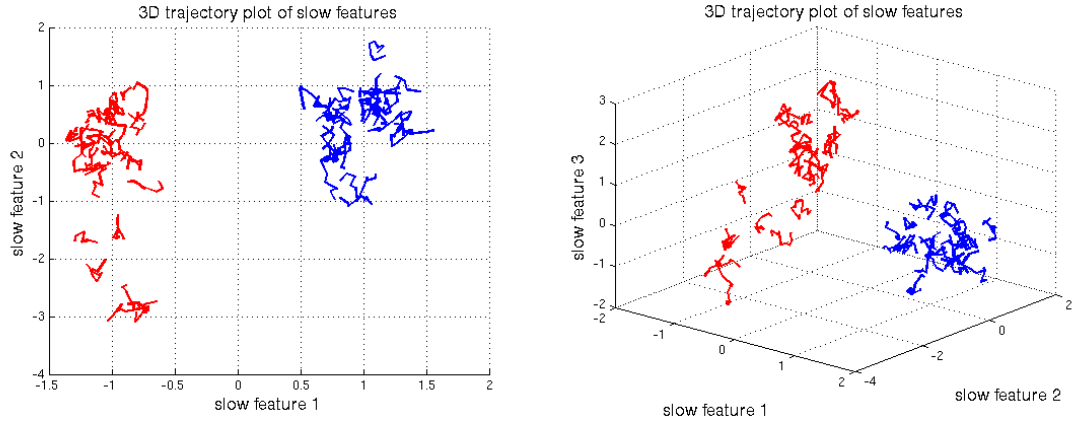
6.3 Additional tests

6.3.1 Single snapshot of the LGN model output

To test how important the non-linear expansion of the LGN output through the V1 model is, only the spiking output of the LGN has been low-pass filtered with a time constant of 30ms and used as input to the SFA algorithm. In Figure 6.12a an LGN response to a single stimulus can be seen, and Figure 6.12b shows the performance of SFA with a mutual information of 0.85 *bit*. Applying SFA to the output works, but the expansion with the V1 model is clearly better. The mutual information when using the V1 output is 0.94 *bit* and therefore increased by 0.09 *bit*.

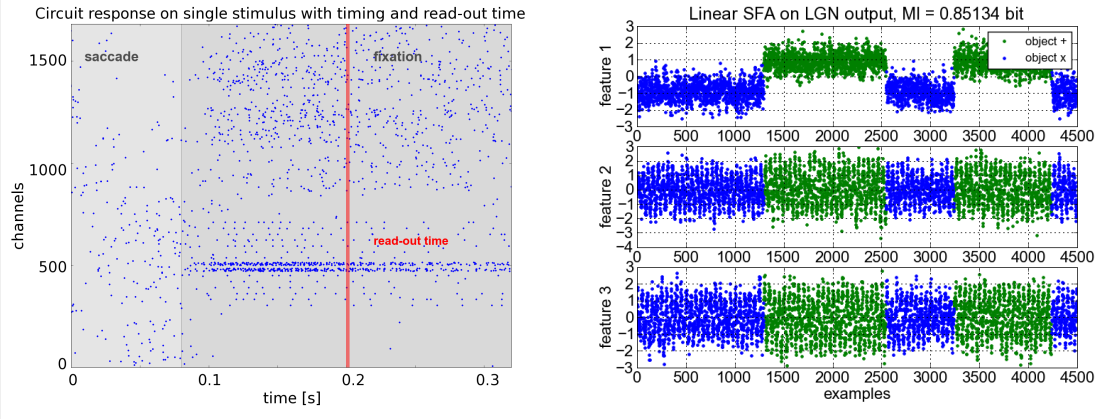
6.3.2 Raw image data

Applying SFA to a sequence of images was already done by Franzius et al. (2008) with a converging hierarchy of layers of SFA nodes. In contrast to Franzius et al. (2008) only one single SFA node is used to extract slowly varying information. In addition to linear



