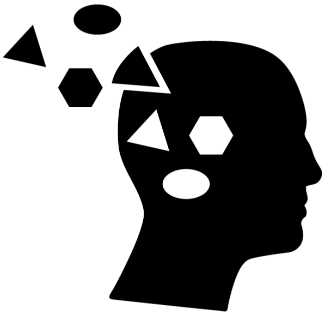
On the Self-organization of a Hierarchical Memory for Compositional Object Representation in the Visual Cortex

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften



vorgelegt beim Fachbereich Informatik und Mathematik
der Goethe Universität
in Frankfurt am Main

von

Evgueni (Jenia) Jitsev
aus Smolensk, Russland

Frankfurt (2010)
(D30)

Vom Fachbereich Informatik und Mathematik der
Goethe Universität als Dissertation angenommen.

Dekan: Prof. Dr. Detlef Krömker
Gutachter: Prof. Dr. Christoph von der Malsburg, Prof. Dr. Rudolf Mester,
Prof. Dr. Jochen Triesch

Für Catherine & Cailie

Abstract

At present, there is a huge lag between the artificial and the biological information processing systems in terms of their capability to learn. This lag could be certainly reduced by gaining more insight into the higher functions of the brain like learning and memory. For instance, primate visual cortex is thought to provide the long-term memory for the visual objects acquired by experience. The visual cortex handles effortlessly arbitrary complex objects by decomposing them rapidly into constituent components of much lower complexity along hierarchically organized visual pathways. How this processing architecture self-organizes into a memory domain that employs such compositional object representation by learning from experience remains to a large extent a riddle.

The study presented here approaches this question by proposing a functional model of a self-organizing hierarchical memory network. The model is based on hypothetical neuronal mechanisms involved in cortical processing and adaptation. The network architecture comprises two consecutive layers of distributed, recurrently interconnected modules. Each module is identified with a localized cortical cluster of fine-scale excitatory subnetworks. A single module performs competitive unsupervised learning on the incoming afferent signals to form a suitable representation of the locally accessible input space. The network employs an operating scheme where ongoing processing is made of discrete successive fragments termed decision cycles, presumably identifiable with the fast gamma rhythms observed in the cortex. The cycles are synchronized across the distributed modules that produce highly sparse activity within each cycle by instantiating a local winner-take-all-like operation.

Equipped with adaptive mechanisms of bidirectional synaptic plasticity and homeostatic activity regulation, the network is exposed to natural face images of different persons. The images are presented incrementally one per cycle to the lower network layer as a set of Gabor filter responses extracted from local facial landmarks. The images are presented without any person identity labels. In the course of unsupervised learning, the network creates simultaneously vocabularies of reusable local face appearance elements, captures relations between the elements by linking associatively those parts that encode the same face identity, develops the higher-order identity symbols for the memorized compositions and projects this information back onto the vocabularies in generative manner. This learning corresponds to the simultaneous formation of bottom-up, lateral and top-down synaptic connectivity within and between the network layers. In the mature connectivity state, the network holds thus full compositional description of the experienced faces in form of sparse memory traces that reside in the feed-forward and recurrent connectivity. Due to the generative nature of the established representation, the network is able to recreate the full compositional description of a memorized face in terms of all its constituent parts given only its higher-order identity symbol or a subset of its parts. In the test phase, the network successfully proves its ability to recognize identity and gender of the persons from alternative face views not shown before.

An intriguing feature of the emerging memory network is its ability to self-generate activity spontaneously in absence of the external stimuli. In this sleep-like off-line mode, the network shows a self-sustaining replay of the memory content formed during the previous learning. Remarkably, the recognition performance is tremendously boosted after this off-line memory reprocessing. The performance boost is articulated stronger on those face views that deviate more from the original view shown during the learning. This indicates that the off-line memory reprocessing during the sleep-like state specifically improves the generalization capability of the memory network. The positive effect turns out to be surprisingly independent of synapse-specific plasticity, relying completely on the synapse-unspecific, homeostatic activity regulation across the memory network.

The developed network demonstrates thus functionality not shown by any previous neuronal modeling approach. It forms and maintains a memory domain for compositional, generative object represen-

tation in unsupervised manner through experience with natural visual images, using both on- ("wake") and off-line ("sleep") learning regimes. This functionality offers a promising departure point for further studies, aiming for deeper insight into the learning mechanisms employed by the brain and their consequent implementation in the artificial adaptive systems for solving complex tasks not tractable so far.

Contents

1 Introduction and motivation 1
1.1 Memory and the cortex : the missing area 51 and other mysteries . . . 3
1.2 Neuronal modeling in machine vision . . . 15
1.3 Objectives and thesis overview . . . 23

2 Elementary cortical module : a neuronal model for unsupervised competitive learning 25
2.1 Fast neuronal dynamics of a cortical module . . . 27
2.2 Homeostatic activity regulation . . . 41
2.3 Activity-dependent bidirectional plasticity . . . 41
2.4 Model parameters and simulation . . . 44
2.5 Unsupervised learning with single and distributed modules . . . 45
  2.5.1 Unsupervised clustering and learning of facial features . . . 46
  2.5.2 Unsupervised feature extraction and gamma cycle coding scheme . . . 54
2.6 Discussion . . . 59

3 A self-organizing hierarchical visual memory: unsupervised learning of a generative compositional object representation 65
3.1 Unsupervised learning of object identity and category . . . 68
  3.1.1 Network architecture, configurations and experimental setup . . . 68
  3.1.2 Assessing network connectivity organization . . . 70
  3.1.3 Results . . . 71
3.2 Processing properties during memory recall . . . 85
  3.2.1 Locking : persistent activity after stimulus removal . . . 85
  3.2.2 Generative pattern completion and attentional mechanisms . . . 86
  3.2.3 Recall and encoding over multiple cycles . . . 89
  3.2.4 Self-generated memory replay in absence of external stimuli . . . 91
3.3 Rapid, non-synaptic learning via excitability regulation . . . 92
3.4 Remarks on scalability . . . 96
3.5 Discussion . . . 102

4 Autonomous off-line memory reprocessing in a sleep-like state and its functional consequences 113
4.1 Off-line memory reprocessing and generalization boost . . . 114
  4.1.1 Off-line regime setup and performance evaluation . . . 114
  4.1.2 Results . . . 114
4.2 Discussion . . . 119

5 Résume and outlook 123
- 5.1 Learning of transformation invariant object representation . . . 124
- 5.2 Memory maintenance via off-line memory replay . . . 126
- 5.3 Further forms of learning . . . 127
- 5.4 Epilog . . . 128

Bibliography 130

Index 163

List of Figures 165

List of Tables 168

Kurzfassung 169

Zusammenfassung in deutscher Sprache 171
- 1 Einführung und Motivation . . . 171
- 2 Modell eines elementaren kortikalen Moduls . . . 173
  - 2.1 Neuronale Mechanismen für unüberwachtes kompetitives Lernen . . . 173
  - 2.2 Unüberwachtes Lernen mit einzelnen Modulen . . . 174
- 3 Ein selbstorganisierendes hierarchisches Gedächtnisnetzwerk für kompositionelle Objektrepräsentation . . . 176
  - 3.1 Unüberwachtes Lernen der kompositionellen Gesichtsrepräsentation . . . 176
  - 3.2 Besonderheiten der Verarbeitung im Gedächtnisnetzwerk . . . 178
- 4 Autonome Gedächtnisverarbeitung in einem schlafähnlichen Regime . . . 179
- 5 Résume und Ausblick . . . 180

Danksagung 183

Curriculum Vitae 184

1

Introduction and motivation

Among the great number of unresolved mysteries about the higher functions of the brain, its ability to learn from the experience is a particularly fascinating one. We learn things permanently, every day we gain new memories and may happen to loose some old ones. Most of this learning happens seemingly effortless, in absence of any special instruction or explicit reinforcement. This ability is not unique to the nervous system of higher primates. All vertebrate animals are to a certain degree flexible in their behavior, being able to acquire complex memories specific to relevant situations or tasks and benefit from the experience made previously if the setting reoccurs. The basis of learning is conserved in evolutionary very old neural mechanisms, as even such comparably primitive creatures as sea slugs show the same basic learning phenomena on neuronal level that are encountered in much more complex vertebrate organisms [Hawkins et al., 1983, Brembs, 2003, Antonov et al., 2003, 2010].

The biological systems seem to have successfully adopted mechanisms of learning a long time ago to secure the survival in complex and uncertain environments (Fig. 1.1). The neuronal processes behind these mechanisms are of acute interest for neuroscience, as there the memory formation and learning were always in the central focus of research. At the same time, unraveling the same mechanisms would be of great use and importance for the fields of artificial intelligence and machine learning. The success of biological systems in solving very complex tasks suggests that whatever principles govern the learning, those surely would be worth mimicking in various domains of technical application.

Unfortunately, there is still an obvious lack of understanding how the basic principles of learning are implemented in the nervous system. This lack is clearly evident in the absence of any artificial systems that were flexible enough to learn autonomously how to cope with the posed tasks without heavy supervision by the human operator. The dominating approach to designing artificial information processing systems is still the hard-wiring of subtask-specific routines that are already known to provide the right intermediate steps toward the solution of a given problem. For many classical problems of artificial intelligence, like object or speech recognition, this hard-wiring approach simply does not bear any fruit if the system has to deal with sensory streams of natural complexity in an uncertain environment. This task setting seems to be immune against algorithmic decomposition into well-defined input-output subroutines that ultimately have to deliver a final answer (e.g., probability distribution over object identities and the details concerning the object appearance and composition) to a provided request (e.g., "what is in the middle of the image"). In contrast to the hard-wiring approach, the learning paradigm does not

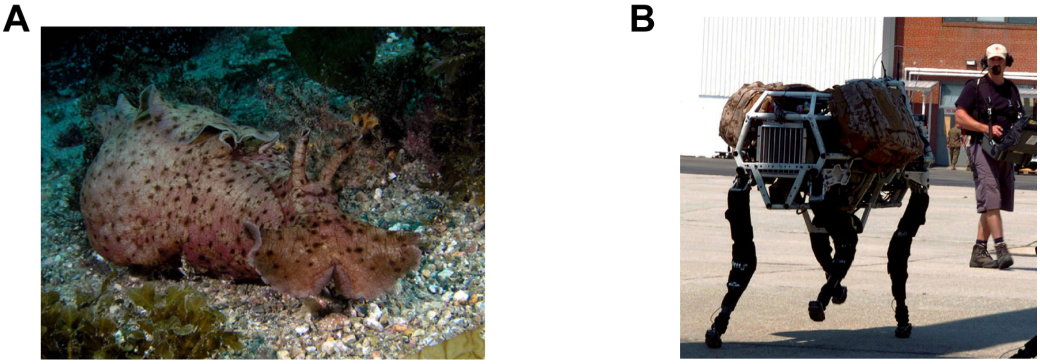


Figure 1.1: Small slug beats BigDog? **(A)** Sea slug, *Aplysia Californica*. The primitive nervous system of this animal (about $20,000$ clearly identifiable, large neurons) is capable of all basic forms of learning, like non-associative habituation and sensitization, and associative classical and operant conditioning [Bailey and Chen, 1983, Walters and Byrne, 1983, Carew et al., 1983, Brembs et al., 2002]. Equipped with these mechanisms, the sea slug is able to adapt perfectly to the environment it lives in. **(B)** BigDog by Boston Dynamics. This quadruped robot is currently one of the most advanced adaptive walkers. The learning is restricted only to locomotion, though. The human operator is necessary to guide the robot along the specified routes.

require an intelligent designer to program the hard-coded solution. Instead, it uses task-related data and examples to figure out how to deal with the task successfully. This is what the brain is adept at. And ideally, this is what an artificial system has to be capable of in order to perform task-solving without any supervision.

The difficulty to develop understanding of learning as phenomenon of adapting the function to the demands of the task is particularly remarkable in light of the great optimism spread at the beginnings of artificial intelligence in 50's. The full solution of the general problem of intelligent processing was prophesied at that time by a number of leading researchers to be achieved within twenty or slightly more years [Simon, 1965, Minsky, 1967]. It is easy to see today that it hasn't worked that way. The experimental neuroscientists were more careful to make predictions of that kind, maybe because they were permanently confronted with the tremendous complexity of cortical circuits and signaling in their everyday experience. A lot of progress was made there by studying the phenomena of adaptation on cellular and synaptic levels, adopting the hypothesis that learning is ultimately caused by and reflected in the changes of intrinsic properties of cells and their synaptic contacts [Bailey and Kandel, 1993, Feldman, 2009]. On the systemic level of larger networks, however, there is no consistent view available on how the distributed cortical networks interact and get coordinated in the processes of memory encoding, consolidation and retrieval.

Obviously, both neuroscience and machine learning research share the difficulty to comprehend learning on the level beyond adaptation of simple isolated subroutines. The difficulties to achieve an essential breakthrough in most unresolved classical problems of artificial intelligence can be arguably traced back to this common deficit. A showcase of such a long-standing problem is the problem of visual object recognition. In what follows, let us take a look on this problem from perspectives of both neuroscience and machine learning, where a lot of effort has been spent to arrive at a functional model, without being able to come up with a successful one so far.

1.1 Memory and the cortex : the missing area 51 and other mysteries

The question about the biological nature of learning and memory formation has a very rich and long tradition, reaching to the time of great Greek philosophers. Already Aristotle discussed the phenomenon of memory in his work. He noted that memory content is based on, but not equivalent to, past perceptual experience, and he also hypothesized physiological conditions to underlie defects of the memory [Aristotle, 1990]. Assumptions about the physiological nature of the memory processes had to be made on the basis of pure thought experiments until quite recently, when the first methods to assess the microanatomy and electrophysiology of the nervous system were developed in the late 19th - early 20th century. One of the very first attempts to create maps of the cortex were the cytoarchitectonic studies carried out on the brain of humans, monkeys and other species by Brodmann, Economo and Koskinas using the Nissl staining method [Brodmann, 1909, von Economo K. and Koskinas, 1925]. Those maps were based solely on the anatomical properties of neurons and their region-specific organization in the brain tissue, resulting in description of the cortex in terms of different segregated areas (Fig. 1.2 **(A)**). Brodmann assigned each area a number, choosing the numbers rather arbitrarily. For instance, it may be of relevance for the conspiracy theorists that the area 51 is missing among the assignments in the map for the human brain. The cortex has for sure enough naturally given secret places though, so there is absolutely no reason to assume yet another one created intentionally by a prominent neuroanatomist.

The cytoarchitectonic maps were created without drawing any potential relation to the functionality of the corresponding cortex areas. Therefore it is remarkable that many of those mapped areas indeed turned out much later to subserve certain coarsely defined functionalities in the cortical processing. For example, the hierarchically organized ventral pathway of the visual cortex involved heavily in the visual object recognition contains stages that correspond pretty well to the cytoarchitectonically defined Brodmann areas (Fig. 1.2 **(B)**). So, the primary sensory visual area V1 responsible for the low-level analysis of the visual input corresponds to Brodmann area 17, and the higher posterior (PIT) and anterior (AIT) areas of the inferotemporal cortex (IT) thought to contain neurons selective for complex forms and whole objects reside in Brodmann area 20. This match indicates that already the coarse anatomical organization may provide useful hints about the cortical function. However, if the aim is to understand the information processing and its changes caused by learning, other methods are required.

In modern neuroscience, two approaches to studying memory processes turned out to be particularly fruitful : behavioral lesion studies and direct assessment and manipulation of neuronal and synaptic signaling in vitro and in vivo. By using lesion studies, it became possible to characterize involvement of specific areas in specific higher brain functions. For instance, a very important breakthrough in memory research was the identification of a structure crucially important for the formation of the declarative long-term memory - the medial temporal lobe (MTL), including the hippocampus, the surrounding cortex areas (parahippocampal, perirhinal and entorhinal cortex) and the amygdala (Fig. 1.3). This breakthrough was possible due to a tragic consequence of a surgical bilateral removal of MTL performed on patient H.M. as part of his epilepsy treatment. The tragic consequence was the complete abolishment of H.M.'s ability to form memories of novel facts, of new persons he met after the surgery or generally of episodic events of his day, while the old memories from his previous life remained intact [Scoville and Milner, 1957, Milner et al., 1998, Squire, 2009].

Subsequent studies with H.M. and further animal experiments made clear that only this certain type of memory formation, namely declarative or explicit memory with the capacity of conscious recall and report, was heavily impaired by the MTL lesion. Another type of memory - implicit memory - was not affected. This memory system, which was dissociated from the explicit, declarative system in

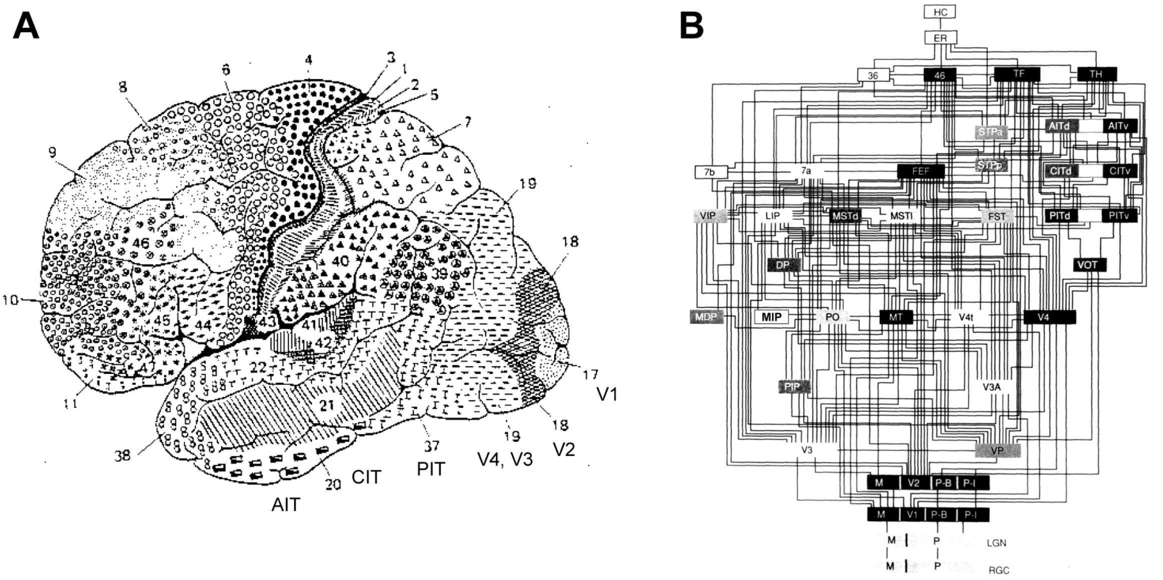


Figure 1.2: Brodmann areas and mapping of cortical pathways. **(A)** 1909, Korbinian Brodmann subdivided human cerebral cortex in 52 discrete areas according to cytoarchitectonic observation of the brain tissue [Brodmann, 1909]. Many of these areas were found later to have a specific functional role in processing. Depicted are the areas V1(17, primary visual cortex)-AIT(20, anterior inferotemporal cortex) of the ventral visual pathway. **(B)** 1991, The functional map of pathways and areas involved in visual processing, established by Felleman and van Essen [Felleman and Essen, 1991]. Using combined neuroanatomical, electrophysiological and tracing methods, it was possible to develop a notion of a processing hierarchy, where the areas can be assigned to either lower (V1, V2) or higher (PIT, AIT) hierarchy stages, depending on their synaptic distance from the sensory thalamic nuclei and the interareal connectivity pattern. Correspondingly, the connectivity arriving in an area could be classified in terms of feed-forward (bottom-up) or feed-back (lateral and top-down) synaptic connections, originating either from an up- or downstream area. At the highest level of the visual processing hierarchy are the entorhinal cortex (EC) and the hippocampal formation (HC).

these studies, includes all types of acquired skills which cannot be accessed explicitly via conscious recollection (Fig. 1.4). These skills comprise procedural motor skills, habits, and also perceptual skills in different modalities, like audition or vision. The subjects with MTL lesion are thus able to perform different learning tasks as well as healthy probands, as long as learning does not require conscious access to the acquired skills [Cohen and Squire, 1980, Squire and Zola-Morgan, 1991]. For example, the subjects can learn to solve complex mechanical puzzles, or complete correctly words shown before if a partial cue (e.g., first three letters) is available, but if asked whether they ever practiced the task in the past, they are not able to remember ever having done it [Milner et al., 1968, Squire, 1987].

