Bio-inspired landmark recognition models

ADVANCED SEMINAR

submitted by
Stefan Urban

NEUROSCIENTIFIC SYSTEM THEORY

Technische Universität München

Supervisor: Marcello Mulas, PhD
Final Submission: 18.01.2016

Technische Universität München
Neurowissenschaftliche Systemtheorie

Advanced Seminar

Stefan Urban 3664589

Bio-inspired landmark recognition models

07-Oct-2015

Problem description:

Computer vision aims at the development of algorithms that can match human visual skills. It is essential for an autonomous robotic system and it is usually based on engineered solutions. However, a promising strategy to achieve high level performance is to mimic the way the human brain processes visual information [1,2]. In this advanced seminar the student shall review working models of the visual system with particular emphasis on those specifically developed for robotic applications.

Task:

More precisely, the student shall:

- Describe what is known about how the brain processes visual information
- Review the scientific literature about bio-inspired algorithms for feature extraction and landmark recognition
- Evaluate the performance and the biological plausibility of each algorithm
- Discuss the main advantages and disadvantages in comparison with state-of-the-art algorithms

Bibliography:

[1] Kerr, D.; McGinnity T.M.; Coleman S.; Clogenson M.: A biologically inspired spiking model of visual processing for image feature detection. Neurocomputing 158, 2015

[2] Siagian C. and Itti L.: Biologically inspired mobile robot vision localization, IEEE Trans. Robotics, vol. 25, no. 4, pp. 861-873, Aug. 2009.

Supervisor: Marcello Mulas, PhD

(Jörg Conradt)
Professor

Abstract

Human vision still outperforms the best object recognition algorithms there are. It is believed that mimicking the retinal and visual cortex system will help with building better performing models. For this, first the biological foundation is analyzed and then transfered to the basic HMAX model. After that Serre et al.'s extension will lead to a pretty good representation of the first layers in the visual cortex. Finally we will discuss the application requirements for self-localization.

Contents

1 Introduction **2**

2 Human Visual System **3**
2.1 Retina 3
2.2 Optic Chiasm and Lateral Geniculate Nucleus 5
2.3 Visual Cortex 6

3 Biologically Inspired Computational Models of the Visual System **9**
3.1 Basics 9
3.1.1 Signal Propagation 9
3.1.2 Simple and Complex Cells 10
3.2 Feed-Forward Models 11
3.2.1 HMAX model for object recognition 11
3.2.2 Extention of HMAX with Template Matching 12
3.3 Biological plausibility 14

4 Application: Localization **16**

5 Conclusion **17**

List of Figures **18**

Bibliography **19**

Chapter 1

Introduction

One of the main topics in computer vision is object recognition. Despite many years of research that brought us all kinds of artificial algorithms, very basic problems remain nearly unsolvable.
An example would be this simple decision task: Is there a bird in an image? Humans can answer this after seeing a picture for only very short duration although they never saw the image before. More surprisingly it also does not matter if a person knows that exact type of bird or if he had seen it before. Viewing positions and angles, photometric effects, scene settings and changing body shapes have little effect on the overall test performance.
It is obvious this is the product of a long evolutionary process that lead to these high efficient object recognition system. So if it works this well, the naive solution is to design new algorithms the same way the human visual system works.
To gain an understanding of what these new models should achieve, we are starting to look at the physiology in chapter 2. After covering the structure of the retina and the information transport to the brain, the complexity of the visual cortex will be addressed.
Chapter 3 starts off with some background about what kind of model will be used. The basis for all of them will be the results of Hubel and Wiesel's research that is shortly explained and then used to implement the HMAX model by Riesenhuber and Poggio. Serre et al. refined this model and also discussed the biological plausibilities.
As the main task of this seminar is to find a biological inspired way for self-localization, chapter 4 comes back to that and links the acquired models to that application. It is basically object recognition, but with some boundary conditions and additional work.

Chapter 2

Human Visual System

Until today the human vision is the best-known system for recognizing objects and actions in a scene in a very short time. The basis to understand how it works lies in the anatomy and physiology. Therefore all the parts that belong to the visual cortex are described here.

2.1 Retina

The retina is the only source for visual stimuli in mammals. This is an excellent prerequisite to study neural responses, because all input can easily be controlled. Figure 2.1 shows a cross section through a human eye. Light enters through the lens and falls onto the retina on the other side, where it is transfered into neuronal signals. The foeva is the part of the retina with the highest density of cells and is the area with which humans most consciously see.