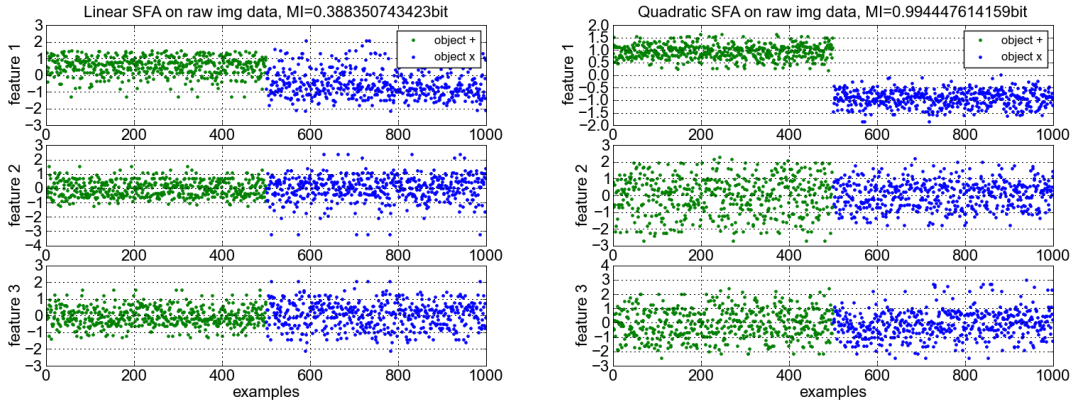
(a) 2D plot of slow features (b) 3D plot of slow features

Figure 6.11: 2-dimensional and 3-dimensional trajectory plot of slow features. In both cases one can see that the objects are absolutely separable. The 3D plot has been created from multiple readouts. I used 10 snapshots from every single stimulus coming from time point 195ms to 204ms. The stimulus duration was 320ms. Red points/lines correspond to object '+', blue ones to object 'x'.



(a) LGN response with single snapshot (b) SFA on LGN output

Figure 6.12: In a) one can see the spike train response of the LGN output on a single stimulus. In b) the results of SFA are shown.



(a) linear SFA on raw image data

(b) quadratic SFA on raw image data

Figure 6.13: SFA results on raw image data. In a) one can see the result of a linear SFA applied to raw image data. In b) the results of a quadratic SFA on raw image data can be seen.

Simulation	**MI, linear SFA**	**MI, quadratic SFA**
Single snapshot of V1	0.94 *bit*	–
Multiple snapshots of V1	0.99 *bit*	–
Single snapshot of LGN	0.84 *bit*	–
Raw image data	0.38 *bit*	0.99 *bit*

Table 6.1: Position-invariant object classification with SFA. The performance (measured by MI) of all combinations of pre-processing circuits and SFA methods used in this thesis are summarized.

SFA, I also tested a quadratic expansion. Before actually computing the SFA output, the images which are represented by a matrix, have to be transformed into a vector. This is done by concatenating every single row vector of the matrix to one large row vector. Additionally, Gaussian noise with a variance of $\sigma = 0.05$ is added to the input data and a principal component analysis is used to reduce the input space. Only the first 100 principal components are taken into account, because they contain nearly all of the variability of the data. The results can be seen in Figure 6.13. With a mutual information of only 0.38 *bit* for the linear SFA the result is quite bad. The quadratic expansion was much better and reached a mutual information of 0.99 *bit*.

It can be seen, that the V1 circuit as a non-linear preprocessing step gains a lot of performance compared to the result computed by linear SFA on raw image data.

In Table 6.1 the results of all different simulations are compiled.