These findings highlighted another important point about memory processes. The MTL was essentially important for forming new memories, not for recalling old ones already stored. This led to further distinction of the memory systems into short-term, or working, and long-term memory. The working memory was intact in patients with MTL lesion, and they were able to encode any incoming stimulus properly and keep it on short term as long as they had an opportunity for active rehearsal (e.g., repeating a telephone number again and again, [Sidman et al., 1968]). What was not working properly was obviously the transfer from short into the long-term storage, a process termed memory consolidation. This also indicated that the MTL itself is not the place where long-term memory content is residing. As a great deal of learning turned out to happen in implicit, automatic fashion, it became suggestive that the long-term storage may also be modified in a direct way, without recruiting the MTL systems responsible for declarative memory.

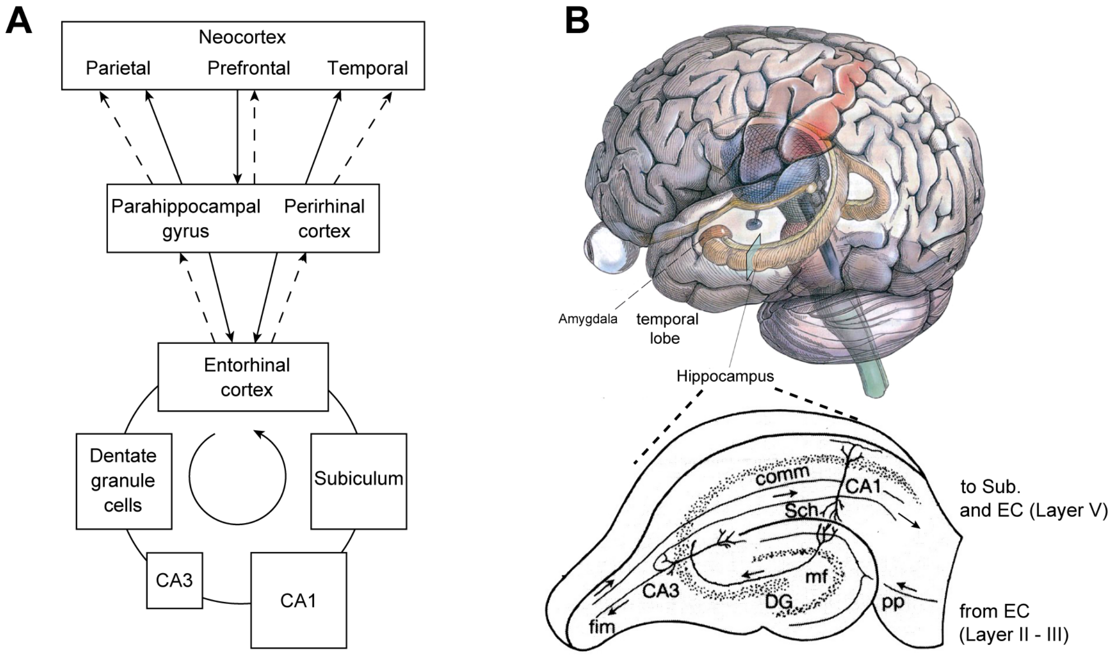


Figure 1.3: Medial temporal lobe and hippocampal formation. **(A)** Main components and pathways of the medial temporal lobe important for the formation of declarative memory. The entorhinal cortex (EC) has a key role as the interface between the neocortical areas and the phylogenetically older hippocampus, which has only three layers instead of six horizontal layers in the neocortex. EC relays the activity from the neocortex to hippocampus and receives the hippocampal output to route it back to the neocortex. This scheme is presumably used for establishing associations between the content stored in different uni- and polymodal cortical areas, linking them together into a coherent memory trace. (Taken from [Rolls, 2004], with permission) **(B)** A coronal slice through the hippocampus (elongated structure running medially through the temporal lobe, amygdala anterior at its head), revealing main excitatory input-output pathways. pp: perforant pathway from EC Layer II-III converges on the granule cells (GC) of the dentate gyrus (DG). The GC synapses can undergo associative long-term potentiation (LTP) depending on pre- and postsynaptic activity. mf: mossy fiber pathway from GC to hippocampal region CA3. LTP is non-associative (presynaptic only) at mf synapses. fim: fimbrial pathway heading to the CA3 region of the contralateral hemisphere. comm: commissural pathway, fibers arriving from the CA3 region of the contralateral hemisphere. Sch: Schaffer collateral pathway to CA1 region. Both comm and Sch have associative LTP at the synapses. The output goes from CA1 to Subiculum and then to the EC Layer V. Distinct CA regions are thought to be a part of a vast hippocampal network that contributes in an essential way to the memory formation and consolidation process. The amygdala is sometimes considered to be a part of the hippocampal formation, as it is able to modulate memory formation by signaling motivational and emotional cues relevant for memorizing a content. (Adapted from [Kandel et al., 2000])

The implicit nature of skill acquisition is particularly obvious for perceptual learning. In the visual cortex, complex objects are thought to be rapidly decomposed into their constituent parts along the processing hierarchy of the ventral pathway during memory recall or during the encoding of a novel object [Fujita et al., 1992, Tsunoda et al., 2001, Fiser and Aslin, 2005, Reddy and Kanwisher, 2006, Connor et al., 2007]. This decomposition procedure has to rely on the visual experience made before. Still, it is impossible to consciously access the skills used for this decomposition. Consequently, the long-term memory for visual content is thought to be distributed along the visual processing hierarchy. To study the function of the visual memory system, one obviously has to go at least down to the level of cortical circuits and local operations instantiated along the visual pathways. On the other hand, one has to stay on the coarser scale of networks to address the question about the coherent object representation used by the visual cortex to encode, store and retrieve its content.

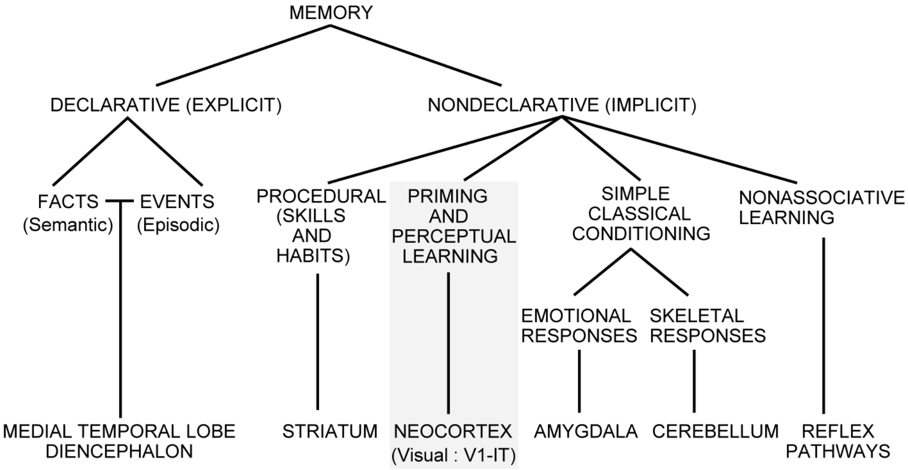

MEMORY
DECLARATIVE (EXPLICIT)
NONDECLARATIVE (IMPLICIT)
FACTS (Semantic)
EVENTS (Episodic)
PROCEDURAL (SKILLS AND HABITS)
PRIMING AND PERCEPTUAL LEARNING
SIMPLE CLASSICAL CONDITIONING
NONASSOCIATIVE LEARNING
EMOTIONAL RESPONSES
SKELETAL RESPONSES
MEDIAL TEMPORAL LOBE DIENCEPHALON
STRIATUM
NEOCORTEX (Visual : V1-IT)
AMYGDALA
CEREBELLUM
REFLEX PATHWAYS

Figure 1.4: Various forms of memory. The studies subsequent to the H.M. case revealed two main forms of memory, declarative (explicit) and non-declarative (implicit). While declarative memories can be consciously accessed, or recollected, the implicit memories are expressed in behavioral performance without the ability of conscious recollection. The perceptual visual memory, which will be the topic of this thesis, falls into this second memory form. The visual long-term memory is presumably subserved by the ventral pathway of the visual cortex. The memory traces for visual objects are most probably laid down along the hierarchical ventral pathway, storing low-level perceptual elements in lower areas like V1, V2 and combining them to more and more complex forms on the way to the inferotemporal cortex IT, where neurons show selectivity for specific objects. (Modified from [Squire and Stark, 2008], with permission)

A substantial amount of research was done on both levels. Concerning the cortical circuits, a classical finding is the vertical and horizontal organization of the neuronal cells that make up the microcircuits of different cortical areas [Douglas and Martin, 2004, Thomson and Lamy, 2007]. Different types of the neurons populate six horizontal layers (layer I-VI) spreading from the surface to the white matter of the cortex. Each layer has not only a particular type of cells residing in it, but also a highly specific synaptic connectivity arriving from the outside (Fig. 1.5). For instance, layer IV, containing excitatory spiny stellate cells (SSC), receives predominantly bottom-up inputs from the excitatory pyramidal (PY) neurons of preceding upstream areas, while layer II/III containing tightly coupled groups of excitatory pyramidal cells receives mainly top-down and lateral recurrent inputs from higher areas or from other excitatory populations in the same area. Further, there is a local tendency of PYs to cluster their bodies and bundle their ascending apical dendrites vertically on the way toward the cortical surface [Peters and Sethares, 1996] (Fig. 1.6 (**A**), (**B**)). These segregated vertical cell clusters seem to be also bound together by short-range inhibitory and excitatory lateral connectivity [Mountcastle, 1997, Rockland and Ichinohe, 2004].

The notion of vertical organization gave rise to the hypothesis of an elementary unit of cortical processing termed minicolumn [Lorente de No, 1938, Szentágothai, 1978, Mountcastle, 1997, Jones, 2000, Buxhoeveden and Casanova, 2002]. Recent studies of local microcircuitry suggest a more complicated picture of the fine-scale excitatory subnetworks that may correspond to these hypothetical elementary processing units (Fig. 1.6 (**C**)). Such subnetworks do not seem to obey the anatomical principles of vertical organization. Instead, they comprise tightly coupled sets of PYs from layers II/III with rather arbitrary layout within the layers [Yoshimura and Callaway, 2005, Yoshimura et al., 2005, Song et al., 2005, Haider and McCormick, 2009]. The subnetworks are thus defined in terms of high probability of functional coupling between the constituent neurons and by the common afferent input they receive from the SSCs of the layer IV. These subnetworks are thought to form bigger clusters, or modules, in

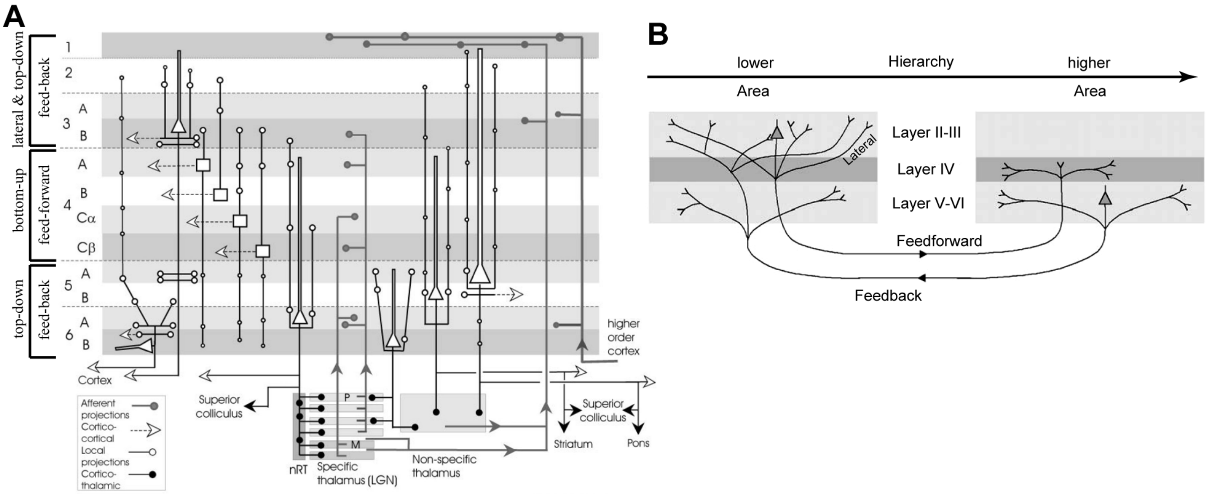


Figure 1.5: Cortical microcircuits, cell and synaptic specificity. **(A)** A simplified generic layout of a circuit for a primary sensory area (primary visual area V1 shown). The excitatory pyramidal cells (PY) are situated in layers II-III and layer V-VI. The excitatory spiny stellar cells (SSC) from layer IV are the main recipients of the thalamocortical synapses originating from neurons in lateral geniculate nucleus (LGN, receives input from retina over the optic nerve through to parallel separate channels, P (parvocellular) and M (magnocellular), that are preserved in V1). SSC contact PY in layer II-III, which in turn project to PY in layer V. PY in layer V project back to PY in layer II-III, creating a local excitatory feedback loop. PY in layer II-III receive also feedback from the higher areas and send lateral connections within the same area and feedforward connections to the next area of the processing hierarchy. PY from layer V and VI provide output for various subcortical structures. (Modified from [Thomson and Bannister, 2003], with permission) **(B)** A generic view on signal exchange between areas in a hierarchical pathway. An upstream area sends feedforward signals from PY layer II-III to the area downstream, where the signals arrive mainly at SSC layer IV. The downstream area sends feedback signals that arrive at PY layer II-III of the upstream area. Within an area, feedback exchange occurs also over the lateral connections between PY in layer II-III. (Modified from [Bullier, 2003], with permission)

which unspecific excitation and inhibition links the segregated subnetworks together.

Revision of the classical hypothesis of a minicolumn does not give up the notion of tightly coupled neuronal populations forming elementary processing units within a larger functional module. Such modules can indeed be found everywhere along the visual processing hierarchy (Fig. 1.7, 1.8). In the primary visual cortex V1, which is arguably the most thoroughly studied area of the neocortex, each module processes a tiny part of the visual field and contains neurons selective for a full range of orientations and spatial frequencies [Hubel and Wiesel, 1977, DeValois and DeValois, 1990]. On the top of the ventral pathway, the areas of inferotemporal cortex (IT) were found to be composed of modules with neurons that prefer stimuli of much higher complexity than the lower visual areas [Fujita et al., 1992, Tanaka, 1996, 2003, Sato et al., 2009a]. In these stages on the top of the visual processing hierarchy, the neurons tend to respond to complex shapes or even whole objects or faces (found in particular areas dedicated to face processing, like occipital face area OFA and fusiform face area FFA [Perrett et al., 1992, Kanwisher et al., 1997, Tsao et al., 2006, Liu et al., 2009]), ignoring most simple stimuli like oriented bars or gratings (Fig. 1.9). The modules in these areas cluster neurons with selectivity to similar, but different stimuli, so that each module can be considered as a container for a number of related shapes or objects.

Bearing this picture of the hierarchically organized, recurrently interconnected stages of distributed modules in mind, what can be hypothesized about object representation arising in such a neuronal architecture and about the processes shaping the memory structure there? Each module can be seen

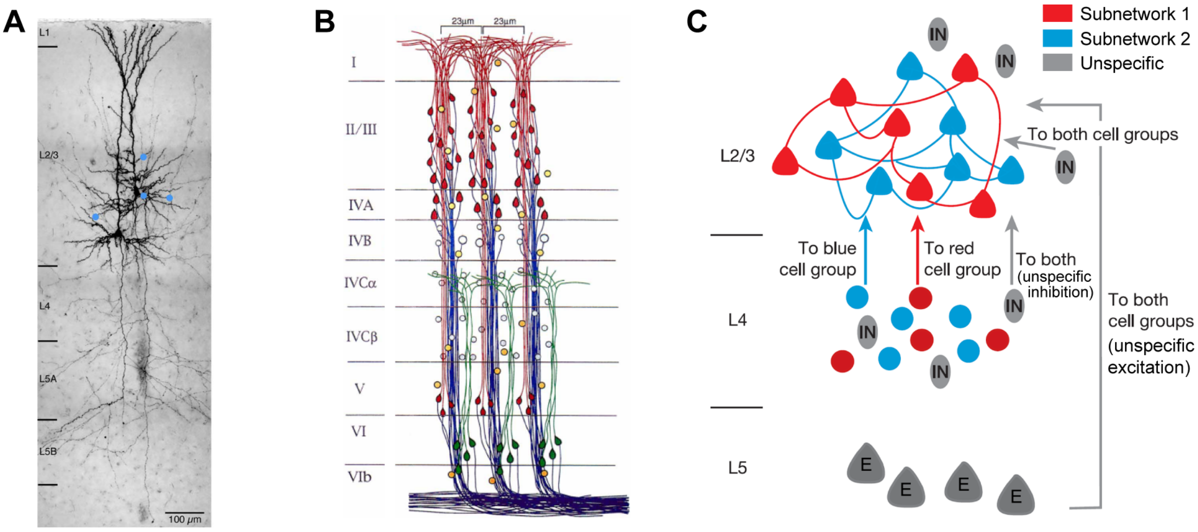


Figure 1.6: Functional excitatory fine-scale subnetworks formed by tightly coupled PY. **(A)** Tight coupling between two biocytin-labeled PY from layer II-III in the somatosensory cortex of a rat (barrel cortex). The presynaptic PY (left) contacts the postsynaptic PY at different dendritic locations (blue dots). The probability of the reciprocal connection within such pair is substantially higher than between two arbitrary neurons. (Taken from [Lübke and Feldmeyer, 2007], with permission). **(B)** The classical view of minicolumns. Adjacent clusters of PY bodies and their ascending apical dendrites form vertical structures (with spacing between the clusters of about $23\mu m$ in primary visual cortex of monkey) that were hypothesized to be an elementary unit of cortical processing. The minicolumns (also termed pyramidal modules) were thought to share the same afferent input, being embedded into a larger functional structure (macrocolumn) that extends over $600-800\mu m$ and contains $80-200$ minicolumns depending on the cortical area. (Modified from [Peters and Sethares, 1996], with permission) **(C)** Revised view on excitatory fine-scale subnetworks of PY with non-vertical layout composing together a functional module within a small cortical patch $(600-800\mu m)$. Two schematic subnetworks are shown. A subnetwork is defined by strong connectivity between a subset of PY from layer II-III. These cells receive common afferent input from a portion of SSC on layer IV. All subnetworks receive unspecific excitation from PY in layer V (E) and unspecific inhibition from the inhibitory fast-spiking interneurons (IN). (Modified from [Yoshimura et al., 2005], with permission)

as a vocabulary of visual elements of different complexity depending on the processing stage. These vocabularies were developed in the course of experience with the visual world. Whatever image falls upon the retina, it gets immediately interpreted by the responses distributed across the recurrently interconnected vocabularies. These stimulus-evoked responses correspond to the sparse activation along the visual pathway observed in experimental studies [Young and Yamane, 1992, Weliky et al., 2003, Quiroga et al., 2005, 2008]. An arbitrary object can be represented in this fashion as a composition of reusable universal parts taken sparsely from the overcomplete set offered by the established vocabularies. Recent experimental work suggests that also the relations between the different reusable elements, of which the objects are composed, are explicitly extracted from the visual input and captured in the structure of the long-term visual memory [Fiser and Aslin, 2002, 2005].