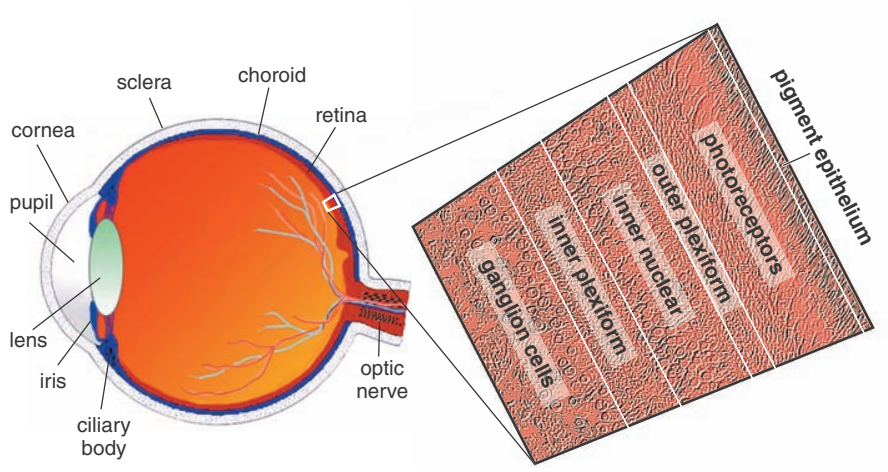

Figure 2.1: Cross section of the human eye [Kolb, 2003]

There is already processing happening in the layers of the retina before the infor-

mation goes onto the optic nerve. The photo receptive cells, the rods and cones, generate an analog signal in the outer plexiform layer. This is the input of the bipolar cells that transport the signal to the inner plexiform layer. With interaction between bipolar and horizontal cells, the visual system is able to adapt to different illuminations. In the inner plexiform layer the dentrites of ganglion cells and amacrines combine the data from different bipolar cells before it get converted to action potentials and sent out of the retina.
Overall the information is encoded into a string of about 1 million ganglion axons that are called the optic nerve. It transports the information to the optic chiasm.

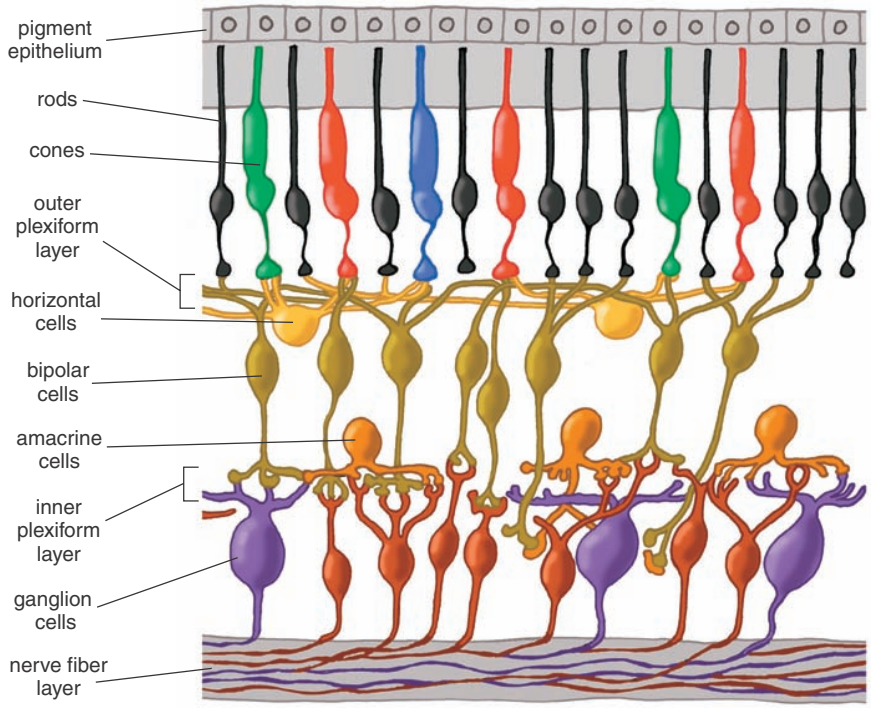

Figure 2.2: Cross section of the retina [Kolb, 2003]

So what is the output of the retina? Ganglion cells are overall connected to many rods and cones each, but their specific function differs drastically. Mostly there are cells that fire when the center of their receptive field is illuminated and the surroundings are not, respectively the other way around (OFF- and ON-center cells). Additionally there are ganglion cells that are responsible for special low level functions. For example the focusing of the eye or the signal inhibition when the eye or head moves and would overflow the cortex with neural information (object motion sensitive cells).