6.4 Modelling the experiment of DiCarlo

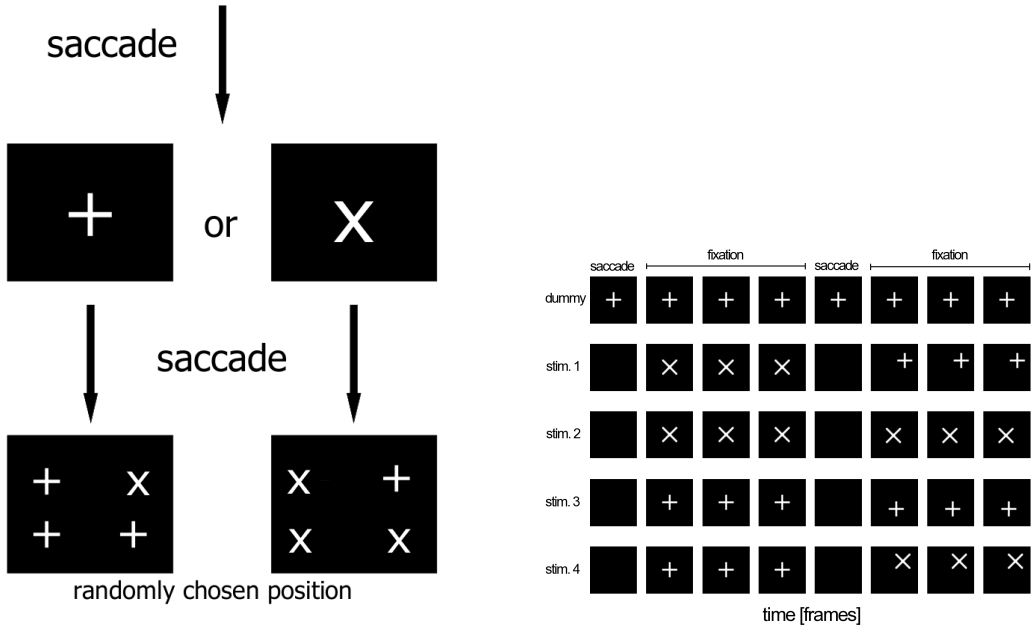
Li and DiCarlo (2008) provided evidence that unsupervised temporal slowness learning is substantial for building tolerant object representations. With targeted alteration of stimuli objects, monkeys experienced a new altered visual world. They present different objects to a fixated monkey and only if a specific object (the target) appears at a specific retinal position, the object was switched. The non-target objects never switched. This new world in turn caused drastic position tolerance changes of IT (inferior temporal cortex) neurons after only one hour of experience. A possible explanation is that contiguous retinal images mainly consist of images concerning to the same object. Task of this work is to show that these changes in position tolerance can be demonstrated by SFA and a detailled patch of V1 neurons.

6.4.1 Input generation

The input is generated in the same way as described in Section 6.1, but a single stimulus is defined different now. The stimulus now represents a saccade lasting for 80ms, followed by a fixation period with an object randomly drawn by the Markov model from Figure 6.3 that is always placed at the center of the retina. There is an additional saccade afterwards to a random position in the receptive field and a fixation period with another object. The first shown object will be switched to it's counterpart if the randomly chosen position is within the first quadrant (upper right corner) of the receptive field and the same object identity will be chosen otherwise, see Figure 6.14 for details. In the following I will refer to the different periods of a single stimulus in the following way:

- first saccade: S1
- first fixation period: F1
- second saccade: S2
- second fixation period: F2

If the setup in this thesis behaves in a similar way than the IT neurons from the experiment of Li and DiCarlo (2008), the SFA method is expected to classify objects that switch their identity during F2 as the same object that appears during F1. The algorithm learns these two different object representations as one single representation of the same object.



(a) Timing of switching objects

(b) Sample sequence of switching objects

Figure 6.14: In a) one can see the timing of a single stimuli when objects should switch identity according to their position. First, a saccade is made (80ms), then an object is chosen randomly according to the Markov model in Figure 6.3. A second saccade (80ms) is made followed by another object. This object is either the same as in the first fixation period, or the counterpart of it if it appears in the upper right corner. In b) one can see a sample sequence consisting of a dummy sequence in the first row and 4 subsequent stimuli. The dummy sequence is used to initially activate the neural circuit. Each row represents a stimulus consisting of two saccades and two fixation periods. It can be seen that the objects switch identity if they appear in the upper right corner during the second fixation period.