If put together, this evidence provides a view of an explicitly compositional, generative nature of object representation employed by the visual cortex. Such representation offers rich description of any object identity in terms of a hierarchy of parts. Besides its rich expressive power, representation of this kind has another crucial advantage for memory formation. Because it relies on universal, reusable vocabularies, it can instantiate objects never experienced before as collections of best fitting elements, immediately providing a substrate for a memory trace in combinatorial fashion (Fig. 1.10). Further, in the memory recall situation, the explicitly captured higher-order relations among the vocabulary el-

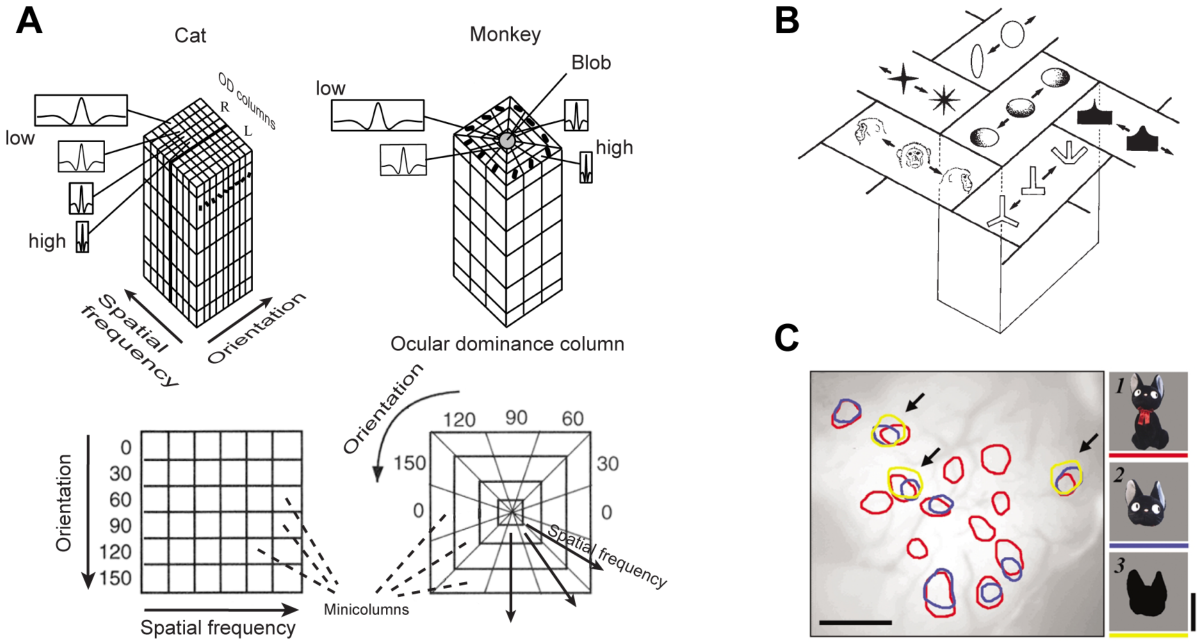


Figure 1.7: Functional modules along the ventral visual pathway in the visual cortex. **(A)** A schematic module from the primary visual cortex V1 of cat and monkey. Each module contains clusters of neurons selective for a specific orientation and spatial frequency. Further, the neurons are grouped into ocular dominance columns according to the eye (left, right) they receive their afferent input from. (Adapted from [DeValois and DeValois, 1990]) **(B)** In the higher visual area IT a module is thought to contain clusters of neurons that are selective to similar, but not identical complex forms or objects. These modules may thus subserve invariant recognition of an object, while at the same time providing capability to emphasize subtle differences between the object instances. (Taken from [Tanaka, 2003], with permission) **(C)** Experimental evidence for the parts-based object representation in the higher visual cortex. Complex objects were presented to a macaque monkey, simplifying the appearance subsequently by removing some parts, and the neuronal responses of the dorsal TE (corresponds roughly to PIT) were visualized through combined intrinsic optical imaging (IOS) and extracellular recording from the cortex. A cat image elicited a number of active spots in the area. Showing simplified versions elicited a subset of those spots, with an apparent correspondence between the subset size and the degree of parts reduction. More complex coding was also observed, where removing the object parts evoked spots not active during the presentation of the full object. This indicates that parts-based object representation does not necessarily rely on a simple linear combination of the part-specific modules. (Taken from [Tsunoda et al., 2001], with permission)

ements would provide valuable contextual support for the recognition process. This is because these relations can restrict the search for a potential memorized candidate on only the few relevant element combinations that correspond to memorized content instead of searching through all of them, effectively avoiding a combinatorial explosion [Fiser and Aslin, 2005, Oliva and Torralba, 2007].

To make use of the multiple advantages of a compositional object representation, the visual cortex has of course to provide all essential prerequisites. While the modules of the early visual areas like V1 could in principle develop their vocabularies even before experiencing natural stimulation, being driven by the self-generated retinal activity observed already in prenatal state [Albert et al., 2008], it is obvious that more complex vocabularies of higher areas containing different shapes or the selectivity of the modules for whole objects can only be formed by learning from natural visual input. The same applies to the relations between vocabulary elements, which have to be extracted from the higher-order statistics of the visual scenes and captured in the recurrent lateral and top-down connectivity within and between the visual cortical areas. The contextual support provided by learning such high-order relations is in general of crucial importance for correct interpretation of visual stimuli embedded in a

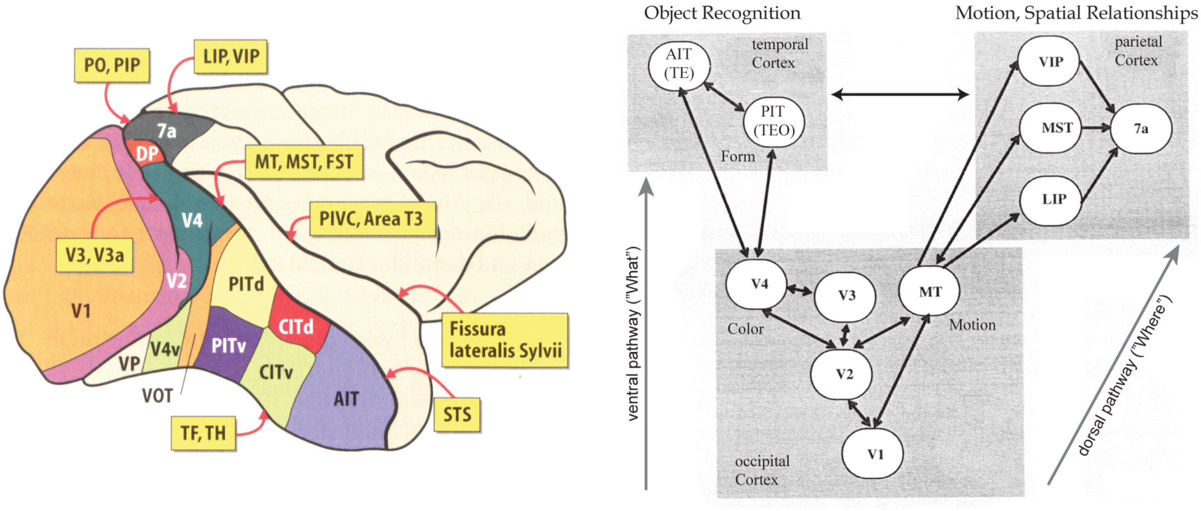


Figure 1.8: Parallel processing pathways in visual cortex and their functional segregation. **(A)** Lateral view on a macaque brain emphasizing the ventral visual pathway V1-AIT. Both the ventral and the dorsal pathway start in the occipital lobe in V1. While the ventral pathway heads toward temporal lobe, the dorsal pathway goes for the parietal lobe. The regions of the dorsal pathway are buried within the superior temporal sulcus (STS) on this view. (Taken from [Birbaumer and Schmidt, 2005], with permission) **(B)** A simplified view on the functionally segregated ventral and dorsal pathways. Although both routes participate in intense signal exchange, early lesion and subsequent neurophysiological studies assigned to the ventral pathway a major role in analysis and recognition of visual appearance (form vision, "what" pathway), while the dorsal pathway was shown to be crucial for identification of spatial relationships in the visual scene and for visuomotor guidance (space vision, "where" and "how" pathway) [Mishkin et al., 1983, Goodale and Milner, 1992]. (Modified from [Rolls and Deco, 2002], with permission)

larger context (e.g., object or scene). Their local appearance is usually highly ambiguous and can be correctly interpreted only if consulting additional contextual cues mediated by the connectivity formed during the previous experience with the visual input (Fig. 1.11). This perceptual learning is most probably of the ongoing nature, potentially affecting each level of the processing hierarchy including the early visual areas like V1 and V2 [Ahissar and Hochstein, 1997, 2004, Gilbert et al., 2009].

The question how this learning may happen in detail on the neuronal network level is clearly relevant not only for the visual modality, but for every kind of cortical processing. Still, most knowledge about the neuronal mechanisms subserving learning concentrates on the level of a single cell. The plasticity mechanisms described there include various types of possible activity-dependent modification of synaptic connections converging on the neuron [Bear et al., 1987, Lisman, 1989, Artola et al., 1990] and of the intrinsic properties of the neuron itself [Desai et al., 1999, Zhang and Linden, 2003, Marder and Goaillard, 2006, Maffei and Turrigiano, 2008] (Fig. 1.12). Now, a memory trace that should capture the compositional representation of an object clearly has to comprise multiple neuronal populations and their interconnectivity. Thinking about the cortical module as a building block within the visual processing hierarchy, it becomes necessary to consider how the process of memory formation recruits these distributed modules and how the changes necessary for the creation of a memory trace are coordinated across them.

An attractive hypothesis in this context states that activity formation and plasticity may be coordinated by the ongoing cortical rhythms in the gamma range $(30 - 100Hz)$. These gamma rhythms are presumably generated locally by the interaction between PYs and fast-spiking inhibitory interneurons (FS) within the microcircuits [Traub et al., 1996, Tiesinga and Sejnowski, 2009] (Fig. 1.13). More

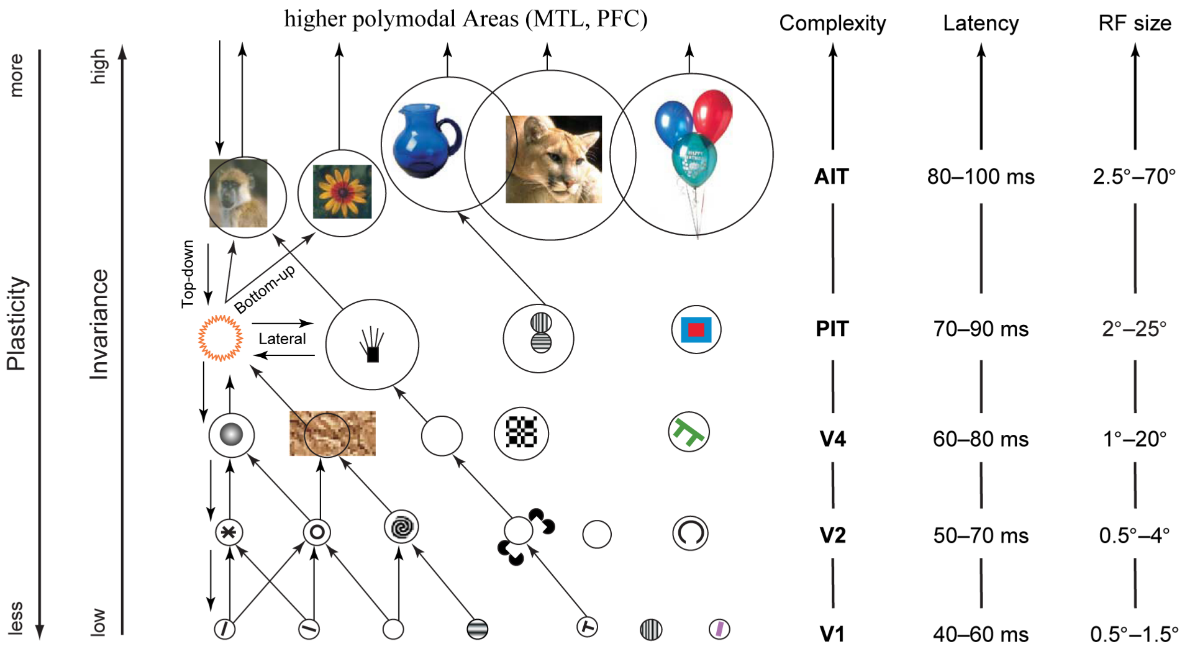


Figure 1.9: Processing hierarchy of the ventral pathway of the visual cortex. Starting from the primary visual cortex V1, the neurons become progressively selective to more and more complex visual stimuli along the ventral pathway. The receptive field size increases as well as the tolerance against different transformations like shift, scale etc., rendering the responses of the neurons more and more invariant with increasing hierarchy level. The response latencies for a feed-forward sweep after stimulus onset are shown on the right. The recurrent processing mediated through lateral and top-down connectivity contributes heavily to the stimulus interpretation by refining the initial coarse hypothesis formed by the feed-forward sweep. In addition, the ongoing cortical activity across all stages shapes most probably the stimulus-evoked responses, such that recurrent connectivity pre-imposes an experience-dependent bias on the visual object recognition even before stimulus onset. Perceptual learning is thought to involve all processing stages in task-dependent manner, so that connectivity modification is also possible in the early visual areas if a task requires so. In natural situations, most of the adult visual learning is assumed to take place in the higher areas of inferotemporal cortex, PIT-AIT. (Modified from [Rousselet et al., 2004], with permission.)

specifically, there is evidence that cortical processing is composed of discrete atomic fragments, each fragment being a single cycle of an ongoing gamma rhythm [Pöppel, 1997, VanRullen and Koch, 2003, Fries et al., 2007, Luo and Poeppel, 2007]. In each cycle, the local operation performed by the module may resemble a type of competitive computation termed soft winner-take-all (sWTA)[Douglas and Martin, 2004]. This competitive computation is thought to select and amplify a very small number of units within a module that are able to become active in course of one gamma cycle [Zhang and Ballard, 2002, de Almeida et al., 2009]. The cycles are assumed to run in synchrony across the distributed modules contributing to a current perceptual task, even if the processing modules are separated by long distances in the cortex. The sparse activity formed across the processing hierarchy during the cycle can potentially provide a neuronal substrate for a memory trace. Remarkably, there is evidence showing that synaptic potentiation and depression occur predominantly in certain phases of the ongoing rhythm [Huerta and Lisman, 1993, 1995, Wespatat et al., 2004]. This alignment of plasticity events to a common rhythm could then hypothetically be used for linking the units activated within a cycle to an assembly, forming a memory trace for the currently processed object.

Another interesting hypothesis is that the participation of different units in memory traces could be regulated by a local homeostatic mechanism. A candidate for such mechanism could be plasticity of

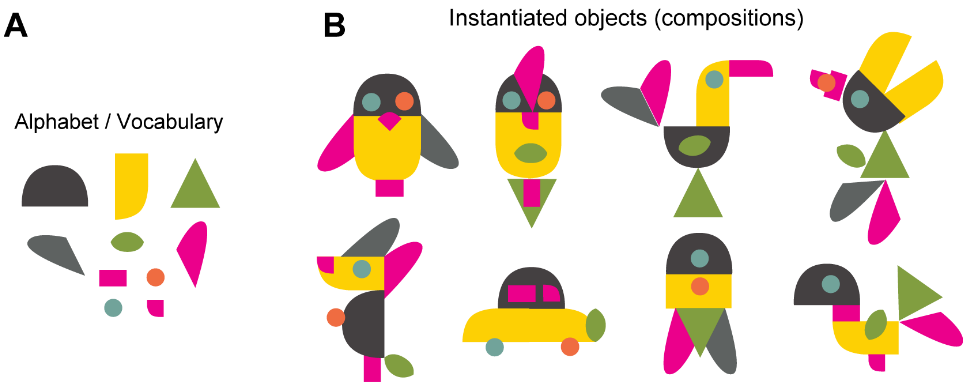


Figure 1.10: Vocabulary of universal elements and combinatorial representational power. Using a vocabulary of universal primitives, it becomes possible to instantiate a great number of arbitrary objects, composing them from the existing primitives of much lower complexity. The simplified scenario visualized here is also inspired by the classical ideas put forward in [Biederman, 1987]. **(A)** A vocabulary of few elements. **(B)** Instantiated complex objects. (In natural vision, it is more advantageous to have an overcomplete vocabulary, so that only few elements have to be picked out for representing an arbitrary object, which is not the case in this toy example.) Image courtesy of Catherine Lubbers, 2009.

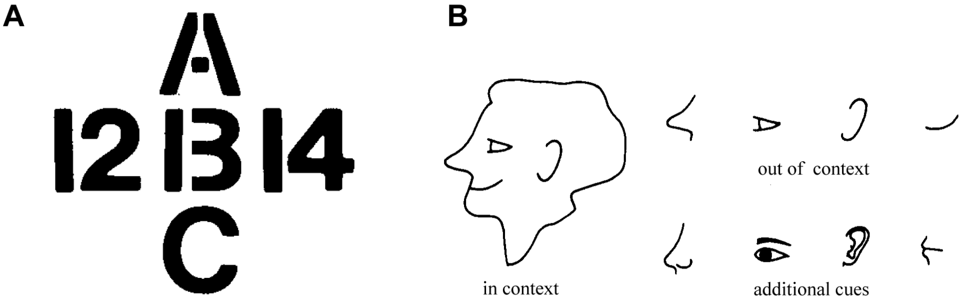


Figure 1.11: Role of context in visual recognition. **(A)** The same stimulus can often have different meaning depending on the context. To interpret such a stimulus correctly, exchange of global contextual cues becomes necessary. **(B)** A local feature becomes interpretable only in a global context learned from previous experience, being meaningless if isolated. Adding more detail can render the feature again interpretable. (Modified from [Coren et al., 1999] , with permission)

intrinsic neuron excitability, which was observed in experiments involving prolonged over- or understimulation applied to neuronal cells [Marder et al., 1996, Desai et al., 1999, Debanne et al., 2003, Maffei and Turrigiano, 2008]. Exposed to over- or understimulation, the neurons were found to adjust their physiological response properties to keep their standard activity level (Fig. 1.12 **(C)**). Such homeostatic activity regulation may provide a simple way to reassure that all neuronal units of the processing hierarchy contribute to an equal extent to the memory encoding and formation, resulting in a balanced usage load across the available neuronal resources.