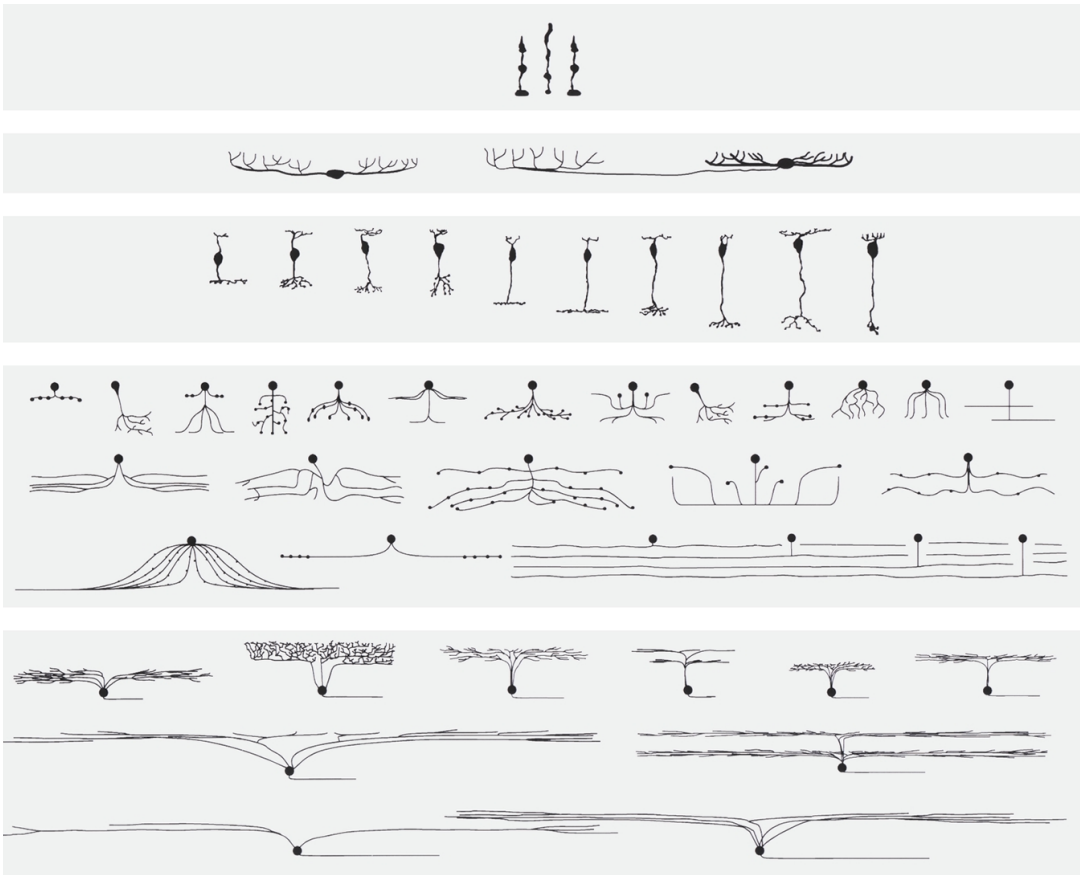

Figure 2.3: Different types of neuroes in the layers of the retina (from top): photo receptors, horizontal cells, bipolar cells, amacrine cells and ganglion cells [Masland, 2001]

2.2 Optic Chiasm and Lateral Geniculate Nucleus

In the optic chiasm the signals on the optic nerve are separated into the left and right half frame of the eye. Then the same side signals from both eyes are sent together to the lateral geniculate nucleus (LGN).

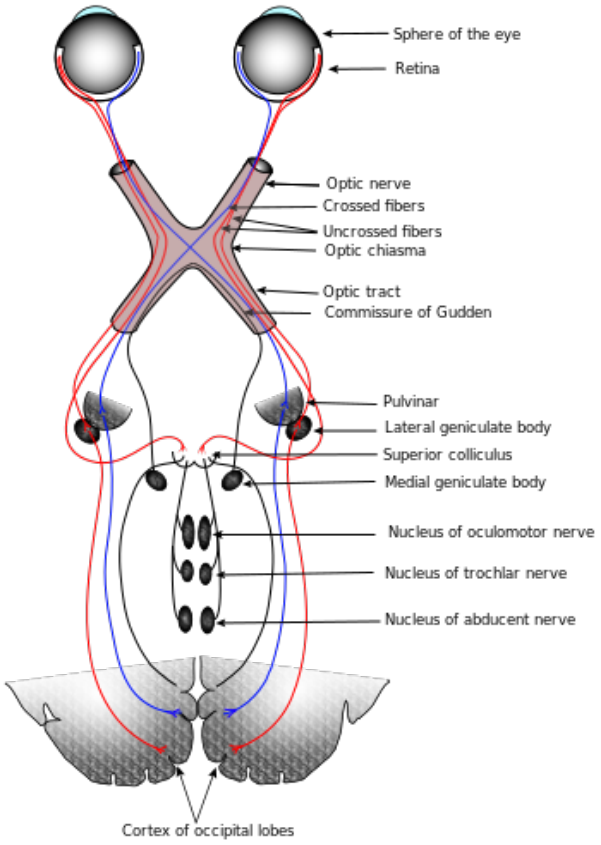

Figure 2.4: Overview of the signal flow from the eyes (top) to the visual cortex (bottom), blue lines indicate the nasal side of retinal signals, red the temporal ones (Source: wikipedia.org)

The major processes that happen in the LGN have to do with combining the signals from the left and right eye (3D-representation of scene). This has little effect on the ability to recognize objects simply because this also works with only one eye.