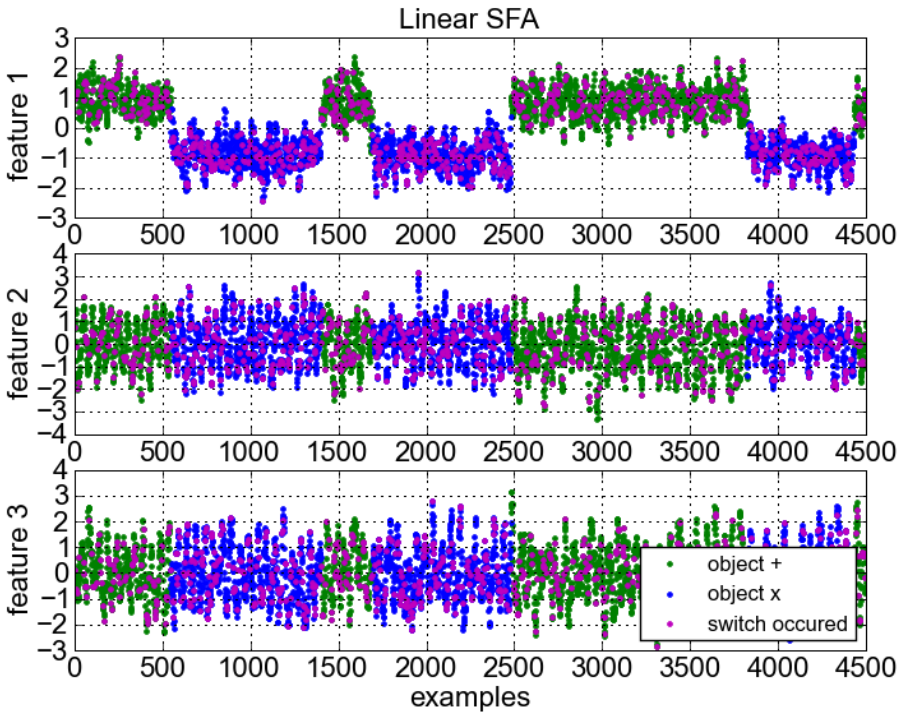


Figure 6.15: SFA on an altered visual world. The slowest feature again distinguishes between the two objects and clearly demonstrates that our simulations show the same behaviour as the experiment of Li and DiCarlo (2008). Green points represent object '+', blue points object 'x' and magenta points indicate snapshots where the object during F2 appeared in the upper right corner and switched it's identity.

6.4.2 Slow feature analysis on an altered visual world

The procedure of extracting the input for the SFA algorithm out of the V1 responses is the same as in section 6.1, except that the snapshot is taken at an appropriate time point out of the second fixation period (F2), where the objects' identities depend on their retinal position. The result of the SFA algorithm can be seen in Figure 6.15. Green points represent object '+', blue points object 'x' and points colored magenta represent samples where the object's identity has been switched due to it's retinal position in the upper right corner. This means that if a magenta point appears in the range of green points ('+'), the retinal position of the object during F2 was in the upper right corner and therefore switched identity to object 'x'. For magenta points in the range of blue points ('x') it is vice versa.

It can be seen that magenta points, where the objects switch identity, occur in the same amplitude distribution then their predecessors, see Table 6.2 for their mutual information.

Objects	Mutual Information
$+ \Rightarrow +$ and $+ \Rightarrow x$	0.011
$x \Rightarrow x$ and $x \Rightarrow +$	0.02

Table 6.2: The mutual information between two different objects is in general computed by the predictions of the two objects from the SFA algorithm and their real identities. For the position invariant case, the SFA algorithm should not distinguish between object identities that stay the same in F2 and objects that switch to their counterpart due to their retinal position in the right upper corner. In more detail this means that if a '+' remains or if it switches to a 'x' should not matter and the mutual information only computed of predictions for the case where the object's identity is a '+' in F1 is very low with 0.011 *bit*. The same result can be seen for object 'x' in the second row of the table. Here only sequences that start with a 'x' in F1 are taken into account and their mutual information is calculated using the predictions of the SFA algorithm and their real identities.