Further, a recent line of research about off-line memory reprocessing in states of sleep and restful waking shows that memory processes are by no means restricted only to phases of active perception and action. It has been known since the fifties that the brain alternates between different global states of activity, switching from wakefulness to different sleep regimes, non-rapid eye movement (NREM) and REM sleep [Aserinsky and Kleitman, 1953, Hobson et al., 1975, Hobson and Pace-Schott, 2002] (Fig. 1.14). However, only recently clear experimental evidence became available from behavioral studies showing effects of the sleep stages on formation and consolidation of memory [Stickgold et al.,

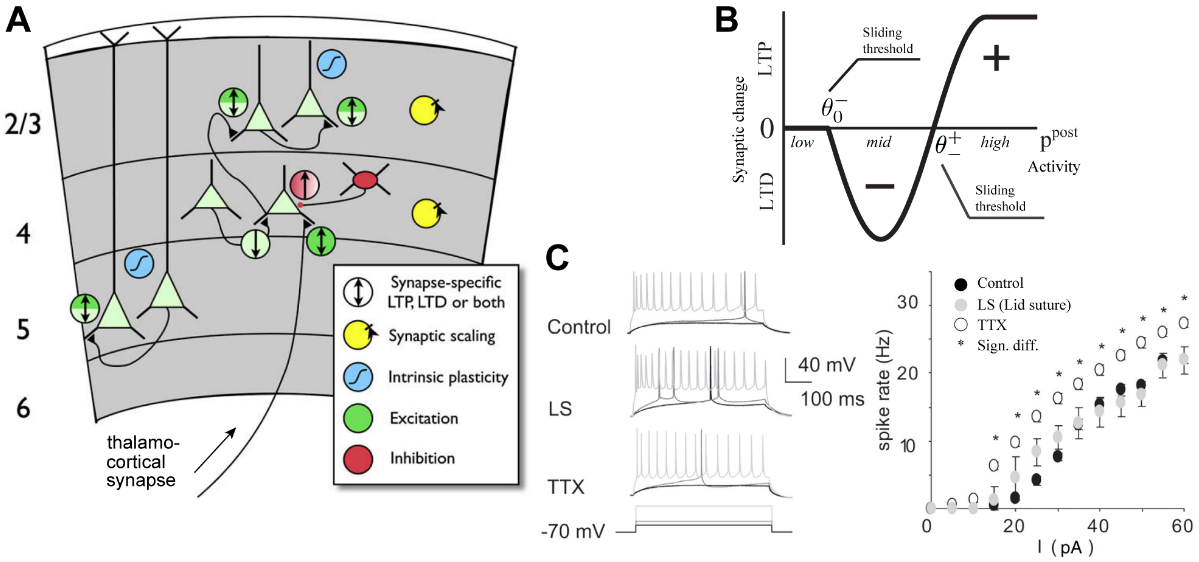


Figure 1.12: Multiple forms of plasticity in a cortical microcircuit. **(A)** A diagram of plasticity in a simplified primary sensory area. Most excitatory synapses are capable of bidirectional plasticity, either potentiating (LTP) or depressing (LTD) their strength in activity-dependent manner (green circles, up and down arrows, light green for pre-, dark green for postsynaptic modification locus, mix indicates both forms). Inhibitory synapses from interneurons onto PY can also be plastic, the locus of the modification is still unknown (red circle). Further, PY are able to regulate their intrinsic somatic properties in response to over- and understimulation, altering a neuron's response to the afferent input in global, synapse-unspecific manner. This homeostatic mechanism is known as plasticity of intrinsic excitability (blue circles). In addition to the somatic homeostatic mechanism, excitatory and inhibitory synapses exhibit homeostatic synaptic scaling to maintain the synaptic weights in the range suitable for ongoing learning (yellow circles). (Modified from [Maffei and Turrigiano, 2008], with permission.) **(B)** Bidirectional plasticity of excitatory synapse as observed in experiments. The synapse undergoes potentiation or depression depending on the level of pre- and postsynaptic activity p^{post}. LTP and LTD zones are determined by sliding thresholds on the postsynaptic side that are updated according to the previous history of postsynaptic activation. This dependence of plasticity on the previous stimulation history is known as metaplasticity. (Adapted from [Artola and Singer, 1993]) **(C)** Plasticity of intrinsic excitability assessed in an in-vivo experiment with mice. PY from layer II-III of the primary visual cortex of the mouse were deprived using two different paradigms. TTX: intraocular injection of Tetrodotoxin heavily decreases the activity transmitted by the optic nerve of the deprived eye; LS : lid suture leads to general reduction of the visual drive and causes decorrelation of the signals arriving from different eyes. Both paradigms cause abnormal underactivation of the PY, which is counteracted by the homeostatic regulation mechanism. **(Left)** Response of PY neuron to depolarizing DC current steps of different amplitude (black, low; light gray, high) for different conditions. **(Right)** Response properties shown for different conditions in the form of frequency-current curves. Asterisk indicates significant differences between Control and TTX condition. (Modified from [Maffei and Turrigiano, 2008], with permission)

2001, Wagner et al., 2004, Stickgold, 2005, Marshall and Born, 2007, Diekelmann and Born, 2010]. This evidence indicates that sleep in general enhances the memory performance on the tasks learned before if compared to the subjects that didn't have opportunity to sleep after learning [Gais et al., 2000, Stickgold et al., 2000, Walker et al., 2002, Fischer et al., 2002, Wagner et al., 2007]. Less clear is whether different sleep stages are relevant for consolidation of different memory types. Some evidence suggests that procedural memory for motor-dependent tasks may be preferentially consolidated during REM sleep or NREM stage 2 [Walker et al., 2002, Peigneux et al., 2003], while declarative memory (e.g. for associative word pair lists) requires reprocessing in NREM, or more specifically slow-wave-sleep stage (SWS, NREM Stage IV) for the consolidation [Gais and Born, 2004a,b, Marshall et al.,

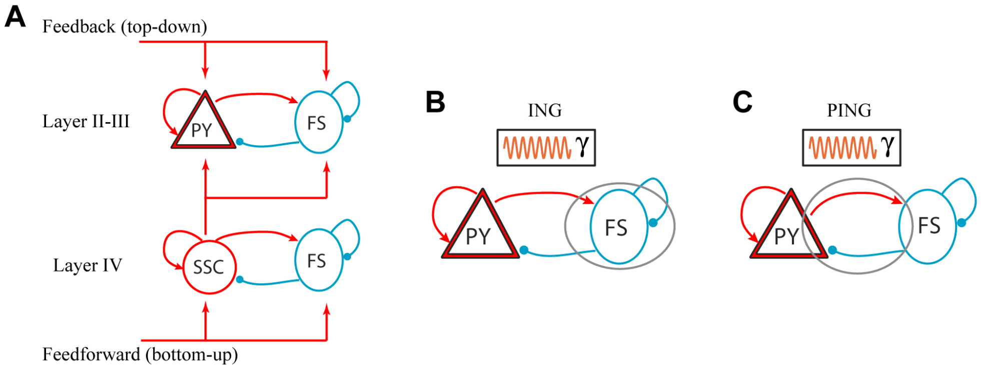


Figure 1.13: Different strategies of local rhythm generation in cortical microcircuits. **(A)** A hypothetical generic cortical microcircuit, receiving feed-forward and feed-back afferents from up- and downstream areas. In layers II-III and IV there are reciprocally connected excitatory PY and fast-spiking inhibitory FS cells. Red excitatory, blue inhibitory transmission. **(B)** Hypothetical possibilities of local γ-rhythm $(30 - 100Hz)$ generation: ING (interneuron gamma) and PING (pyramidal-interneuron gamma). The ING mechanism requires only the FS to become active for gamma rhythm generation. The synchrony is developed due to the self-inhibitory connections that stop FS cells from firing once an activation volley was produced. The FS engaged in the rhythm drives then PY that in turn further facilitate the common rhythm via signal exchange with FS. For the PING mechanism, the synchronous activation of PY and subsequent signal exchange between PY and FS are essential for proper rhythm generation. (Adapted from [Tiesinga and Sejnowski, 2009])

2006, Rasch et al., 2007]. Other studies indicate in turn that both sleep stages are important for memory consolidation in such tasks as visual texture discrimination and motor skill acquisition [Gais et al., 2000].

Although behavioral improvement in cognitive tasks can be clearly stated in studies that concentrate on the memory function after sleep, there is no consistent explanation how this improvement comes to pass. A finding which recurs among many different studies is memory replay in the form of spontaneously generated activity observed in the cortex, hippocampus and striatum during the SWS and REM sleep [Wilson and McNaughton, 1994, Skaggs and McNaughton, 1996, Dave and Margoliash, 2000, Louie and Wilson, 2001, Pennartz et al., 2004, Euston et al., 2007, Ji and Wilson, 2007, Lansink et al., 2008]. This memory replay is considered to be one of the hallmarks of off-line memory reprocessing. A similar type of replay is also observed in wake periods a short time after active task-related learning [Foster and Wilson, 2006, Yao et al., 2007, Axmacher et al., 2008b, Fuentemilla et al., 2010]. The replay further seems to be coordinated between hippocampus and neocortex by the ongoing theta rhythms synchronized across the regions where the activity is generated, both in wake and in sleep regime [Siapas and Wilson, 1998, Louie and Wilson, 2001, Axmacher et al., 2008b, Wagner et al., 2009, Düzel et al., 2010].

It becomes suggestive to put this self-generated memory replay in relation to the improvement of memory function assessed after off-line reprocessing. It is conceivable that such reactivation of memory traces may help to stabilize weak, labile newly formed memories or that it may perform another kind of maintenance like refreshing of older content or avoiding interferences by reducing the overlap between strongly conflicting memory traces [Káli and Dayan, 2004, Norman et al., 2005]. Another hypothesis postulates the transfer of the initially encoded memories from the fast learning subsystem of the hippocampus to the slow learning system of the neocortex to occur during such replay episodes [Siapas and Wilson, 1998, Hasselmo and McClelland, 1999, Rasch and Born, 2007]. This transfer may then explain why recall of well consolidated memories becomes subsequently independent of hippocampus with time. Overall, it becomes obvious that a functional model of memory formation and

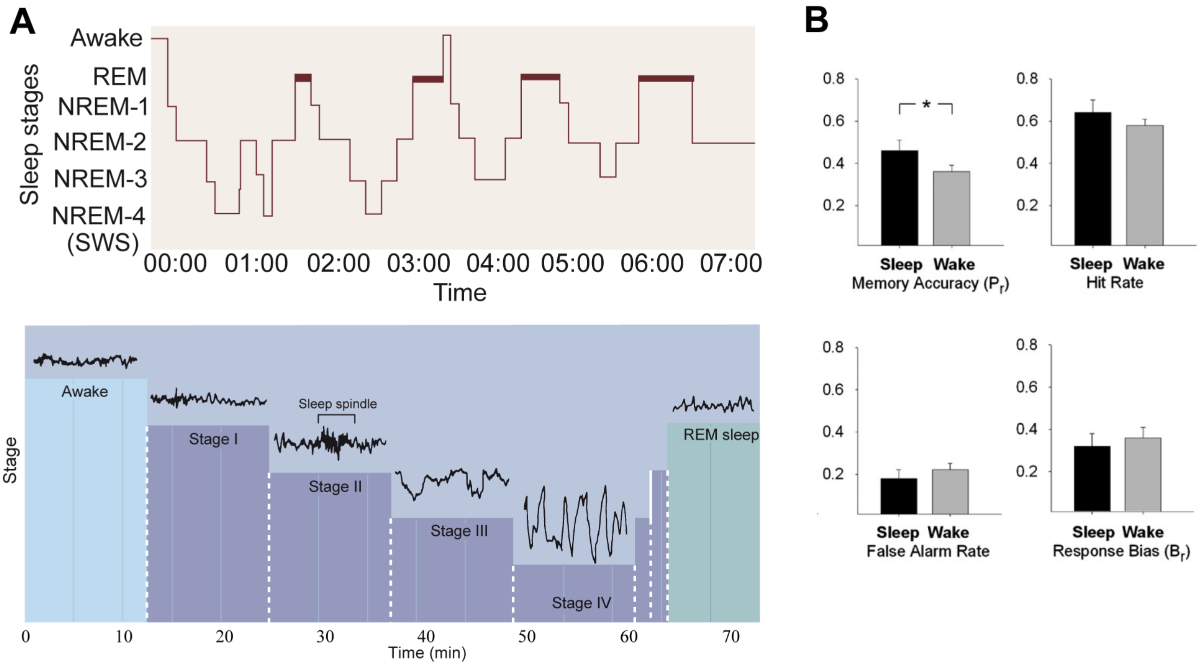


Figure 1.14: Sleep and memory function **(A)** Different sleep stages in the brain are characterized by distinct types of global activity formation reflected in EEG. The brain runs through a full sleep cycle in about $70 - 90$ minutes. The NREM sleep stages are first to be visited, culminating in slow-wave-sleep (SWS, NREM Stage IV) characterized by large amplitude slow oscillations clearly seen in EEG. From this stage, a fast transition to rapid eye movement sleep stage (REM) is performed. There, the EEG signal resembles closely the pattern observed during wakefulness. Different states of brain function are also characterized by different levels of neuromodulation. Cholinergic activity is high during wakefulness and REM, whereas its level is very low during the SWS. Noradrenergic and serotonergic neuromodulation is in turn almost absent in REM sleep. (Taken from [Purves et al., 2004, Stickgold, 2005], with permission.) **(B)** An experiment showing the improvement in visual recognition tasks after sleep over non-sleep. The subjects were shown a number of persons they had to memorize from natural face images. Following learning, one group were allowed to sleep at night, while the other group were kept awake (23:00-7:00 h) in the laboratory. All subjects were given an opportunity to make a recovery sleep the next day after the learning. The recognition tests were performed the night after the recovery sleep. Sleep after learning clearly boosts memory performance in face recognition task ("was the face presented before?") compared to the condition without sleep ($^*p < .05$). (Taken from [Wagner et al., 2007])

maintenance has to include not only the periods of active on-line learning, but also account for the functional improvement caused by reprocessing of memory in off-line states.

1.2 Neuronal modeling in machine vision: learning from predictions that went wrong

The development of artificial vision systems has always been driven by the goal of approaching the functionality of the natural visual system. Indeed, the methods of computational neuroscience and machine learning could potentially offer a way to obtain functional proof for hypothetical mechanisms of processing and learning employed in the brain. Such functional proof would involve constructing a system to handle visual objects just by being exposed to natural visual input, in largely unsupervised fashion. At the current state of research, this ideal seems to be still rather distant.

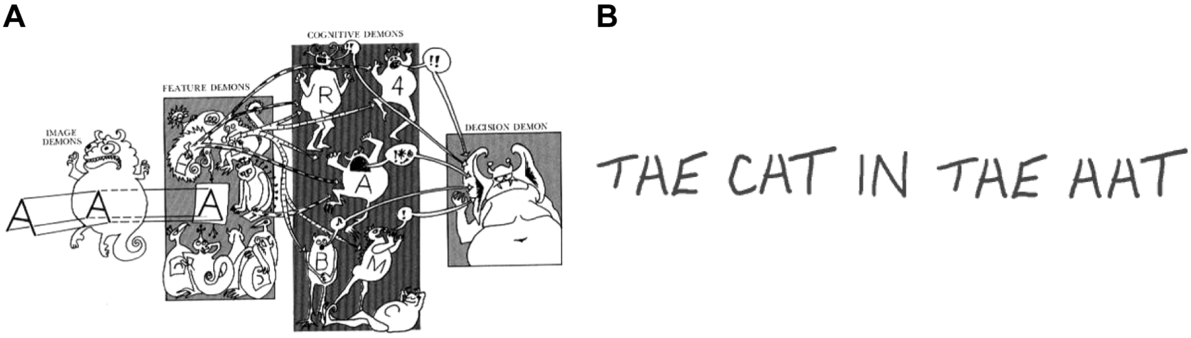


Figure 1.15: Pandemonium model of hierarchical processing and its fundamental drawback. **(A)** Pandemonium and letter recognition. The main principle of this feed-forward architecture lived on in a great number of successive neuronal modeling work. The feature demons analyze low-level visual features like oriented edges. The cognitive demons are selective to specific arrangement of edges and shriek if the preferred letter seems to be in the input. The decision demon picks out the one who shrieks loudest, making the decision about the outcome of recognition. **(B)** The purely feed-forward processing employed in the Pandemonium architecture will fail if trying to interpret ambiguous letters in this phrase. Different interpretation as "A" or as "H" can be done correctly for the same stimulus only if taking into account the global context of the word the local letter is part of. The cognitive demons have no opportunity to talk to each other (lateral feedback exchange) or to receive shrieks back from decision demon (top-down feedback), being thus unable to make a valid decision about ambiguous parts of a whole in the absence of contextual support.

This current state is somewhat at odds with bold predictions made at the beginnings of the artificial intelligence research in the 50's, when full solution of the vision problem was expected soon. Remarkably, models proposed then already contained the basic capabilities of learning and even hierarchical processing [Rosenblatt, 1958, Selfridge, 1958, Minsky, 1961]. For instance, in Selfridge's Pandemonium model [Selfridge, 1958], distributed processing units, the "demons", were embedded in a hierarchical feed-forward structure (Fig. 1.15). The responses of the units to the inputs (demons' shrieks) propagated from bottom to the top of the hierarchy where a decision demon was capable of choosing one of the cognitive demons that shrieked the loudest, implementing thus a winner-take-all (WTA) selection procedure. The units were also able to modify the strength of their connections according to their own activity, the activity of the sender unit and a training signal, amounting to a supervised learning procedure. It seemed that many important ingredients were already in place to begin with construction of a basic cortex-like learning machine for object recognition, and still, more than 50 years later, not even this basic functional model is in reach. So, the demons were hiding in the details as usual.

In spite of this, it cannot be said that no significant advances were made in the field of artificial vision. Impressive success was achieved especially in the tasks where the perceptual situation was restricted to a well-defined setting with very low uncertainty, for example in case of visually guided inspection of routine assembly processes running on a conveyor belt or in the case of autonomous car control on an empty highway. However, it remained symptomatic for most systems that they require a great amount of supervision by human operators and cannot learn autonomously from sensory streams of natural complexity in generic situations.

Some of the demons that kept the progress from going were identified by the scientific community. It has been realized for instance, that learning suffers generally from the curse of dimensionality, or combinatorial explosion, if operating on the raw sensory data (like light intensity values) without appropriately converting it into a more suitable compact representation [Geman et al., 1992, Fiser and Aslin, 2005]. Learning from raw sensory data is absolutely intractable, because the number of free

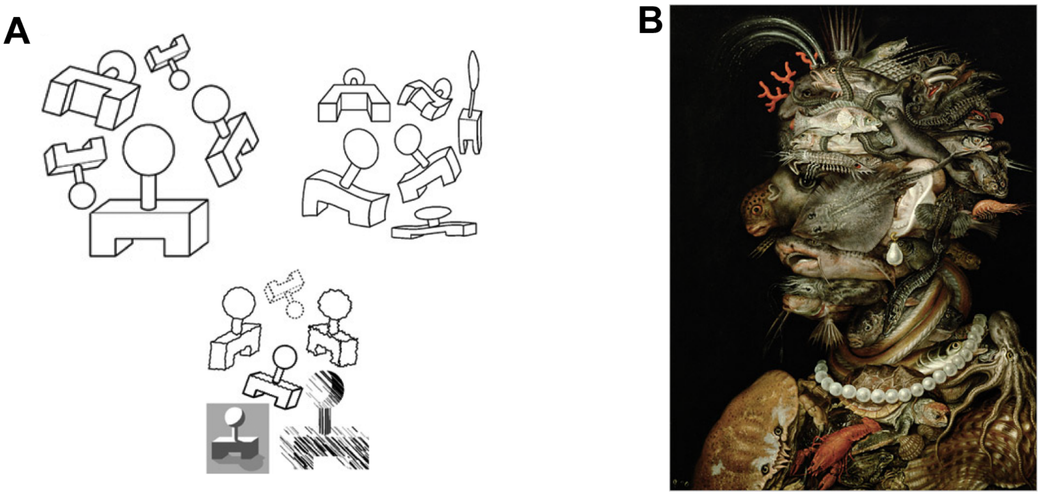


Figure 1.16: Invariance problem in vision **(A)** The same object can undergo transformations that changes its image, the recognition process has still to determine the correct identity. This is very difficult if the object representation does not provide explicit ways to treat transformations and changes they induce in the local appearance. **(B)** Some transformations may completely change the original local appearance, while the global object identity (here a face) stays still the same. Giuseppe Arcimboldo. Water. 1566

parameters, or dimensions, describing each data sample often exceeds by far the number of examples the learner can hope to get to see in a life time. Another classical problem is posed by the fact that the natural sensory data in its raw form, like for instance the retinal activation pattern, looks never the same for the same object identity under different viewing conditions. Already a simple translation of the object changes the raw sensory signal drastically, without even mentioning the effects of other more sophisticated transformations that may occur to the image of the object in the real world setting. So, the question arises how to construct an object representation which is invariant to the natural transformation and at the same time can detect them in the sensory image. This invariance problem is in general one of the hardest issues in object recognition (Fig. 1.16).

These problems gave rise to more advanced methods for processing the raw image data. Feature extraction methods, like for instance Gabor filtering [Daugman, 1985, Turner, 1986, Jain and Farrokhnia, 1991, Lades et al., 1993, Wiskott, 1995] or scale-invariant feature transform (SIFT, [Lowe, 1999, 2004]) were developed to transform the raw pixel image into a number of useful descriptors, or features, reducing substantially the dimension of the data space where learning has to operate and providing some tolerance against simple transformations. One often refers to a bag of features (BoF) when talking about the extracted features set. Such filtering can be generally understood as operation that chooses a more suitable set of basis functions to span the original space. Searching a good basis that captures the most important properties of the original data is also an underlying procedure for a number of classical dimensionality reduction techniques like PCA and ICA [Jutten and Hérault, 1991, Turk and Pentland, 1991b, Hyvärinen and Oja, 2000, Bartlett et al., 2002].

Many BoF approaches, especially the local Gabor transform, have close resemblance to the hypothetical filtering operation performed by the early visual area V1 [Marcelja, 1980, Jones and Palmer, 1987] (Fig. 1.17). The local Gabor transform and SIFT deliver both a fortunate object representation, decomposing the object into a collection of local features of lower complexity. These features possess certain types of tolerance against image transformations. For instance, Gabor features are invariant against global changes in illumination, small changes in scale and local shifts. SIFT is an even more sophisticated transformation that provides both translation and scale invariance of the descriptors.