2.3 Visual Cortex

The human visual cortex is the biggest single part of the brain. This emphasizes the major role vision has for humans. In about 20 distinct areas the incoming signals from the retina are processed in a mostly hierarchically manner.
Many different regions were identified to perform distinct tasks. Figure 2.5 is an example of a map of the connections between these areas. Most of them are bidirectional.

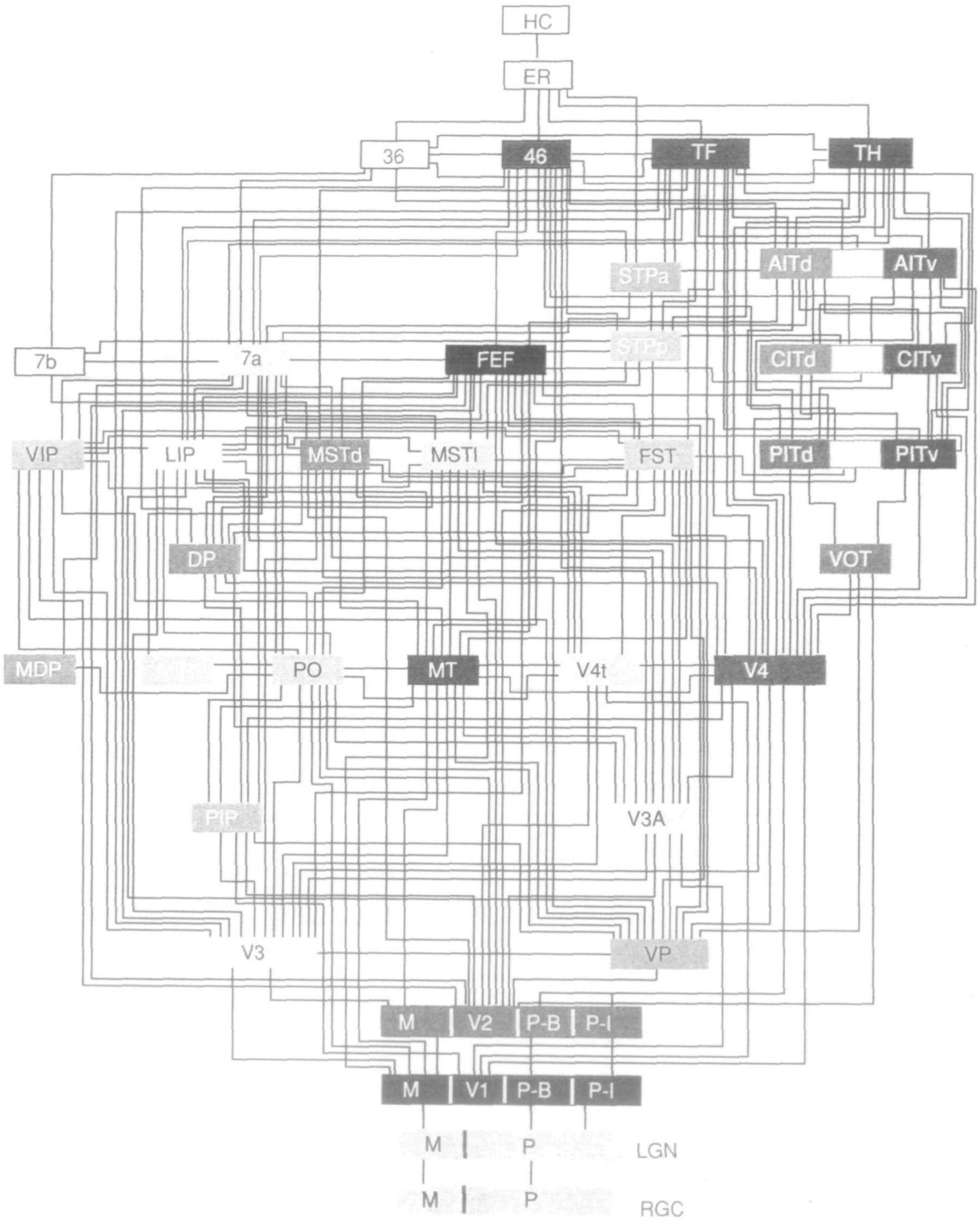

Figure 2.5: Visual cortex map from retina (bottom) to hippocampus (top), most connection are reciprocal [Felleman and Van Essen, 1991]

Typically the visual cortex is separated in multiple major areas which are: the primary visual cortex V1 or the striate cortex and the extra-striate areas which are

V2 to V5. The information processing is divided into two pathways: The dorsal visual pathway deals with calculating the location of objects and is the connection of the visual cortex with the pariental lobe. The more important one is the ventral visual pathway that conducts the task of what the eyes see and leads to the temporal lobe.