So if an object appears on the upper right corner of the retina (and switches identity) or somewhere else without a switch of identity doesn't really matter. The object will always be classified into the same class as it's predecessor.

I also simulated the case where the SFA algorithm is trained without any sequences where the object switch their identity. If a stimulus is then applied to the SFA algorithm including objects that switch their identity, the switched object of F2 should be classified as it's counterpart in F1. The algorithm should e.g. detect a plus in the upper right corner, which followed a cross from F1 as a plus and vice versa with the other object, see Table 6.3 for their mutual information. The MI for that case is computed by SFA that is trained on data where no switching occurs, but is tested on data that also includes sequences where objects switch their identity. Figure 6.16 illustrates an example using object '+' that either remain it's identity or switches to object 'x' if it appears in the upper right corner during F2. As the algorithm is only trained on objects which do not switch their identity, the 'x' during F2 is recognized as a 'x'.

Objects	Mutual Information
$+ \Rightarrow +$ and $+ \Rightarrow x$	0.88
$x \Rightarrow x$ and $x \Rightarrow +$	0.8853

Table 6.3: During the training phase of the SFA algorithm, no sequences occur where the objects switch their identity, regardless of their retinal position. Test data instead also includes sequences where the objects switch their identity according to their retinal positon. The mutual information is calculated in the same way as it is done for Table 6.2 however no position invariance should occur now. The mutual information between the objects is now very high.

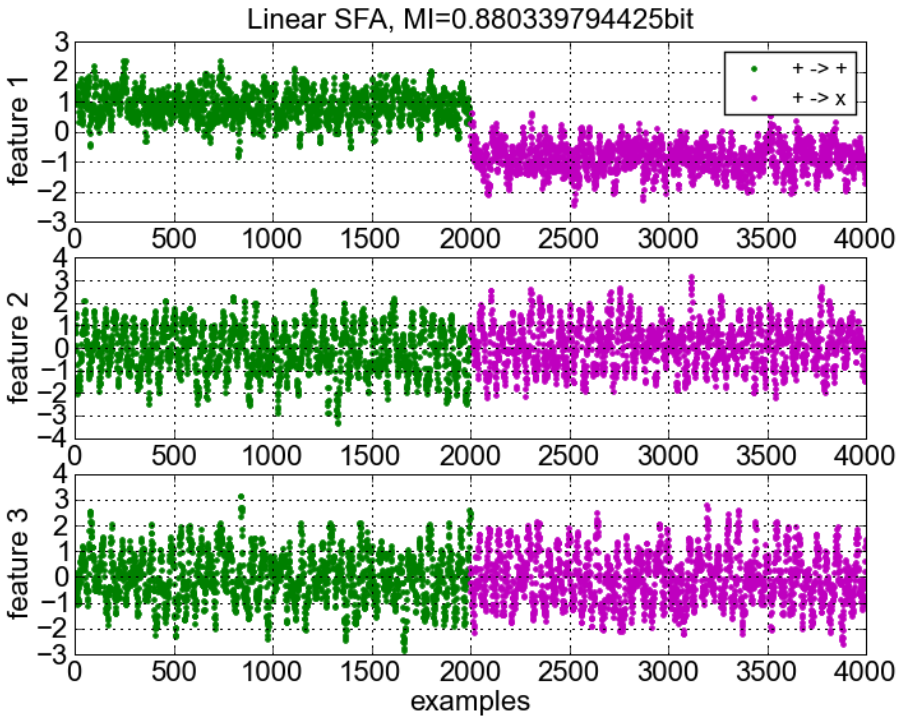


Figure 6.16: If the training data does not include sequences that switch due to their retinal position, but testing data does, the algorithm distinguishes between these objects. An object that switches identity if it appears in the upper right corner will simply be realized as it's counterpart. Green points represent the SFA output of stimuli consisting of object '+'. Magenta points represent outputs when a '+' switched to a 'x' during F2