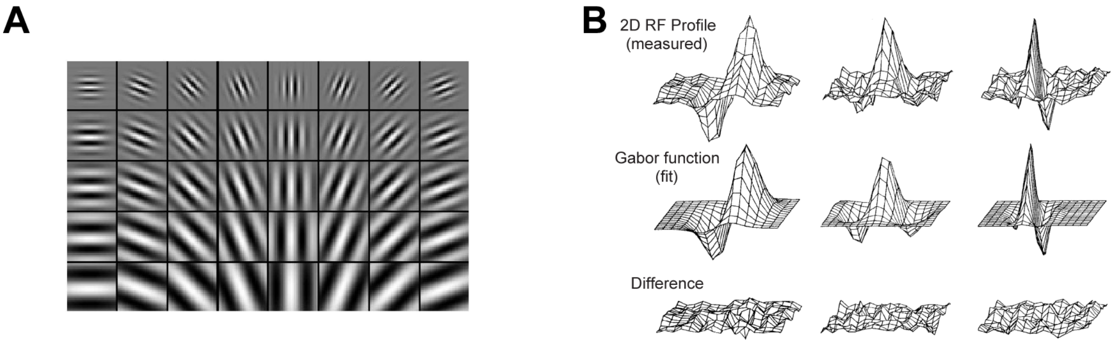


Figure 1.17: Gabor filtering provides low-level visual features for object recognition **(A)** A collection of Gabor filters can be used to transform images into a set of local features suitable for object recognition. Often, only amplitudes of the filter responses are taken, ignoring the phase, which makes the representation robust to local shifts and resembles an operation performed by complex cells of the primary visual cortex. **(B)** Static receptive fields of simple V1 neurons can be coarsely approximated by Gabor functions (Modified from [Jones and Palmer, 1987]).

The BoF approach does not solve the invariance problem though, addressing only some partial issues mentioned before. This is due to fundamental restriction of this kind of object representation, which largely ignores different types of relations between the local features that are inherently present in the visual data. The previously discussed compositional nature of the visual objects is thus not captured by this data structure. Furthermore, many BoF approaches do not involve a hierarchical representation and require to create an unique feature set to store each object for later recognition. The memorized feature sets are thus disjunct, so that elementary features cannot be reused for representing different objects, as opposed to a strategy that relies on an universal feature vocabulary. This is not only a very inflexible method to use memory resources, but it also impairs the ability of the system to generalize over novel data and build categories, as the natural similarity between the stored objects cannot be easily recovered from such representation.

These disadvantages were addressed in a number of extended approaches, two of which are of particular interest here. The first (HMAX architecture, [Riesenhuber and Poggio, 1999]) is a prominent representative of hierarchically organized feed-forward neuronal architectures in the tradition of previous classical work [Rosenblatt, 1958, Selfridge, 1958, Fukushima, 1980]. It is of special interest to look at this extended design because it demonstrates on the one hand how impressively far one can get with that type of rather simple neuronal architecture if using an universal feature vocabulary and hierarchy for the compositional object representation, while on the other hand showing the old problems still persisting. The second approach (elastic graph matching, EGM, [Lades et al., 1993, Wiskott et al., 1997]) has its roots in machine vision and has been used successfully in various real world applications like face recognition, holding state-of-the-art status there for a long time so far. This approach is remarkable, because it not only employs universal vocabularies of reusable features, but also introduces a simple kind of topological relation between them.

The original HMAX architecture was proposed quite recently. Its initial version used a static vocabulary of handcrafted low-level features to enable object recognition [Riesenhuber and Poggio, 1999, 2000]. The extension of the system which is discussed here is able to learn this low-level vocabulary from natural images [Serre et al., 2007b,c]. The architecture is an example for a wide class of neuronal multi-layered networks with purely feed-forward connectivity, that is, a strictly unidirectional signal flow propagating the responses of the units from bottom to the top of the hierarchy (Fig. 1.18).

The main novelty of the HMAX architecture was the feed-forward alternation of two basic pooling

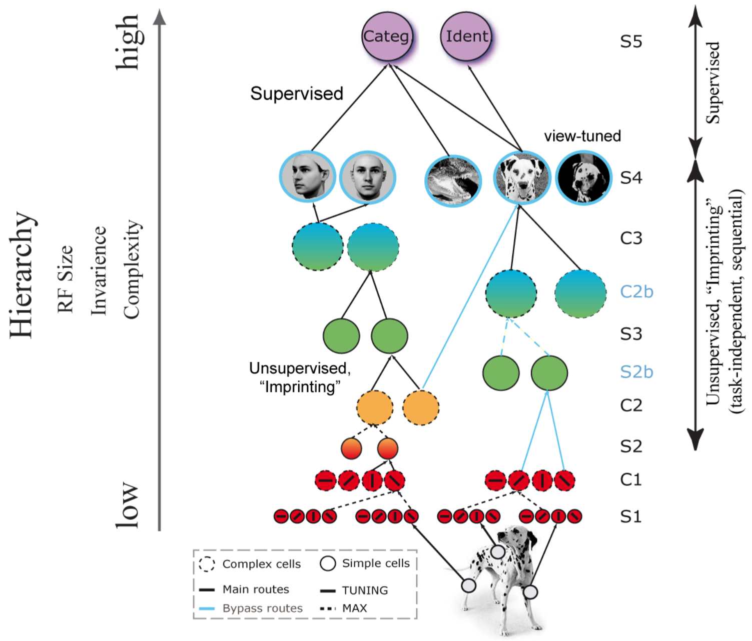


Figure 1.18: HMAX architecture. The hierarchical feed-forward processing uses alternating weighted sum and maximum pooling operations to achieve increasing complexity in the selectivity and increasing invariance in the response of the units. Unsupervised learning for the lower stages is done in sequential manner, training one stage at a time on all the image data from the training set. This procedure uses "imprinting" of unit receptive fields, setting them to a current randomly selected image patch and then freezing, without providing an incremental rule for synaptic modification. For the highest hierarchy level, where the categorization and identity units are residing, supervised learning with corresponding class and identity labels is necessary to tune the receptive fields accordingly. (Modified from [Serre et al., 2007b], with permission.)

operations over the responses of the units : the weighted sum and the maximum. These operations are performed on subsequent S and C layers, or stages, of the network (the naming is borrowed from the classical work on the Neocognitron architecture [Fukushima, 1980]). The layers are composed of units that are dedicated to processing of progressively complex visual features with increasing hierarchy level. The hierarchy starts with the lowest layer S_1, where the units signal the responses of local image Gabor filtering operation performed on different scales. The next layer C_1 contains units each pooling over a retinotopically defined neighborhood from the preceding S_1 layer via a max operation, selecting only the strongest response from the small set of S_1 units they are looking at. The units of the C_1 layer come to possess thus larger receptive fields than S_1 units, already acquiring small degree of tolerance for shift and scale. The next layer S_2 pools over C_1 via a weighted sum operation, and the C_2 uses again the maximum operation to pool over responses of S_2. The processing continues in the same fashion until the top hierarchy level is reached. There, the layer S_4 contains view-tuned units that are selective for a particular target object from the image database used for learning. Building upon S_4 there is an additional layer of categorization units, which are linear classifiers pooling over object-specific units from S_4 to represent a specific category, like planes or faces.

The system has still great impact on the computational neuroscience community, because it was indeed able to learn and show impressive recognition performance comparable to manually constructed state-of-the-art machine vision systems on a wide set of natural images. The architecture employed object decomposition by a hierarchy of visual elements of different complexity, establishing a universal vocabulary that could be shared among different objects. However, a number of fundamental weaknesses remained. One crucial limitation was that processing and learning were restricted to feedforward connectivity only. The relations between the visual elements of the hierarchy were captured only implicitly in the bottom-up connectivity of the units sensitive to specific conjunction of the elements from the immediate preceding layers. During recognition, there is consequently no way to interpret locally ambiguous information about visual features on lower layers in a context-dependent fashion, that is, by taking into account the evidence mediated by the responses of the units on the higher or on the same level of the hierarchy. In many situations, this inability leads to accumulation of locally ambiguous responses that cannot be resolved further, leading to wrong interpretation not only of the corresponding features, but eventually also of the whole object. Furthermore, there is no way to correct wrong local decisions once they are made, because the signals required for the potential refinement of the initial hypothesis are not able to travel back, lacking recurrent connectivity in the network. These disadvantages may be the reason why the system could discriminate well only between highly distinct objects (e.g horses vs. planes vs. faces).

Another weakness is the learning procedure itself. First, it lacks flexibility and neurobiological plausibility in many points. The learning of the low-level feature vocabulary is done by a procedure termed imprinting, which just sets synaptic weights of the units to Gabor responses from randomly picked local image patches and then freezes the weights, without modeling how the weight modification may be implemented in realistic, ongoing incremental fashion [Serre et al., 2007b]. Further, the learning has to be done for each S network layer separately in sequential manner. During a single learning stage, a given layer is modified being exposed to all the images from the training set while the rest is fixed. This sequential procedure starts with the lowest layer and finishes at the top. Again, the ongoing incremental nature of the biological learning and the resulting flexibility are missing there. This kind of difficulty persists also in other related approaches, like convolutional networks [Ranzato et al., 2007] or deep belief networks [Hinton et al., 2006]. Moreover, concerning the degree of supervision, the learning is only unsupervised for the lower S stages. The top categorization stage still requires the labels of the identity of the objects from the presented images.

This being said, the approach has to be acknowledged for its proximity to the important principles of cortical processing and the demonstrated functionality on the natural image data. Most researches agree meanwhile that this architecture can be only a model for the initial recognition process that creates a first coarse interpretation of the stimulus by a rapid feed-forward sweep through the processing hierarchy [Serre et al., 2007a,c]. For the successive refinement of this initial hypothesis, the recruitment of the recurrent lateral and top-down connectivity that is massively present in real cortical networks becomes necessary [Lamme and Roelfsema, 2000, Delorme et al., 2004].

The notion that an explicit preservation of the spatial relationships between the visual elements, of which natural objects are composed, can greatly enhance the capability to perform recognition was used in the elastic graph matching algorithm (EGM, [Lades et al., 1993, Wiskott et al., 1997]). There, the data structure taken to represent an object is a graph (Fig. 1.19). The nodes of the graph may carry a number of local features, or jets, while graph edges capture the topological relations between the local features in the nodes. Two types of graph were distinguished in the original approach [Wiskott, 1995]. The model graph was an instantiation of a single specific object, e.g. a face of an individual person. In distinction, the so-called bunch graph served in turn as a generic model for a whole object category. In the face recognition task setting, the bunch graph carried local vocabularies for facial regions (eyes, nose, mouth, etc) in form of representative feature sets on each of its nodes. The edges between the nodes were labeled with the average distances measured between the respective facial features.

Using the bunch graph structure, it was possible to locate a face in the scene and to determine person-related properties, like gender, race, whether the person wears glasses or a beard and so on [Wiskott, 1995, Wiskott et al., 1997]. To identify the person, a subsequent search in the model graph gallery was

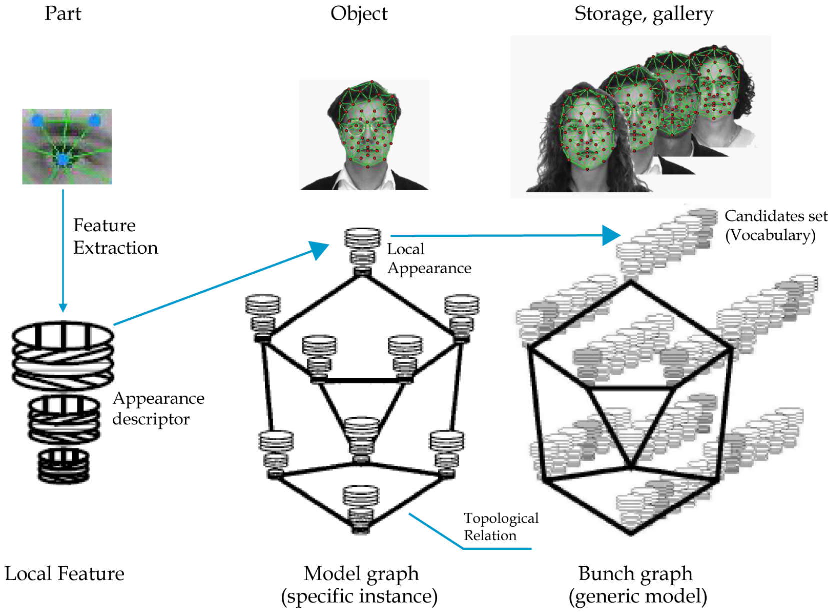


Figure 1.19: Elastic graph matching for face recognition. An individual face is represented by a graph structure, holding local appearance in the nodes and topological relations between the local features in the edges. A model graph captures a face of a particular person, a set of model graphs is the memory domain, or a gallery. A bunch graph is a structure which provides general description of the face class. It contains vocabularies of local features, or candidates, on its nodes, picked manually to deliver a possibly full generic description of local face appearance. The edges carry the average distances between the local features. Bunch graph can be used to detect a face in the image and to identify properties of a particular face, like gender, race and so on. The property identification is done on the basis of the best fit feature candidates selected from each node given a face image. The labels of the winner candidate features determine then per major vote the properties of the presented face. For the person identification, a sequential search through the model graph gallery becomes necessary, where each person is stored separately in a dedicated model graph. (Adapted from [Wiskott et al., 1997])

necessary. Importantly, the similarity function used for the localization of the face in the image involved not only the similarities between the local feature sets and the Gabor transformed image patches, but also the degree of the topological match of their arrangements. The resulting similarity landscape put on the image exhibited clear unambiguous extrema in the regions where face images were positioned, because any dubious correspondences between local model and the image features could be ruled out by the topological term.

This approach and its extensions turned out to work very well in different object recognition settings [Wiskott et al., 1997, Kotropoulos et al., 2000, Loos et al., 2003, Westphal and Würtz, 2009]. However, with respect to learning, major unresolved problems remained. The local feature vocabularies were handcrafted by picking suitable features manually from face images. The same applied for the edges between the nodes. Moreover, the topological constraint was the only kind of contextual support employed for the recognition. The relations between different vocabulary elements in terms of being parts of the same object identity were ignored, so that learning of a specific object was basically creating a disjoint model graph with the object features on the nodes. Each object was thus stored as a separate collection of dedicated features, without the possibility to share features between different memorized objects. The universal feature vocabularies in the bunch graph were used only for face detection, but not for the individual face recognition. This happened in the later phase by going sequentially through all the model graphs stored in the memory (gallery) and comparing them with the graph extracted from the

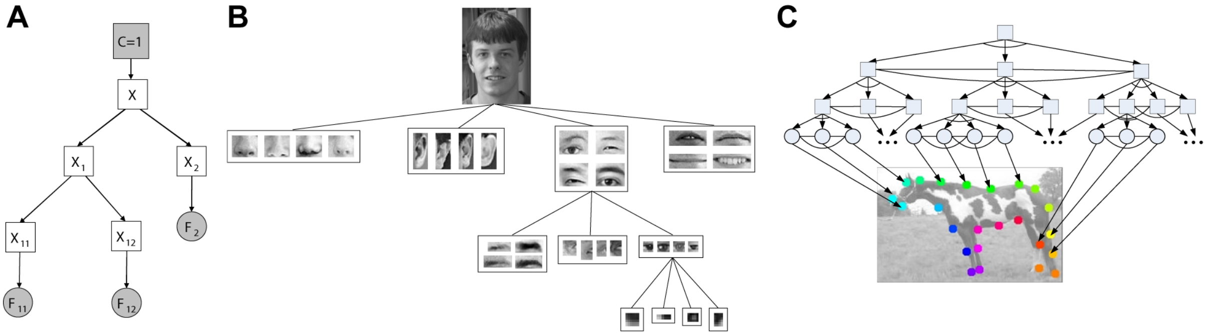


Figure 1.20: Hierarchical graphical models for object recognition **(A)** A simple hierarchical model, where spatial relations between the nodes of the same level are not explicitly incorporated. C is the class node, X denotes the entire object, X_i the object parts and F_{ij} the observed image features. Given the observed features, the inference procedure can compute efficiently the most likely values of the X and X_i. **(B)** An example of a compositional parts-whole hierarchy for a face object. Each part node contains a set of possible appearances. (Taken from [Epshtein and Ullman, 2007], with permission.) **(C)** A hierarchical graphical model with explicit spatial relations between the nodes. (Taken from [Zhu et al., 2010], with permission.)

position located on the image in the first phase. In summary, the graph data structure was proven to be a versatile and powerful object representation in real-world applications, leaving the issue of learning largely open.

The success of both approaches in performing object recognition on real images was encouraging. At the same time it was obvious that they both still suffered from what I may term *representational poverty*. The representational poverty (RPV) is a fundamental inability of the employed object representation to provide essential information that is crucial for proper handling (memorizing and recognizing) of visual objects of natural complexity. This property can be called fundamental in the sense that it cannot be removed by improving the object representation via learning, because of the severe limitations put on the form of the representation initially. RPV is clearly present in the HMAX feed-forward architecture, as it cannot possibly capture the relations between the elements from different vocabularies explicitly in the absence of lateral and top-down connectivity. RPV is also existent in the original bunch graph structure, where it is not specified how to link the elements from different vocabularies together that belong to the same object. RPV can be found in many other popular approaches to object recognition, like for instance in Eigenface method based on principal component analysis (PCA) of the face image space [Turk and Pentland, 1991a]. There, the initially chosen representation allows each face to be interpreted as a linear combination of holistic components learned from the image set, so it becomes fundamentally impossible to treat faces naturally as composition of their local parts.

It is of course very hard business to avoid RPV while at the same time keeping the learning procedure tractable. A number of recent methods originating in probabilistic graphical modeling attempt to incorporate a rich compositional object representation for recognition while keeping inference and learning of such a model feasible [Fergus et al., 2003, Fei-Fei et al., 2006, Ommer and Buhmann, 2006, Epshtein and Ullman, 2007, Fidler et al., 2009, Ommer and Buhmann, 2010, Zhu et al., 2010]. In a hierarchical graphical model, the compositional nature of visual objects can be naturally described by specifying a hierarchy of conditional probability densities for smaller components given the more complex components which the smaller components are parts of. The components themselves are represented by the nodes of the model. The different relations, spatial and hierarchical, between the components are held in the connections between the nodes (Fig. 1.20).

This rich compositional description results in loops in the graphical model, which for a long time were thought to be prohibitive for an efficient inference about the components and object identity given an image (or image features) to interpret [Pearl, 1988, Cooper, 1990]. Fortunately, there exist now belief propagation algorithms that allow the approximative inference to be performed efficiently despite the presence of loops in the graphical model [Frey and MacKay, 1997, Murphy et al., 1999, Aji and McEliece, 2000, Yedidia et al., 2000, Kschischang et al., 2001]. Such belief propagation involves typically message passing between the nodes, where messages are the vectors of probability values generated from the nodes given the available evidence. The computation performed on the messages arriving at nodes may be either of sum-product or max-product type, depending on the employed algorithm (belief update or belief revision, respectively, [Weiss, 2000, Kschischang et al., 2001]). Loopy belief propagation leads to a fast convergence of the computed values in the nodes to the approximation of posterior probabilities for the components and the object identity given the observed image data.

While the algorithms for inference in hierarchical graphical models with loops seem to reach an applicable state, the approaches to learning such a model from natural image data are still deep in development, and are currently a matter of ongoing research in machine learning [Fei-Fei et al., 2006, Ommer and Buhmann, 2010, Zhu et al., 2010]. Remarkably, there are also attempts to map the operations employed in general graphical models onto the operations performed by neural circuits. This enterprise seems to be still very much in its infant state, with some noticeable progress recently [Deneve, 2008, George and Hawkins, 2009, Litvak and Ullman, 2009]. One of the most daunting problems is again the issue of the learning mechanisms. In particular, it is largely unclear what neural operations have to work with what plasticity mechanisms to instantiate structure formation in a recurrent network that would support learning of generative, compositional object representations and create a universal memory domain for arbitrary natural visual objects. Any progress towards answering this question would move us a bit closer not only to understanding the learning in circuits of the brain, but also to development of novel technical systems able to mimic this capability. Such systems would be able to solve complex tasks in vision and other related settings of natural signal processing that are not tractable so far.