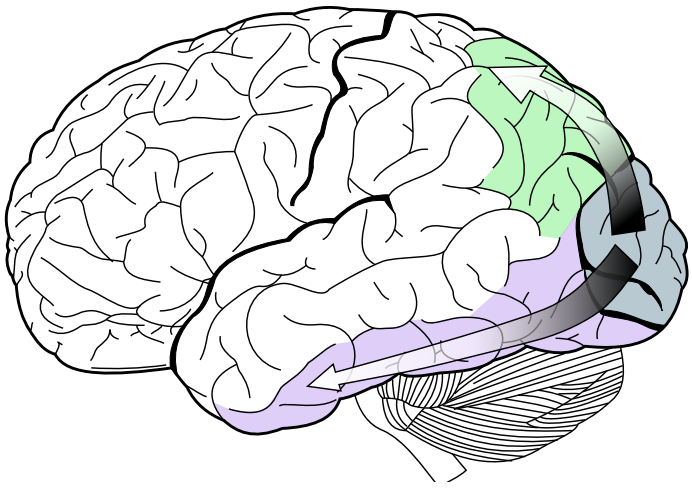

Figure 2.6: Ventral (purple) stream going to temporal lobe and dorsal (green) stream heading to parietal lobe (Source: scholarpedia.org)

In the first layer V1, the primary visual cortex, each neuron only responds to very little areas of the visual field (= receptive field of neuron). Basic information about the orientation and movement of objects are obtained. Beginning with the V2 the signals are combined for invariance over the whole visual field. From then on the ventral stream mainly goes along the regions V4 and ends up in the inferior temporal cortex (IT). In there much the high-level processing happens. With the connection to the memory areas of the brain, here arise stimuli like recognized objects and faces. As the information over more and more neurons are combined in higher visual cortex areas, each neuron represents the content of a larger receptive field. At the end invariance of many interferences is achieved, which means an object can be recognized no matter where for example on the input image it is, or what it's rotation is.
A model to explain this was developed by Hubel and Wiesel with the use of so called simple and complex cells. They are described as part of the following chapter.

Chapter 3

Biologically Inspired Computational Models of the Visual System

Building a model for the visual system will be a compromise. Of course one can start to model every single neuron from the retina to the visual cortex, but this is very unpractical. So the typical step would be to abstract a certain part of the system to a functional description. There are basically two ways to do this: Abstraction top-down and bottom-up.
Top-down means goal-driven, so you could build functional models that abstract viewing tasks, expectations or reasoning. Obviously this is very complicated, therefore bottom-up models, which are stimulus driven, are the way to go.
To realize that kind of model, first of all the signal propagation in the visual system is discussed. Then, with the research of Hubel and Wiesel as basis, the HMAX model will be explained. In [Serre et al., 2007] it will be extended to better fit the biological foundation.

3.1 Basics

3.1.1 Signal Propagation

The only connection of the retina is the optic nerve. It is about 1 million axons transmitting information in one direction, hence only feed-forward.
Each additional layer of the visual path has back some sort of feedback. As mentioned in chapter 2, almost every connection in figure 2.5 is bidirectional. This makes modeling extremely difficult. Thorpe et al. found out that, even every cortex area of the visual system has some sort of higher function level feedback, in the first 150 milliseconds it is not active. Therefore for this short time one can assume a strict feed-forward network. So, despite the fact many features are missing, most models limit themselves to feed-forward.

3.1.2 Simple and Complex Cells

The basis for the following model described in section 3.2.1 is an achievement by Nobel Prize winners Hubel and Wiesel. They discovered two important types of neurons in the visual cortex: simple and complex cells.
The simple cells respond to light or darkness in a certain area of the visual field, that is called receptive field. They combine multiple ON- of OFF-cells coming from the ganglion layer in the retina as one can see in figure 3.1. Thereby orientation selective neuronal signals in V1 can be modeled.

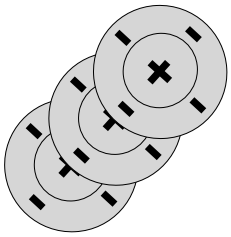

Figure 3.1: Overlap over multiple ON-center ganglion cells that is performed by a simple cell (Source: wikipedia.org)

Complex cells on the other hand combine the output of multiple simple cells and are able to fire when detecting edges or even motion in a specific direction. Figure 3.2 shows a experiment that was performed using a cat as test subject. An electrode was injected into a complex cell in the visual cortex. If the bright white bar was moved in the direction of the drawn arrow, the neuron would fire. If the bar stops or moves the other direction, it would not.

Figure 3.2: Experiment performed by Hubel and Wiesel: Complex cell of a cat's visual cortex, the neuron fires only if a short bar is moved in the arrows direction [Source: youtube.com]

In [Hubel and Wiesel, 1965] the two scientists extended their model with the introduction of so called hypercomplex cells. These are basically cells with a complexity

beyond the normal complex cells and can be found in higher cortical areas.