Chapter 7

Discussion and Outlook

The experiments clearly showed that slow feature analysis is a powerful method to extract position invariant features from a patch of V1 neurons, stimulated by a sequence of two different objects alternating in an adequate manner. The cortical microcircuit turned out to be a promising non-linear preprocessing step to transform the data to a linearly separable space. The SFA algorithm discriminated the two objects with a high accuracy of around 0.94 *bit*. By using additional models of the retina and the LGN the pre-processing of information in our model is biologically quite realistic. In contrast to Franzius et al. (2008) there is no need for an additional classification procedure on top of the SFA algorithm. They used linear regression and two different classifiers to show the classification ability of their model. In this thesis, the output of the SFA algorithm itself provides enough information to distinguish between the different objects. The main advantage of our model is, that it is biologically quite plausible. The circuit is built on experimental data from Thomson et al. (2002) and represents a highly detailled patch of V1 neurons. The spatial and temporal pre-processing by the retina and LGN model strengthen the plausibility.

Although the classification performance is very high, it may still be possible to improve it. For instance with the use of orientation maps. Schuch et al. (2009) used Kohonen's Self-Organizing Map algorithm (Kohonen, 1982) to implement orientation maps across the cortical surface. It is well known from Hubel and Wiesel (1977) that orientation preferences form maps in cats, where neighbouring neurons tend to respond to similar features. Especially for the case of the two objects '+' and 'x', which only differ in their orientation, the use of such maps could increase the performance.

Another possibility is to use additional circuitry that receives input from responses of excitatory neurons from layer 2/3 of the V1 circuit with convergent connectivity. This means that the size of the receptive field of the second circuit represents a larger visual field. This additional circuit together with the V1 model can be seen as a hierarchical

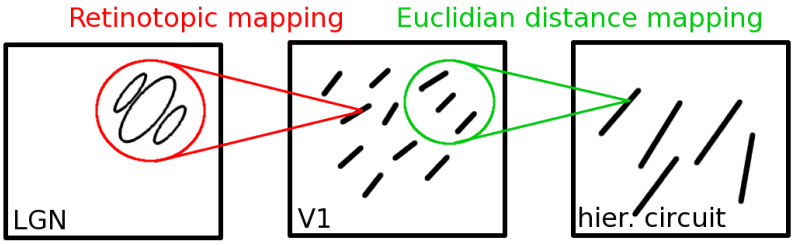


Figure 7.1: Schematic of how a hierarchically expanded circuit may look like. The retinotopical mapping in V1 will be extended by a euclidian distance mapping to the second stage.

structure, see Figure 7.1. The positions of the neurons inside the layer remain the same. The information they receive can be mapped by a euclidian distance to surrounding neurons. This means that a neuron from the additional circuit at a specific position, receives information from all neurons from the V1 circuit that lie inside a specified radius of the currently considered neuron. The connection strength for each neuron is chosen proportional to the inverse of the distance to the currently considered neuron. Neurons that are nearer have more influence than neurons that lie farther away. The SFA algorithm will only be applied to the output of the second circuit. By tuning the circuit's parameter appropriately this hierarchical structure could increase performance.

Bibliography

Bear, M. F., Connors, B. W., and Paradiso, M. A. (1996). *Neuroscience: Exploring the Brain*. Williams & Wilkins.

Berkes, P. and Zito, T. (2005). Modular toolkit for data processing (version 2.0). http://mdp-toolkit.sourceforge.net.

Biederman, I. and Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology*, 18(1):121 – 133.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues.* Barth, Leipzig.

Croner, L. J. and Kaplan, E. (1995). Receptive fields of p and m ganglion cells across the primate retina. *Vision Res*, 35(1):7–24.

Destexhe, A., Rudolph, M., Fellous, J. M., and Sejnowski, T. J. (2001). Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons. *Neuroscience*, 107(1):13–24.

Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction and spatial-view cells. *Public Library of Science (PLoS) Computational Biology*, 3(8):166.

Franzius, M., Wilbert, N., and Wiskott, L. (2008). Invariant object recognition with slow feature analysis. In *ICANN '08: Proceedings of the 18th international conference on Artificial Neural Networks, Part I*, pages 961–970, Berlin, Heidelberg. Springer-Verlag.