1.3 Objectives and thesis overview

The aim of this thesis is to analyze what neuronal mechanisms may underlie the self-organization of a memory domain for compositional object representation in the visual cortex and to present a functional proof in form of a self-organizing neuronal architecture based on the postulated mechanisms. The network architecture will have to demonstrate its functionality by memorizing a substantial number of persons presented from database of natural face images in incremental, unsupervised manner. More concrete, following requirements are put on the system in this work:

- The system has to employ full compositional representation of face objects. That is, it has to learn the local vocabularies of reusable parts, the relations between the parts and higher-order symbols for the global face identity simultaneously.
- The compositional representation has to be of generative kind. It should be possible to recreate the full compositional description of a memorized face in terms of all of its parts given only its higher-order identity symbol or a subset of its parts.
- The system has to learn without supervision, that is, without providing any labels about the identity of the faces presented during the learning.

- The system has to form all kinds of connectivity - feed-forward bottom-up, recurrent lateral and top-down - within and between the network layers simultaneously, using only neuronal plausible, generic mechanisms driving activity and structure formation in the network.
- The unsupervised learning procedure has to be self-stabilizing and life-long, that is, no manually defined stop conditions that freeze learning are allowed.
- The system has to be able to retrieve a memorized face within a short time period that is comparable to the amount of time predicted by psychophysiological experiments on ultra-rapid visual object recognition [Keysers et al., 2001, VanRullen and Thorpe, 2001, Kirchner and Thorpe, 2006].

The thesis is then organized according to the steps undertaken to arrive at the envisaged target :

Chapter 2 introduces a model of a cortical module that will serve as a basic network node in the full architecture. The model is based on basic assumptions about the canonical operations instantiated by neuronal mechanisms of the cortical microcircuit. The local computation performed by the module is the competitive, WTA-like selection and gradual amplification of a small unit subset on the basis of the signals incoming from different hierarchical origins (bottom-up, lateral and top-down). The competitive computation is executed in ongoing gamma rhythm cycles. Equipped with the adaptive mechanisms of bidirectional synaptic plasticity and homeostatic regulation of unit activity, the single module is already able to perform various unsupervised learning tasks, like creating a local appearance vocabulary if exposed to natural face images. Parts of this chapter appeared in [Jitsev and von der Malsburg, 2009].

Chapter 3 presents the memory network architecture and reports the core results of this work. The network architecture is based on two reciprocally interconnected layers of distributed cortical modules introduced in the preceding chapter. The initial undifferentiated connectivity (feed-forward bottom-up and recurrent lateral and top-down) between the modules has to be learned from natural face images presented incrementally from the database without providing any identity labels. The memory network established in course of learning is then examined in terms of the quality of the resulting compositional face representation and in terms of the recognition performance on the alternative face views not shown before. In addition, different processing properties emerging in the network are discussed. Parts of this chapter appeared in [Jitsev and von der Malsburg, 2009].

Chapter 4 analyzes a sleep-like, off-line network regime, where the network spontaneously generates ongoing activity, performing a kind of memory replay. The functional consequences of this off-line memory reprocessing are assessed by comparing the recognition performance of the network before and after the sleep-like state. Parts of this chapter will appear in [Jitsev and von der Malsburg, 2010].

Chapter 5 summarizes the findings of this study and discusses open problems and promising directions on the way to revealing the learning mechanisms in the brain and implementing them in artificial systems capable of autonomous unsupervised learning and object recognition from data streams of natural complexity.

2

Elementary cortical module : a neuronal model for unsupervised competitive learning

In this chapter, I want to lay down the foundation for a highly scalable memory network architecture that supports efficient, large-capacity storage and rapid robust recall of complex visual content. To do so, I introduce here an elementary processing module which will then serve as a basic network node to rapidly carry out computations on and adapt appropriately to the incoming signals. This module is inspired by the cortical microarchitecture, being a model for a local cluster of neuronal populations, which are thought to be the elementary units of neocortical computation. The module contains a number of subunits, each corresponding to a tightly coupled excitatory neuronal population [Song et al., 2005, Yoshimura et al., 2005, Haider and McCormick, 2009]. In its core, the model builds on the previous work of Lücke, where unsupervised competitive learning was successfully implemented in a single cortical module, or column [Lücke, 2005, 2009]. Here I profoundly modify the previous formulation and introduce novel adaptive mechanisms, making the module suitable for the operation in a hierarchical, multi-layered memory network architecture.

The module performs a flexible winner-take-all-like (WTA-like) operation on the incoming signals, choosing a unit which receives the strongest excitatory drive to be the winner in a frame of a decision cycle. The decision cycle is an atomic fragment of ongoing processing. The selected winner unit is amplified in its activity during the decision cycle, reaching a graded activity state that corresponds to the strength of its afferent inputs. The rest of the units is suppressed in successive fashion, dropping off from their graded activity levels one after another in inverse order of the strength of their afferent inputs. This operation corresponds to the picking out a suitable unit to be a representative for the given input signal, such that module units can be interpreted as a basis, an alphabet or set of codebook vectors for the input space.

In order to show that this basis can be formed appropriately in unsupervised fashion, I provide adaptive mechanisms to implement in the module an advanced form of competitive learning that exploits the WTA-like operation. Those mechanisms are synaptic and intrinsic plasticity. The activity-dependent synaptic plasticity has a bidirectional character, either potentiating, or depressing, or leaving a given synapse unchanged. The kind of modification depends on the level of the postsynaptic activity of the unit and on the state of the sliding thresholds that separate the postsynaptic activity range into

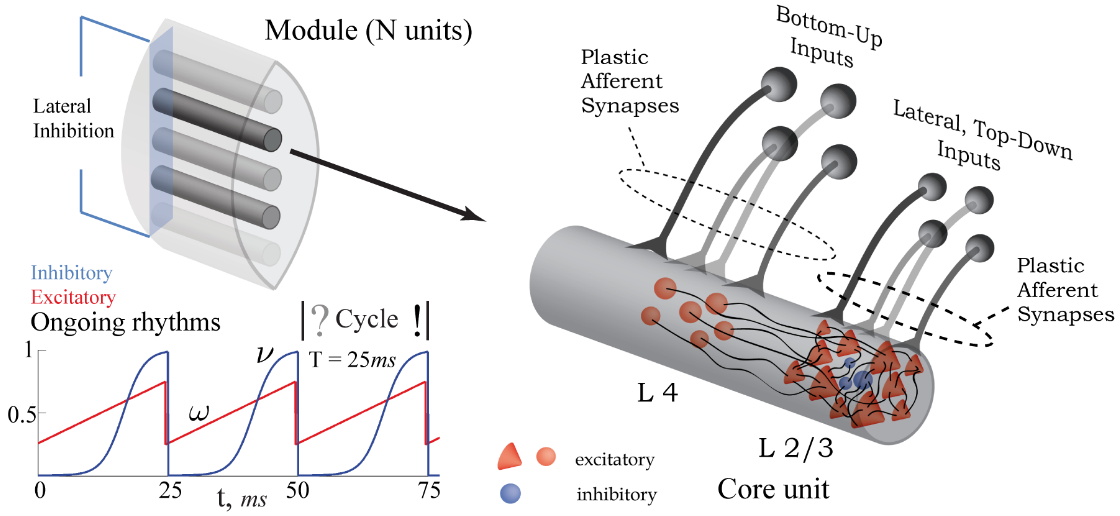


Figure 2.1: A schematic illustration of the cortical module. Each unit within the module represents a population of specifically and tightly coupled excitatory pyramidal neurons that receive common excitatory afferents. The afferents connect through plastic synapses and form distinct groups according to their functional origin (bottom-up synapses arrive at the cortical layer IV, while the lateral and top-down synapses make their contacts with the pyramidal neurons on the layer II/III). The units compete by lateral inhibition. The strength of the lateral inhibition is modulated, being low at the beginning and high at the end of a decision cycle, by oscillation which is in the gamma range ($T = 25ms$).

corresponding zones. Combined with the WTA-like operation, this kind of plasticity strengthens the synapses that store the features that discriminate input patterns, while attenuating those synapses which convey features common to all patterns from the input space. This “caricaturizing” effect should help to differentiate between input patterns that have to be considered as distinct, despite being highly similar in their raw appearance. Thinking ahead about the role of the module in a multi-layered recurrent network, the plastic synapses are separated into functionally different groups according to their origin within the network hierarchy (bottom-up, lateral, top-down). The intrinsic plasticity modifies the excitability level of a unit, down- or upregulating its excitability depending on the history of its previous activation . This homeostatic activity regulation corresponds to effectively adjusting the *a priori* probability of the units to be a winner of a given decision cycle by taking into account the previous wins. Doing so, the mechanism achieves a uniform usage load across units in the module.

Endowed with these mechanisms, the module is prepared for the needs of robust unsupervised learning if embedded within a hierarchical multi-layered network. It has to accomplish two main tasks as a node in such a network. One task is to find a good representation for the locally arriving signals, performing a type of advanced clustering of the incoming data and forming a suitable basis to span the locally accessible input space. Such representation has to provide a versatile vocabulary of universal elements, or parts, which can be reused again and again for combinatorial encoding and storage of novel complex objects. The other, at least as important task is to establish relations between these local vocabularies, associatively linking those local parts together that are constituents of the same object stored in long-term memory. These associative links between modules within and across the layers of hierarchical network define the global appearance of stored objects, capturing explicitly their compositional identity as a set of associated reusable parts.

Apart from its intended function in the network, already a single, isolated module possesses the full functionality of competitive processing and learning on the input data from natural images. To demonstrate this functionality, I will show how the single module successfully copes with a number of non-trivial unsupervised learning tasks. Furthermore, I will show an interesting connection between the coding scheme that emerges from the module's dynamics and the hypothetical role of the gamma cycle as a frame for the probabilistic coding in the cortex. This coding scheme can be seen as a result of a repetitive winner-take-all algorithm executed in the successive gamma cycle frames. Bearing in mind its main application as a network node, the unsupervised learning capabilities of single isolated module are impressive on their own and provide strong support for the quality and versatility of its functional design.

2.1 Fast neuronal dynamics of a cortical module

We start by describing the internal structure and dynamics of the cortical module, called simply the module in following. A module contains a number of subunits I call *core units* (or simply *units*), which receive common excitatory afferents and are bound by common lateral inhibition (Fig. 2.1, 2.9). Acting as an elementary processing unit within the module, each core unit corresponds to a tightly coupled population of excitatory pyramidal neurons ("pyramidal core") as documented in cortical layers II/III and V [Peters et al., 1997, Mountcastle, 1997, Rockland and Ichinohe, 2004, Yoshimura et al., 2005, Song et al., 2005, Shepherd and Svoboda, 2005, Haider and McCormick, 2009]. The two important mechanisms shaping the unit dynamics are the tunable self-excitation and global lateral inhibition, both modulated by ongoing excitatory and inhibitory background rhythms in the gamma range. Interaction between the unit self-excitation and the lateral inhibition defines the competitive character of the WTA-like operation performed by the module. Moreover, due to the self-excitability, the units are capable of sustaining the activity even in the absence of the afferent input.

A cortical module containing a set of N core units is modeled by a set of N stochastic differential equations, each describing the dynamic behavior of the unit's activity variable p. The activity variable p corresponds to the population rate of the excitatory pyramidal cells ("pyramidal core") that make up the unit. The basic form of the equation, leaving out the afferent inputs and neuronal threshold noise for the time being, follows the previous computational study on a cortical column [Lücke, 2005, 2009] with an essential modification in the lateral inhibition interaction term:

$$\tau \frac{dp}{dt} = \underbrace{\alpha p^2(1-p)}_{\textit{self-excitation}} - \underbrace{\beta p^3}_{\textit{self-inhibition}} - \underbrace{\lambda \nu (\max(\vec{\mathbf{p}}_t) - p) p}_{\textit{lateral inhibition}} \tag{2.1.1}$$

where τ is the time constant, α the strength of self-excitation, β the strength of self-inhibition, λ the strength of the lateral inhibition between the units, $\max(\vec{\mathbf{p}}_t)$ is the activity of the strongest unit in the module. The time-varying parameter ν defines the ongoing background inhibitory rhythm in the gamma range. This rhythm modulates the strength of the lateral inhibition, influencing the competitive interaction between the units within the module. Unlike in the previous formulation, here an important modification of the lateral inhibition term makes the strong units less susceptible to the influence of lateral inhibition. The strongest unit within the module is actually able to escape the global inhibitory influence, as in the case of such a winner unit the interaction term becomes zero. So, even if the lateral inhibition strength is growing, the winner unit remains unaffected and is able to keep or amplify its activity level p_{win} (Fig. 2.2, 2.4 **(B)**, 2.7**(D)**). Choosing appropriate values for the parameters α, β and λ, the population activity values p can be restricted to the natural interval between 0 and 1. This allows both interpretations of the variable as either the population rate or the probability of an arbitrary neuron from the population to generate a spike.

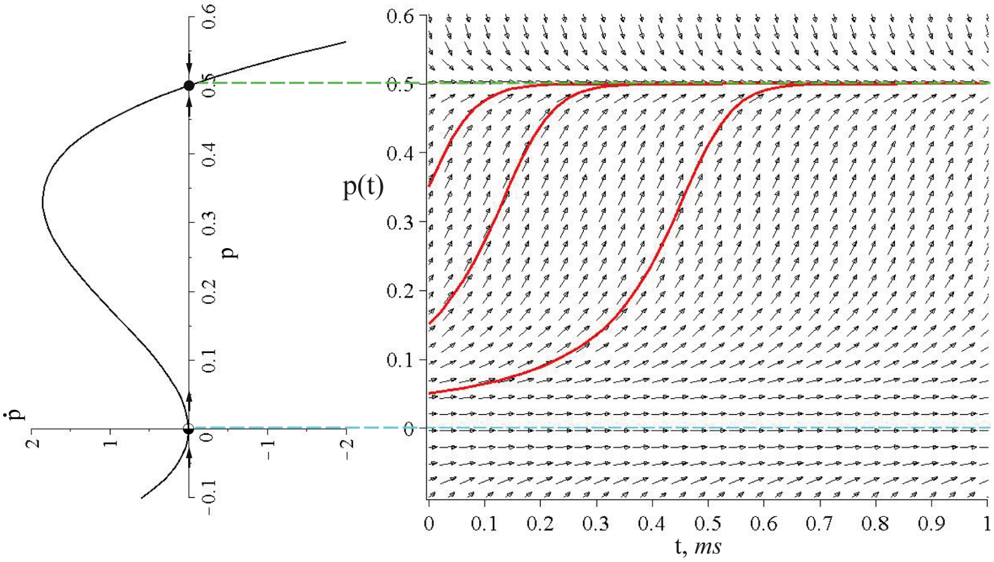


Figure 2.2: Vector field of the winner unit. **Left**: on the line and **right**: the slope field in the (t, p) plane, with solutions drawn starting at different activity levels. The parameters are $\alpha = \beta = 1$, $\lambda = 2$, $\tau = 0.02ms$ (this set will be used throughout this work). If deflected from the semi-stable state at zero, the winner unit follows the flow towards the stable active state at $p = 0.5$. The lateral inhibition modulated by the inhibitory signal ν has no impact on the winner unit activity.

The main components of the dynamics equation can be set in correspondence to the neurophysiological mechanisms that govern the activity formation within the cortical microcircuit. The self-excitation is due to the tight coupling within the fine-scale subnetwork formed by the relatively small population of excitatory pyramidal neurons (in order of hundred cells) on the layer II/III. These neurons are functionally related not only by their tight coupling [Yu et al., 2009, Haider and McCormick, 2009], but also by sharing common afferent input. This common input originates from roughly same subpopulation of spiny stellate cells from the layer IV, as found by local stimulation via focal glutamate uncaging and subsequent recording from the neuron pairs [Yoshimura et al., 2005]. Notice, that being a variable reflecting the population activity of highly specific interconnected pyramidal cells that form a fine-scale excitatory subnetwork, the variable p is *not* a mean-field-like population activity in the classical sense. (see Sec. 2.6 for more elaborate discussion on that).

The different forms of inhibition are mediated by diverse types of inhibitory interneurons populating the horizontal cortical layers. The self-inhibition that balances the effects of self-excitation is most probably provided by fast spiking interneurons residing in layers II/III and IV, like small basket cells or double bouquet cells that target the soma or the proximal dendritic region of excitatory pyramidal neurons [Markram et al., 2004, Burkhalter, 2008]. The lateral inhibition within the module requires inhibitory interneurons with larger axonal extent in both horizontal and vertical direction. This points to large basket cells and Martinotti cells as potential candidates to mediate this form of inhibitory interaction. These inhibitory neurons possess a very expansive axonal arborization, projecting horizontally over millimeter range and contacting excitatory neurons over the full vertical range of cortical layers [Markram et al., 2004, Burkhalter, 2008].

As the formulation of module dynamics is quite hard to comprehend at the first glance, let us take a look on a more accessible situation to grasp the basic properties of the module's dynamical operation. To this end, we assume the module to be in a state where units activities are ordered descendently according to their level, so that it holds $p_{win} > p_i, \forall i : i \neq win, i = 1 \ldots N, win \in \{1, \ldots, N\}$, where N is the number of units in the module and win is the index of the unit with the strongest activity. We can write down the equation for the dynamics of the winner unit:

$$\tau \frac{dp_{win}}{dt} = \alpha p_{win}^2 (1 - p_{win}) - \beta p_{win}^3 \tag{2.1.2}$$

where the lateral inhibition term vanishes being zero (Fig. 2.2). The stationary states of the winner unit are easily obtained by solving $dp_{win}/dt = 0$. Those states are $[\, p_{win} = 0,\, p_{win} = \frac{\alpha}{\alpha+\beta}\,]$. We can use a simple technique of linearizing the system around the stationary points to analyze the stability of these states. Differentiating the winner unit equation gives

$$f'(p_{win}) = \tau \frac{d^2 p_{win}}{dt^2} = p_{win}(2\alpha - 3p_{win}(\alpha + \beta)). \tag{2.1.3}$$

Now, by substituting the stationary states in the differentiated equation we can look on the sign of the result to obtain the stability character of the fixed points. We get $f'(0) = 0$ and $f'(\frac{\alpha}{\alpha+\beta}) = \frac{-\alpha^2}{\alpha+\beta} < 0$ for $\alpha, \beta > 0$. Thus, the non-zero activity state turns out to be stable for the winner unit. The zero activity state is semi- or half-stable, which can be checked by setting $\alpha = \beta = 1$ and plotting the vector field graph of Eq. 2.1.2 (Fig. 2.2). The graph shows that the activities below zero are attracted to the zero state, while activities above zero are repelled away from the inactive state, drifting towards the stable non-zero activity state (which is $p_{win} = 0.5$, given the chosen values of α and β). Note, that the vector field flow of the winner unit is independent of the time-varying parameter ν, which vanishes together with the lateral inhibition term.