3.2 Feed-Forward Models

3.2.1 HMAX model for object recognition

The research of Hubel and Wiesel was not originally intended for object recognition. It was extended by Riesenhuber and Poggio to form a hierarchical model that uses simple and complex cells as building blocks: the HMAX model. As mentioned in 3.1.1, as most models this one is strictly feed-forward only.

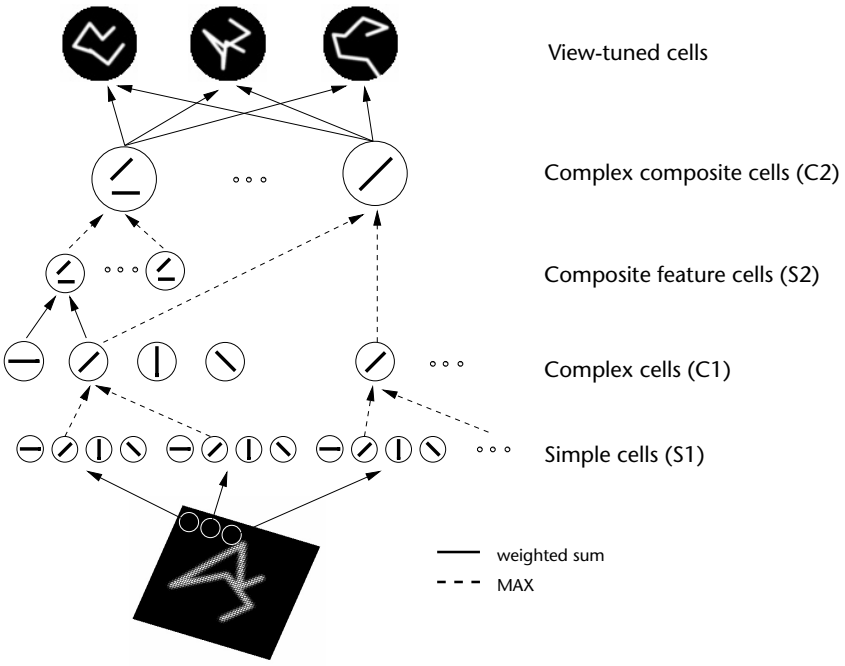

Figure 3.3: HMAX model by [Riesenhuber and Poggio, 1999]

As one can see in figure 3.3, the model consists of 5 layers. The first four layers are alternating simple and complex cells. The first stage features a lattice of blocks that are sensitive to orientation in their respective receptive field. The second layer creates position invariance as it pools over simple cells with the same orientation angle. At S2 these complex cell outputs are composed to selective features such as edges or borders. After another layer that increases invariance over other various transformations, view-tuned cells can describe abstract high-level features like the presence of a specific object. Scale invariance is not shown in this figure, but it is achieved by utilizing different scales of first layer of simple cells and involvement in the subsequent complex pooling operation.

These pooling operations performed by the complex cells are implemented using a nonlinear maximum. This tends to be better than the summation as alternative because it disregards clutter and interferences.
For object identification, supervised learning of the HMAX model works like this: An image of the object is inserted into the model. The response of the second complex cell layer (C2) is recorded and turned into an appropriate view-tuned cell. But there is a need for learning of objects with many pictures differing in perspective, illuminance, position and other transformations. This easily leads to a vast amount of example images for one object.
The two main problems when developing a model for the visual cortex are resolved pretty good. HMAX can provide a high level of invariance in scale, position, rotation and other transformation of the object by simply pooling over the respective cell outputs. Also the selectivity is very high because even on higher layers very sharp edge and border information is provided.

Performance Tests showed that the HMAX model performs very good with natural image sets like Caltech101. But [Pinto et al., 2008] argues these images are not at all representative for the task, this model was developed. They feature objects from many different viewing angles in various lightning conditions which is good. This is also stated by the Caltech101 website: "Most images have little or no clutter. The objects tend to be centered in each image."[1] On real-world image variation, that very well include many interferences and will be the main application environment, the performance is dropping drastically.

3.2.2 Extention of HMAX with Template Matching

A few years later Serre et al. extended the HMAX model with template matching functionality which is more closer the biological reality in the inferotemporal cortex ([Poggio and Bizzi, 2004]).
The basic assumptions are the same as for the standard HMAX model: the processing is hierarchal (scale and position first, then other transformations), feed-forward processing only and the receptive field is increasing the higher layer a neuron is. But one major thing changes now: Learning is assumed to happen almost at every layer which has to be realized as well as possible.