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

Fukushima, K. (1987). Neural network model for selective attention in visual pattern recognition and associative recall. *Applied Optics*, 26(23):4985–4992.

Furmanski, C. and Engel, S. (2000). Perceptual learning in object recognition: object specificity and size invariance. *Vision Res.*, 40(5):473 – 484.

Gazeres, N., Borg-Graham, L. J., and Fregnac, Y. (1998). A phenomenological model of visually evoked spike trains in cat geniculate nonlagged X-cells. *Vis Neurosci*, 15(6):1157–1174.

Harris, C. M., Wallman, J., and Scudder, C. A. (1990). Fourier analysis of saccades in monkeys and humans. *J Neurophysiol.*

Häusler, S. and Maass, W. (2007). A statistical analysis of information processing properties of lamina-specific cortical microcircuit models. *Cerebral Cortex*, 17(1):149–162.

Hubel, D. (1987). *David Hubels Eye, Brain and Vision.* Scientific American Library.

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phys.*, 160:106–154.

Hubel, D. and Wiesel, T. (1977). Ferrier lecture. Functional architecture of macaque monkey visual cortex. *Proc R Soc Lond B Biol Sci*, 198(1130):1–59.

Ince, R. A. A., Petersen, R. S., Swan, D. C., and Panzeri, S. (2009). Python for information theoretic analysis of neural data. *Front Neuroinformatics*, 3(4).

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J Neurophysiol.*, 73(1):218 – 226.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Trans Neural Netw*, 14(6):1569–1572.

Klampfl, S. and Maass, W. (2009). A theoretical basis for emergent pattern discrimination in neural systems through slow feature extraction. *Submitted for publication.*

Klampfl, S. and Maass, W. (2010). Replacing supervised classification learning by Slow Feature Analysis in spiking neural networks. In *Proc. of NIPS 2009: Advances in Neural Information Processing Systems*, volume 22. MIT Press.

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

Kolb, H., Fernandez, E., and Nelson, R. (1996). Webvision the Organization of the Retina and Visual System. http://maxlab.neuro.georgetown.edu/index.html.

Li, N. and DiCarlo, J. J. (2008). Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. *Science*, 321:1502–1507.

Logothetis, N. K. and Steinberg, D. L. (1996). Visual object recognition. *Annu Rev Neurosci.*, 19:577–621.

Lund, J. S., Angelucci, A., and Bressloff, P. C. (2003). Anatomical structures for functional columns in macaque monkey primary visual cortex. *Cereb Cortex*, 13(1):15–24.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.*, 14(11):2531–60.

Morrone, M. C., Ma-Wyatt, A., and Ross, J. (2005). Saccadic eye movements cause compression of time as well as space. *J Vis*, 5:741–754.

Oram, M. W. and Perrett, D. I. (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7:945 – 972.

Purves, D., Augustine, G. J., Fitzpatrick, D., C.Hall, W., LaMantia, A., McNamara, J. O., and White, L. E. (2008). *Neuroscience*. Sinauer Associates, Inc.

Riesenhuber, M. and Poggio, T. (1999,). Hierarchical models of object recognition in cortex. *Nat. Neurosci.*

Schuch, K., Rasch, M., and Maass, W. (2009). Statistical characterization of the spike response to natural stimuli in monkey area v1 and in a detailed laminar model for a patch of v1.

Shannon, C. E. and Weaver, W. (1949). *The mathematical theory of communication.* University of Illinois Press.

Thomson, A. M., West, D. C., Wang, Y., and Bannister, A. P. (2002). Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2 - 5 of adult rat and cat neocortex: triple intracellular recordings and biocytin labelling in vitro. *Cerebral Cortex*, 12(9):936–953.

Tovee, M. J., ET, E. T. R., and Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *J Neurophysiol.*, 72(3):1049 – 1060.

Wiskott, L. and Berkes, P. (2003). Is slowness a learning principle of the visual cortex? *Zoology*, 106(4):373 – 382.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14:715–770.