Assuming now that the winner unit is in its stable state, we can take a look on the behavior of the weaker units using exactly the same technique. The equation for these challenger units reads

$$\tau \frac{dp_i}{dt} = \alpha p_i^2 (1 - p_i) - \beta p_i^3 - \lambda\nu(\frac{\alpha}{\alpha + \beta} - p)p, \tag{2.1.4}$$

with $i = 1 \ldots N, \forall i : i \neq win$. Here, the lateral inhibition term is present. Again, by solving $dp_i/dt = 0$ we obtain the stationary states, which are $[\, p_i = 0,\, p_i = \frac{\lambda\nu}{\alpha+\beta},\, p_i = \frac{\alpha}{\alpha+\beta}\,]$. In addition to the two fixed points we already saw in case of the winner unit, there is one more stationary state which depends on the time-varying parameter ν (Fig. 2.3, 2.4 **(A)**). Let us again analyze the stability of these stationary states. Differentiating the challenger unit equation gives

$$f'(p_i) = \tau \frac{d^2 p_i}{dt^2} = p_i(2\alpha + \lambda\nu) - 3p_i^2(\alpha + \beta) - \lambda\nu \frac{\alpha}{\alpha + \beta} \tag{2.1.5}$$

Substituting the stationary states $[\, p_i = 0,\, p_i = \frac{\lambda\nu}{\alpha+\beta},\, p_i = \frac{\alpha}{\alpha+\beta}\,]$ into the differentiated equation yields $[\, f'(0) = -\frac{\alpha\lambda\nu}{\alpha+\beta},\, f'(\frac{\lambda\nu}{\alpha+\beta}) = \frac{\lambda\nu(\alpha-\lambda\nu)}{\alpha+\beta},\, f'(\frac{\alpha}{\alpha+\beta}) = \frac{\alpha(\lambda\nu-\alpha)}{\alpha+\beta}\,]$. If we make a general assumption $\alpha, \beta, \lambda, \nu > 0$, we see that the stability of both non-zero activity states is subject to change, depending on the value of the parameter ν :

$$\begin{aligned}
f'(0) &= -\frac{\alpha\lambda\nu}{\alpha + \beta} < 0, \\
f'(\frac{\lambda\nu}{\alpha + \beta}) &= \frac{\lambda\nu(\alpha - \lambda\nu)}{\alpha + \beta} \quad \begin{cases} > 0, \ 0 < \nu < \alpha/\lambda, \\ = 0, \ \nu = \alpha/\lambda, \\ \text{undef. for } \nu > \alpha/\lambda \end{cases} \\
f'(\frac{\alpha}{\alpha + \beta}) &= \frac{\alpha(\lambda\nu - \alpha)}{\alpha + \beta} \quad \begin{cases} < 0, \ 0 < \nu < \alpha/\lambda, \\ = 0, \ \nu = \alpha/\lambda, \\ > 0, \ \alpha/\lambda < \nu \end{cases}
\end{aligned} \tag{2.1.6}$$

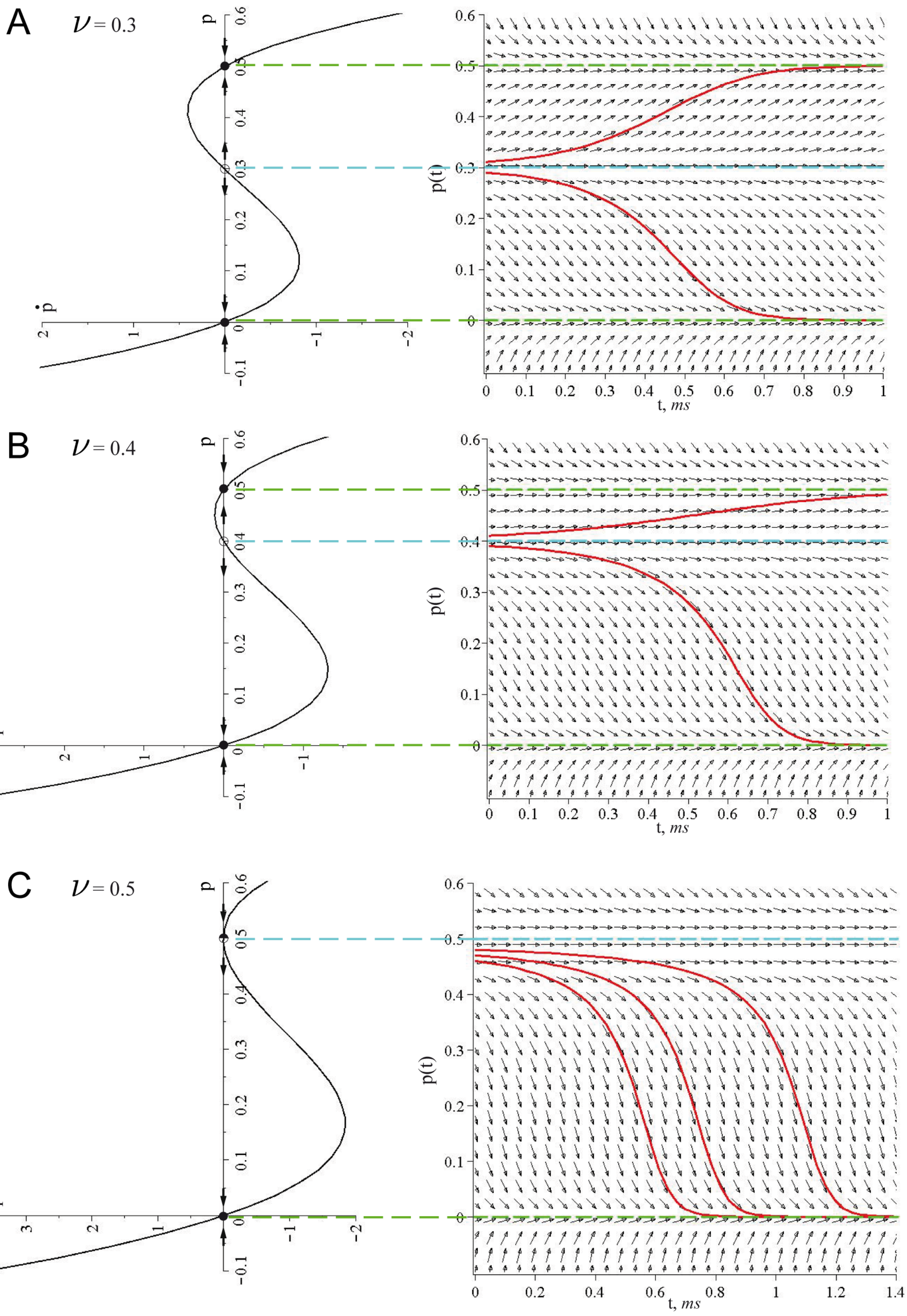


Figure 2.3: Vector field of the challenger unit, using the same nomenclature as in Fig. 2.2. The qualitative behavior of the challenger unit depends on the inhibitory signal ν. **(A)**, **(B)** Below the critical value $\nu_c = 0.5$, the challenger unit can reach the stable active state $p = 0.5$ if its own activity is larger than ν, otherwise it gets deactivated. **(C)** At the critical value bifurcation occurs and $p = 0.5$ looses its stability. Any activity state below follows thus the flow in the direction of the stable zero state and the challenger unit becomes deactivated.

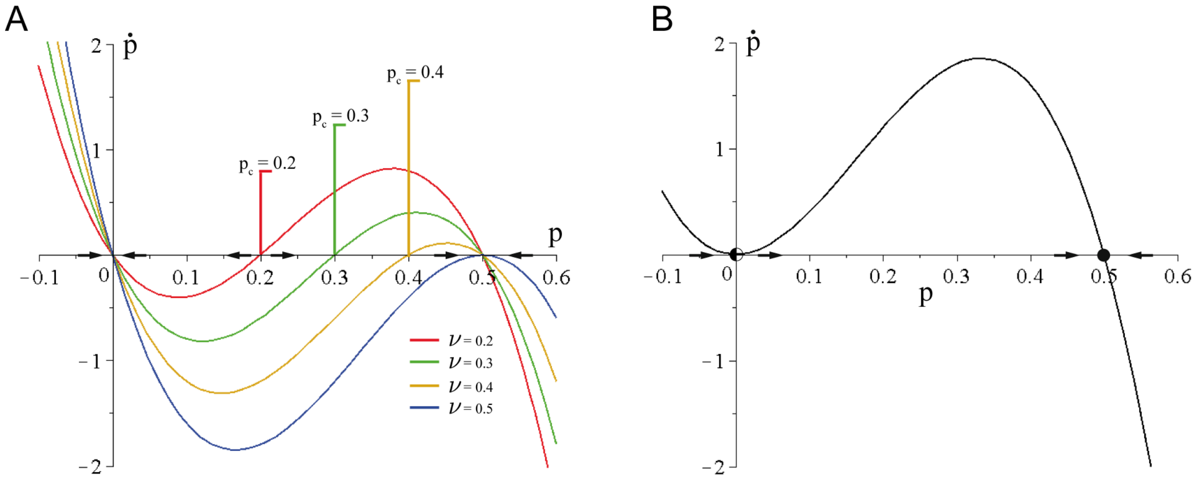


Figure 2.4: Comparing the behavior of winner and challenger unit. **(A)** A challenger unit stays active only if its activity is above a critical level $p = \nu$ that rises with increasing inhibition. At $\nu_c = 0.5$ any challenger unit gets deactivated, except for the symmetric state case where it manages to catch up to the winner unit at $p = 0.5$. **(B)** The winner unit is always able to reach and stay in the stable active state, as it escapes the influence of lateral inhibition (Eq. 2.1.1, 2.1.2).

$(f'(\frac{\lambda\nu}{\alpha+\beta})$ is undefined for $\nu > \alpha/\lambda$ because the state vanishes under that condition violating the assumption $p_i < p_{win}$). So, the system undergoes a bifurcation as the parameter ν crosses a critical point of structural instability ν_c, given by the ratio between the self-excitation and lateral inhibition strength:

$$\nu_c = \frac{\alpha}{\lambda} \tag{2.1.7}$$

To get a clearer picture what happens as ν approaches the critical point, let us set the equation parameters to well-defined values and plot the vector field of the challenger unit for increasing values of ν. I choose here $\alpha = \beta = 1$, $\lambda = 2$ to get the critical point at $\nu_c = 0.5$. Restricting further ν to the interval $(0, 1]$ we get the stability characteristics for the stationary states $[\,p_i = 0,\ p_i = \nu,\ p_i = 0.5\,]$ in form of

$$\begin{aligned}
f'(0) \quad &= -\nu < 0, \\
f'(\nu) \quad &= 2\nu(0.5-\nu) \quad \begin{cases} > 0, \ 0 < \nu < 0.5, \\ = 0, \ \nu = 0.5, \\ \text{undef. for } \nu > 0.5 \end{cases} \\
f'(0.5) \quad &= \nu - 0.5 \quad \begin{cases} < 0, \ 0 < \nu < 0.5, \\ = 0, \ \nu = 0.5, \\ > 0, \ \nu > 0.5 \end{cases}
\end{aligned} \tag{2.1.8}$$

Now, we see that for $\nu < \nu_c$ there are two stable states $[\,p_i = 0,\ p_i = 0.5\,]$, one inactive and one active, with one intermediate unstable active state $p_i = \nu$ that subdivides the activity range in two subregions (Fig. 2.3, 2.4 **(A)**). The activities from the subregion to the left of the unstable state are attracted towards the stable inactive zero state, while the activities from the subregion to the right of the unstable state are driven to the stable active state. This *bistability* persists until the parameter ν reaches the critical point. As ν approaches $\nu_c = 0.5$, the intermediate unstable active state moves away from the zero, moving closer and closer to the stable activity state $p_i = 0.5$. Thus, the left deactivating subregion gets larger and larger, while the right activating subregion becomes smaller and smaller. Thus, to stay active, the challenger units must increase their activity, pursuing the winner unit to catch up at $p_{win} = 0.5$. The units unable to do so become deactivated. Finally, intermediate unstable state and stable active state melt into a semi-stable activity state $p_i = 0.5$ at $\nu = \nu_c = 0.5$. The activation subregion vanishes completely, any activity level below the winner unit is now repelled from the active state and attracted to zero (Fig. 2.4 **(C)**).

This analysis, conducted without taking into account external inputs to the module, already reveals the basic competitive principle behind the computation done by the module. The parameter ν controls the competitive nature of the module's operation. Its critical value ν_c subdivides the computation in two phases. In the first phase ($\nu < \nu_c$), any subset of the core units that manage to resist the growing inhibition can remain active. In the second phase ($\nu > \nu_c$), only one single winner unit is able to stay in the active stable state. For a simple example of a two-units module system, this qualitative change in the system behavior can be visualized in a bifurcation diagram (Fig. 2.6 **(A)**). The activity vector field of the two-units module can be also depicted for different ν values, so that the different attractor landscapes shaped by stable and unstable states of the two distinct phases become apparent (Fig. 2.5).

The extension of the analysis to the general case of N units in the module is straightforward and yields the expected qualitative results about the stability of the system, which undergoes a bifurcation if the strength of lateral inhibition increases and the parameter ν crosses the critical point $\nu_c = \frac{\alpha}{\lambda}$. We can compute the set of stationary states for the following system of differential equations:

$$\begin{cases} \dfrac{dp_1}{dt} = & \alpha p_1^2(1-p_1) - \beta p_1^3 - \lambda\nu(\max(\vec{\mathbf{p}}_t) - p_1)p_1 \ , \\ \dfrac{dp_2}{dt} = & \alpha p_2^2(1-p_2) - \beta p_2^3 - \lambda\nu(\max(\vec{\mathbf{p}}_t) - p_2)p_2 \ , \\ \quad \dots & \\ \dfrac{dp_n}{dt} = & \alpha p_n^2(1-p_n) - \beta p_n^3 - \lambda\nu(\max(\vec{\mathbf{p}}_t) - p_n)p_n \ . \end{cases} \tag{2.1.9}$$

In line with the previous analysis, the stationary states are for $\nu < \nu_c$ of the basic form

$$[\underbrace{(0,\dots,0)}_{\textit{semi-stable}}, \underbrace{(\frac{\alpha}{\alpha+\beta},\dots,0,\dots,\frac{\lambda\nu}{\alpha+\beta})}_{\textit{unstable}}, \underbrace{(\frac{\alpha}{\alpha+\beta},\dots,0,\dots,\frac{\alpha}{\alpha+\beta})}_{\textit{stable}}, \underbrace{(\frac{\alpha}{\alpha+\beta},\dots,0)}_{\textit{stable}}] \tag{2.1.10}$$

and for $\nu > \nu_c$

$$[\underbrace{(0,\dots,0)}_{\textit{semi-stable}}, \underbrace{(\frac{\alpha}{\alpha+\beta},\dots,0,\dots,\frac{\alpha}{\alpha+\beta})}_{\textit{unstable}}, \underbrace{(\frac{\alpha}{\alpha+\beta},\dots,0)}_{\textit{stable}}] \tag{2.1.11}$$

including arbitrary permutations and different multiplicity within the state vectors. Now we again use the linearization of the system around the stationary states to probe for stability of the stationary states. We compute the Jacobi matrix from Eq. 2.1.9 and by substituting the stationary states obtain the

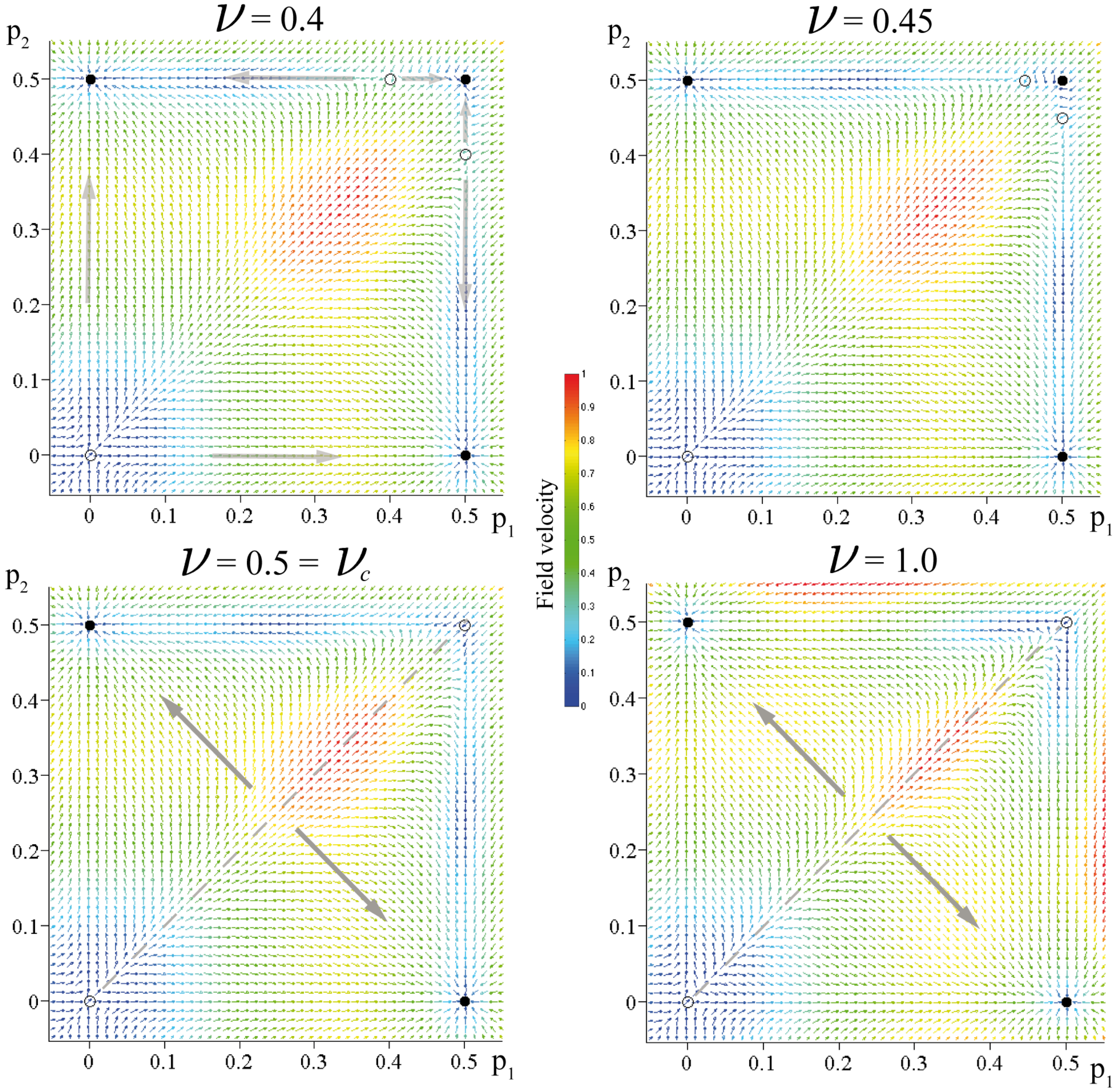


Figure 2.5: Vector field in the phase space of a two-units module and its change induced by the increasing strength of lateral inhibition. With inhibition strength rising, the unstable states $(0.5, \nu)$ and $(\nu, 0.5)$ move closer together and the region in the phase space where both units can stay active shrinks. At the critical point $\nu_c = 0.5$, the approaching unstable states and previously stable state $(0.5, 05)$ coalesce into one unstable state. From then on, the only stable states are $(0.5, 0)$ and $(0, 0.5)$, where only one of the units can remain active.