[1] `http://www.vision.caltech.edu/Image_Datasets/Caltech101`

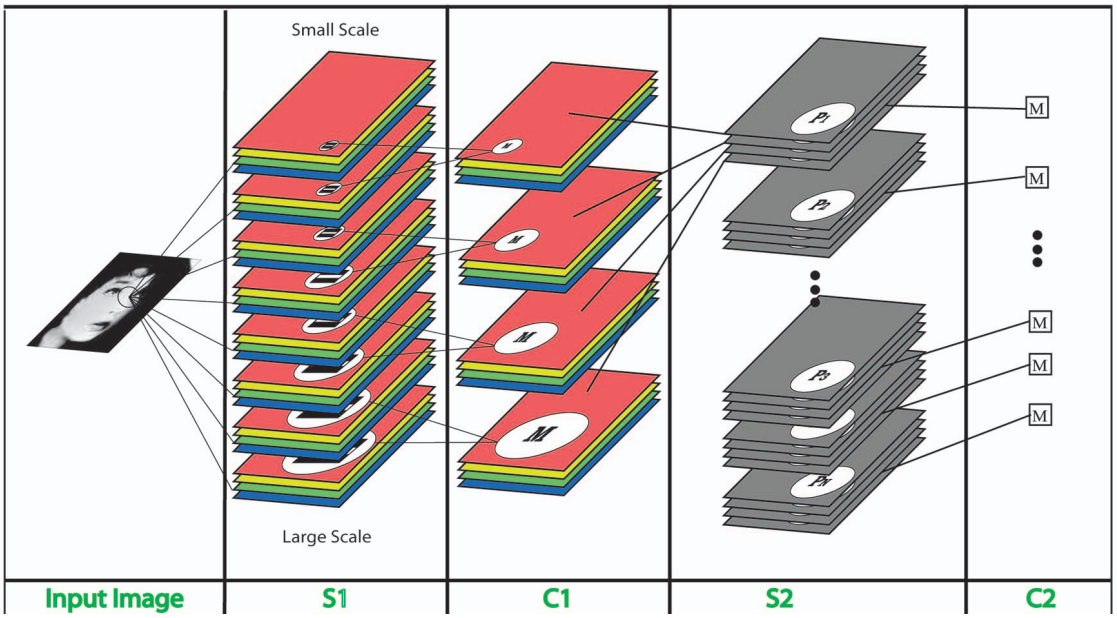

Figure 3.4: System overview of the model, only 8 scales shown (Source: [Serre et al., 2007])

The figure 3.4 shows the proposed model with it's four stages. In the first one, S_1, for an incoming gray-scale image, multiple Gabor filters (16 scales and 4 orientations) are applied:

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

mit

$$x_0 = x\cos\Theta + y\sin\Theta$$
$$y_0 = -x\sin\Theta + y\cos\Theta$$
$$\Theta = 0°; 45°; 90°; 135°$$

The exact parameters can be looked up in the original paper: [Serre et al., 2007].

These filters create 64 different output images. In the next stage, C_1, the complex cells provides the local maximum over position, thus providing shift invariance. During this operation always two of the 16 scales are combined, leaving a total of 32 images (8 sets with 4 different filter orientations).
These are compared with a vast amount of trained feature patches in S_2. This is the template matching that is new to the model. The determination of the training patches is simple: A target set of pictures of objects that should be recognized is acting as input of the model. The feature values are then simply taken from the C_1 output. Later every new image will be compared to all available trained feature values with a radial bias function (RBF):

$$r = e^{-\beta||\mathbf{X}-\mathbf{P}_i||^2}$$

The higher the value r, the better the correspondence to the trained feature.
C_2 finally pools over all S_2 outputs of the same feature patch and saves only the maximum. This leaves a vector with the length of the trained patches.

Performance A major limitation could be the S_1 and C_1 stages that do need much computational effort. But the overall performance shown in , especially in comparison with SIFT, proves it being worth it (see figure 3.5).

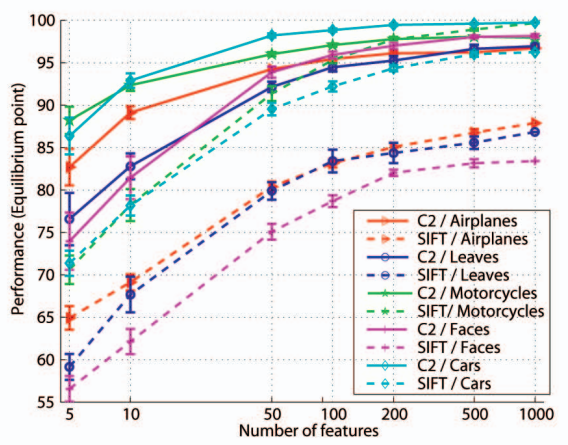

Figure 3.5: Performance of model proposed by [Serre et al., 2007]

All performance tests in the paper were conducted while not tuning one single parameter. These were determined using data about the primary visual cortex. Improvements are very well possible.
As opposed to having a broad categorization performance, specializing on object identification is easier. Reference images have to be learned as explained above. In the process of recognition, the radial basis function will provide a high value just because it is not a feature but more like a specific object patch.

3.3 Biological plausibility

Simple and Complex Cells Many parts of the HMAX model were designed with the experiments of Hubel and Wiesel in mind. These only represent a very basic computational part in the visual cortex. Nevertheless there is the necessity , that it is possible to realize the two main building blocks of the model, the simple and complex cells, in neuronal circuits. If that would not be possible, further investigations based these would be pointless. This was studied in [Serre et al., 2005] with a positive result.

No Feedback One of the main problems the presented models have is the strict limitation to feed-forward networks. As outlined in section 2.3, this is absolutely not true. There are several attempts to achieve this but they mostly have major deficits. This is understandable when looking at a functional map of the visual cortex (figure 2.5). The connections reach up to the hippocampus and are even influenced by cognitive behavior. Also it is arguable if integration of higher brain functions will lead to better object recognition performance.

Learning While making it possible to reduce the amount of images needed for training of object recognition models, humans still are way better. They normally can recognize an object with only one single example as trained reference.

Chapter 4

Application: Localization

One of many possible applications for object recognition in general is self-localization. Therefore stationary objects, called landmarks, have to be recognized and associated with position information.
A simple example: A robot drives through a hallway. There are many different object to recognize that he knows, like a door. But usually there is not only a single door but many different. If recognizing more objects in the surrounding area, like windows, plants or hallway junctions, the possible locations for this kind of specific occurrences is limited. This is normally implemented using a particle filter that is for example explained in [Siagian and Itti, 2009].
But what requirements does this have for the object recognition algorithm? First of all the computational effort should be kept within reasonable limits. Normally the self-localization is not the primary task when there is the need for it. Secondly the image of the camera does not focus automatically onto specific separated objects. There will always be much clutter in the model inputs. One way to overcome this issue is more extensive training. But this is in conflict with the third requirement for object recognition algorithms: Ideally it does not need much object examples to perform well. It is highly unpractical if for indoor navigation every object in each single room has to be trained with hundreds of images.

Chapter 5

Conclusion

In summary the rapid scene analysis is well understood till today, but the real visual system is far more complex then that. Future work has to extend these using more layers (S3, C3) or a basically different feedback approach.
Current algorithms (like in 3.2.2) do a good job when modeling the physiological basis up to the inferior temporal cortex. But as one can so on the map of the visual cortex (figure 2.5) there is still much work to do.

List of Figures

2.1 Cross section of the human eye [Kolb, 2003] 3
2.2 Cross section of the retina [Kolb, 2003] 4
2.3 Different types of neuroes in the layers of the retina (from top): photo receptors, horizontal cells, bipolar cells, amacrine cells and ganglion cells [Masland, 2001] 5
2.4 Overview of the signal flow from the eyes (top) to the visual cortex (bottom), blue lines indicate the nasal side of retinal signals, red the temporal ones (Source: wikipedia.org) 6
2.5 Visual cortex map from retina (bottom) to hippocampus (top), most connection are reciprocal [Felleman and Van Essen, 1991] 7
2.6 Ventral (purple) stream going to temporal lobe and dorsal (green) stream heading to parietal lobe (Source: scholarpedia.org) 8

3.1 Overlap over multiple ON-center ganglion cells that is performed by a simple cell (Source: wikipedia.org) 10
3.2 Experiment performed by Hubel and Wiesel: Complex cell of a cat's visual cortex, the neuron fires only if a short bar is moved in the arrows direction [Source: youtube.com] 10
3.3 HMAX model by [Riesenhuber and Poggio, 1999] 11
3.4 System overview of the model, only 8 scales shown (Source: [Serre et al., 2007]) 13
3.5 Performance of model proposed by [Serre et al., 2007] 14

Bibliography

Jorg Conradt, Pascal Simon, Michael Pescatore, Paul FMJ Verschure. Saliency maps operating on stereo images detect landmarks and their distance, *International Conference on Artificial Neural Networks (ICANN)*, p. 795-800, 2002.

Jorg Conradt, Gaurav Tevatia, Sethu Vijayakumar, Stefan Schaal, On-line learning for humanoid robot systems, *International Conference on Machine Learning*, Stanford, p. 191-198, 2000.

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47, 1991.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160 (1):106, 1962.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*, 28(2):229–289, 1965.

Helga Kolb. How the retina works. *American scientist*, 91(1):28–35, 2003.

Richard H Masland. The fundamental plan of the retina. *Nature neuroscience*, 4 (9):877–886, 2001.

Nicolas Pinto, David D Cox, and James J DiCarlo. Why is real-world visual object recognition hard? 2008.

Tomaso Poggio and Emilio Bizzi. Generalization in vision and motor control. *Nature*, 431(7010):768–774, 2004.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical report, DTIC Document, 2005.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426, 2007.

Christian Siagian and Laurent Itti. Biologically inspired mobile robot vision localization. *Robotics, IEEE Transactions on*, 25(4):861–873, 2009.

Simon Thorpe, Denis Fize, Catherine Marlot, et al. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

License