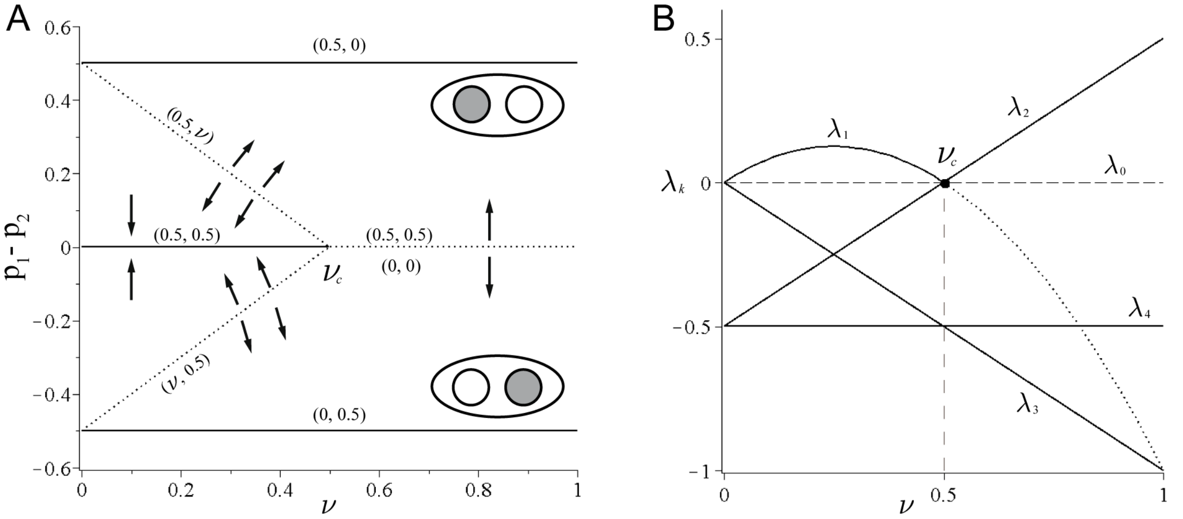


Figure 2.6: **(A)** Bifurcation diagram for a module with $N = 2$ units. Before the critical point is reached, both units can be in active state. After crossing $\nu_c = 0.5$, the only stable states are those with one active and one inactive unit. **(B)** The eigenvalues of the differential equation system for the generic module of N units. The eigenvalue signs depend on the value of ν. At the critical point ν_c the eigenvalue λ_2 changes its sign from negative to positive and the eigenvalue λ_1 vanishes. This modifies system's attractor landscape and determines the qualitative change in module's behavior, allowing only the states with one active unit to be stable.

corresponding eigenvalues :

$$
\begin{aligned}
\lambda_0 &= 0, \\
\lambda_1 &= \frac{\lambda\nu(\alpha - \lambda\nu)}{\alpha+\beta} \quad \begin{cases} > 0, \ 0 < \nu < \alpha/\lambda, \\ = 0, \ \nu = \alpha/\lambda, \\ \text{undef.}, \ \nu > \alpha/\lambda \end{cases}, \\
\lambda_2 &= \frac{\alpha(\lambda\nu - \alpha)}{\alpha+\beta} \quad \begin{cases} < 0, \ 0 < \nu < \alpha/\lambda, \\ = 0, \ \nu = \alpha/\lambda, \\ > 0, \ \nu > \alpha/\lambda \end{cases}, \\
\lambda_3 &= -\frac{\alpha\lambda\nu}{\alpha+\beta} < 0, \\
\lambda_4 &= \frac{-\alpha^2}{\alpha+\beta} < 0
\end{aligned}
\tag{2.1.12}
$$

At the critical point $\nu_c = \frac{\alpha}{\lambda}$, the eigenvalue λ_2 changes its sign from negative to positive and the eigenvalue λ_1 vanishes (Fig. 2.6 **(B)**). This determines the change in the attractor landscape and in the stability of the remaining stationary states. Before crossing the critical point, there are $3^N - 2^N + 1$ stationary states in total, with $2^N - 1$ of those being stable. After crossing the critical point, only N of the remaining 2^N stationary states are stable. Due to the subcritical bifurcation, $2^N - N - 1$ stationary states, which were stable before allowing arbitrary subsets of units to be active, loose their stability during the phase transition. The remaining N stable states are only those which contain only one active winner unit, the rest being deactivated.

Now, in order to make the two-phase winner-take-all computation input-sensitive, we need to specify how the external afferent inputs enter the module. We also have to specify the form and time course for the ongoing excitatory and inhibitory rhythms which will then define the decision cycle, an atomic fragment of processing and learning. For the full formulation of module's dynamics, we take from now on $\alpha = \beta = 1$, $\lambda = 2$, which gives $\nu_c = 0.5$. The qualitative dynamical behavior stays the same in this extended formulation:

$$\begin{aligned}\tau\frac{dp}{dt} = \; & \underbrace{\omega\;\overbrace{(1+I^{LAT}+I^{TD})}^{\textit{Lateral and Top-Down Afferents}}\;p^2(1-p)}_{\textit{self-excitation}} - \underbrace{\beta p^3}_{\textit{self-inhibition}} - \underbrace{2\,\omega\nu(\max(\vec{\mathbf{p}}_t)-p)p}_{\textit{lateral inhibition}} \\ & + \underbrace{I^{BU}p^2}_{\textit{Bottom-Up Afferents}} + \underbrace{\theta p}_{\textit{intrinsic excitability}} + \underbrace{\sigma\eta_t p}_{\textit{threshold noise}} + \underbrace{\omega\epsilon}_{\textit{unspecific excitation}}\end{aligned} \tag{2.1.13}$$

This is now the formulation of module dynamics I will use from now on throughout the thesis, exploiting its favorable properties in the operational regime within a multi-layered hierarchical memory network that will be introduced in the next chapter. At this point, let us focus on the important mechanisms of the full dynamics I haven't yet discussed before. These mechanisms are designed with intention to assure the module's capability to perform computation and learning appropriately being embedded in the network.

One important feature of the dynamics concerns the way how the afferent inputs (denoted with I^{BU} for bottom-up, I^{LAT} for lateral and I^{TD} for top-down afferent input) contribute to the unit's activity (Fig. 2.9 and Fig. 2.1). First, the afferent inputs enter the dynamic equation in multiplicative fashion, being modulated by the activity of the unit. This is different from the previous approach, where the afferent input was additive with a very small coupling strength, serving as perturbation to the internal dynamics of the system [Lücke, 2005, 2009]. This difference is crucial, as modifying the input injection method changes the module's coding scheme that has to reflect the strengths of the incoming inputs. As we will see, the module becomes able to encode the input strength in the time points of successive unit deactivation *and*, importantly, in the graded activity of the units (Fig. 2.7 (**D**), 2.8). This can be done because the multiplicative contribution of the afferent input to the dynamics encodes the input value directly into the activity value of the stable state of the unit, instead of just deflecting the unit activity a bit from its stable state via additive perturbation.

Second, thinking in advance about the role of the module as a node in a multi-layered recurrent network, the afferent inputs are separated in functionally different groups according to their origin within the network hierarchy. This synaptic separation causes functionally different inputs to have different impact on the activity of the unit. The functional difference can be made explicit by taking a glance at the stable state of the winner unit (assuming for clarity $\sigma = \epsilon = \theta = 0$), which takes the value

$$p_{stable} = \frac{\omega(1+I^{LAT}+I^{TD})+I^{BU}}{\omega(1+I^{LAT}+I^{TD})+1}, \tag{2.1.14}$$

As mentioned before, the afferent inputs contribute directly to the graded activity level of the stable state. However, the way how they contribute to it is different. The bottom-up (BU) input I^{BU} contributes to the activity level in a linear fashion, while the contribution of lateral (LAT) and top-down (TD) inputs (I^{LAT} and I^{TD}) to the unit activity is non-linear. The functional role of LAT and TD afferents is opposed to the purely driving character of BU input by using the lateral and top-down inputs for the modulation of the self-excitation term in the unit's dynamics (Eq. 2.1.13). The stronger the input from LAT or TD afferents, the stronger is the self-excitatory coupling within the core unit. Given a sufficient bottom-up input, this modulation potentiates the unit to amplify its activity stronger and faster than the units with lower self-excitatory coupling strength, thus favoring it in the competition for the right to stay active. The separation done here accounts for the different types of information carried by the signals that arrive via respective afferents. The bottom-up signals carry ambiguous and therefore highly uncertain partial information about the stimulus. For stimuli of natural complexity, this uncertainty can be only resolved correctly if the local decision making can in addition rely on many of context-dependent cues. These cues, mediated by the lateral and top-down signals, are the result of learning through previous experience with the stimuli, allowing fusion of local incomplete, ambiguous information into a global coherent stimulus interpretation. The proper integration of this contextual signaling into the local decision making depends critically on the correct treatment of the incoming signals according to their functional meaning. Putting all the different afferents simply together in one synaptic cluster would destroy the proper cue fusion and thus result in breakdown of correct stimulus interpretation.

Remarkably, this kind of synaptic separation is classically evident in the cortical circuits. There, the incoming afferent nerve fibers make intricate synaptic contacts across the whole range of the cortical layers, contacting the target neurons at different sites of the dendritic tree and soma [Spruston, 2008]. Those contacts show very high specificity towards the layer where they arrive at and towards the site on the target neuron they make contact to, dependent on the origin of their presynaptic source. Simply stated, the separation of incoming synapses across the cortical layers follows a generic scheme where bottom-up incoming feed-forward afferents arrive in layer IV on the spiny stellate cells, while the vast majority of LAT and TD feed-back synapses contacts the apical dendrites of pyramidal neurons from the layers II/III and V [Felleman and Essen, 1991, Douglas and Martin, 2004, Thomson and Lamy, 2007, Larkum et al., 2009]. It seems that the cortical organization also takes into account the necessity to separate synapses carrying signals of different hierarchical origin into segregated groups.

As pointed out before in this section, the crucial property of the module dynamics that makes it act like a flexible WTA-like decision unit is the bifurcation the system undergoes when the growing lateral inhibition forces all units but the strongest one to shut down. The course of the decision making in the module is thus shaped by the two ongoing rhythms, the excitatory rhythm ω and the inhibitory rhythm ν, as they both modulate the strength of lateral inhibition between the units and the self-excitatory coupling within those. The excitatory and inhibitory rhythms ω and ν are given by:

$$\omega(t) = \omega_{min} + \frac{mod(t,T)}{T}\,(\omega_{max} - \omega_{min}), \tag{2.1.15a}$$

$$\nu(t) = \nu_{min} + \frac{1}{k \cdot e^{-g\,(mod(t,T)\,-\,0.5\,(T+T_{init}))} + (\nu_{max} - \nu_{min})^{-1}}, \tag{2.1.15b}$$

with the period $T = 25\,ms$, which corresponds to oscillation in the gamma range ($\gamma = 40Hz$). The term $mod(t,T)$ denotes the modulo operation, $\nu_{min} = 0.005$, $\nu_{max} = 1.0$, $\omega_{min} = 0.25$, $\omega_{max} = 0.75$ are the lower and upper bounds for the amplitudes of rhythms, and $T_{init} = 5ms$, $k = 2$, $g = 0.5$ are parameterizing the sigmoid curve that defines the time course of inhibition within a single period of the rhythm (Fig. 2.7 **(C)**). Both inhibitory and excitatory rhythms described here may have variety of sources in the brain, the former being possibly generated by the interneuron subnetwork of fast-spiking inhibitory cells [Whittington et al., 1995, Sohal et al., 2009] and the latter having their origin in activities of fast rhythmic bursting, or chattering, excitatory pyramidal neurons [Gray and McCormick, 1996, Cunningham et al., 2004].

The form of inhibitory rhythm is chosen such that each *decision cycle* spanned by the rhythms is subdivided in two phases at the critical point $\nu_c = 0.5$. Each phase possesses a different coding scheme (Fig. 2.7 **(D)**). In the first phase, $\nu < \nu_c$, the inhibition is low and any subset of the core units that manage to resist the growing lateral inhibition can remain active. Those units reach a certain *graded* activity level that encodes the strength of the incoming afferent input. As the lateral inhibition grows approaching the critical point, some of the previously active units drop off from their graded

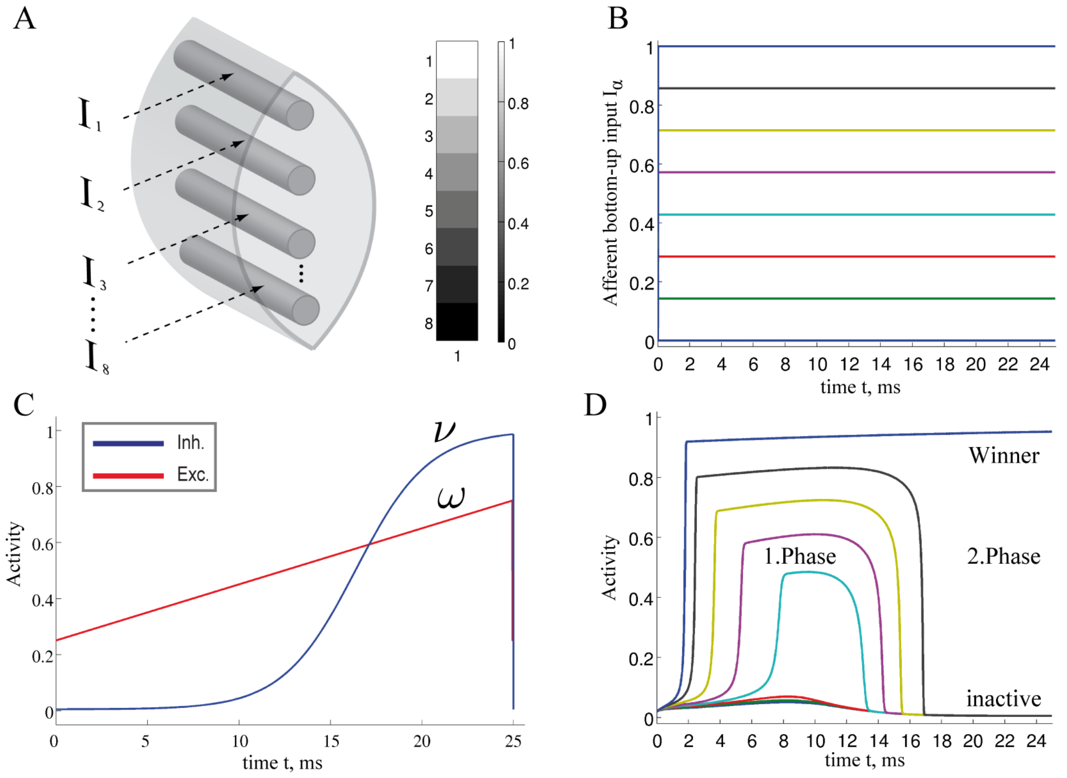


Figure 2.7: An example of 2-Phase WTA coding in a single module. **(A)**, **(B)** Simple bottom-up input in form of linearly increasing scalar values is provided, feeding one value per unit. During the decision cycle spanned by the two signals ν and ω **(C)**, the incoming input is represented by the unit activities p **(D)**. In the first phase the inhibition is low enough so that a subset of units becomes active, representing in a graded fashion the strength of the significantly large input components. In the second phase the inhibition grows stronger, realizing a hard winner-take-all (WTA) behavior, where only the unit with the strongest input remains active.

activity level and get deactivated. The time points of those deactivation events are ordered according to the strength of the incoming input the deactivated units receive. This phase can be thus interpreted as a *selection phase*, where a small number of potential candidates is selected to probe them for the representation of the current input signal. The operation performed by the module in the selection phase can be thus interpreted as soft-winner-take-all (soft-WTA) computation on the incoming inputs. The candidate units provide with their graded activity response an opportunity for the modules in the network to exchange the information about their internal states within and across the network layers. The proper communication between the modules will become crucial later in the network architecture designed in the next chapter. There, it will mediate contextual signaling over lateral and top-down pathways and support the local decision making, guiding it towards a globally coherent interpretation of the current stimulus.

The second phase is the “decision” phase ($\nu > \nu_c$), where due to strong inhibition only one single winner unit remains active within the module, the rest of the units gets suppressed and deactivated. This operation resembles a hard-winner-take-all (hard-WTA) computation. Only the strongest unit from the previously selected candidate set is able to survive, being the most suitable candidate to represent the incoming input signal. The selection of the winner unit can be interpreted as a local decision made by the module on the basis of the accessible information. The winner unit is able to escape the impact of lateral inhibition. In addition, it is supported by the growing excitatory rhythm ω, which further elevates

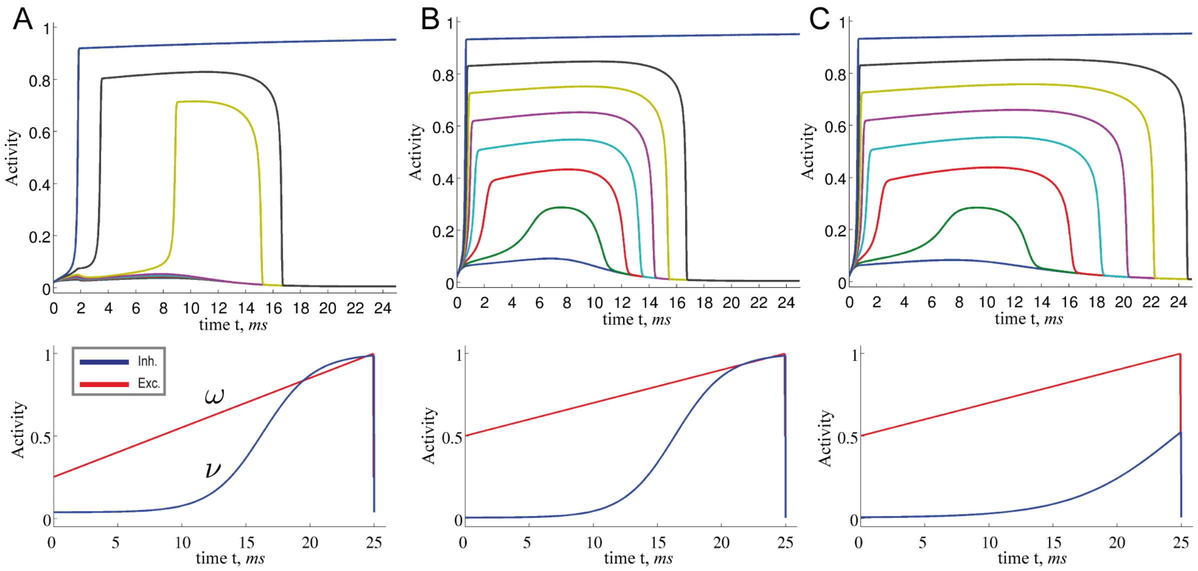


Figure 2.8: Flexibility of 2-phase WTA coding. Tuning the signals ν and ω can increase **(A),** or decrease **(B)** the sparseness, or even change the form and course of the phases **(C)**.

the winner unit, amplifying its activity level. Thus, this second phase can be also called *amplification* phase. The decision about the interpretation of the incoming input is signaled in a very sparse fashion, making its read-out extremely easy. The sparseness and the form of the both phases can be tuned via modification of the ongoing rhythms ω and ν as indicated in Fig. 2.8.

In the mature network connectivity state formed by learning, the contextual influence on the local decision making has to assure the correct interpretation of the global stimulus identity. Importantly, the excitatory rhythm ω modulates not only the strength of the self-excitatory coupling, but also the impact of lateral and top-down signals on the local decision making (see Eq. 2.1.13, 2.1.14). The higher the value of ω, the higher is the influence of the context-dependent signaling on unit activity. Thus, the contribution of contextual information to local decisions is varied from weak to strong by increasing ω in the course of a decision cycle (Fig. 2.7 **(C)**). This is done in order to coordinate properly the impact of different information on the decision making during the cycle course. At the beginning of the decision cycle, the arriving bottom-up signals are given the opportunity to create an initial local hypothesis about the nature of the incoming input. This first estimate does not incorporate exhaustively all the possible contextual biases the system has learned from the experience. This kind of “lazy” computation should help to avoid overhasty interpretations of the incoming signals. Such interpretations would rely too strongly on the structure of the internal representations, without properly considering the alternatives hidden in the actual signals from the outer world. The lazy computation leaves a lot of possible alternatives open how the stimulus could be interpreted, providing a basis for further refinement towards a more elaborate interpretation. The refinement is done by increasing the impact of the contextual signaling on unit activities, which should successively remove those alternatives which are not well compatible with the global contextual support. Close to the end of the decision cycle, the contextual influence is strong and only the most suitable alternative would remain, representing the most plausible interpretation of the current stimulus.

Now, a series of successive decision cycles defined by the ongoing rhythms in the gamma range implements a rapidly repeating, WTA-like decision mechanism. [Zhang and Ballard, 2002, Fries et al., 2007, Börgers et al., 2008, de Almeida et al., 2009]. In each cycle, the incoming input is encoded within the two-phase-WTA coding scheme. At the cycle's end, a winner unit survives the competition and is selected as representative for the given input. To make sure that this operation can keep on going autonomously, it is necessary to give the deactivated units a chance to become active again once they reached the state of zero activity. As all afferent inputs entering the dynamics (Eq. 2.1.13) are modulated by unit activity, a small constant unspecific excitatory drive ϵ is necessary to deflect the deactivated units from the zero state. Note, that the re-activation of the silent units is an intrinsic property of the dynamics. It is *not* necessary to apply the excitatory impulse exactly at the begin of the cycle, the re-activation is performed autonomously by the ongoing rhythms. The tonic unspecific excitation ϵ, modulated by the excitatory rhythm ω, stands for the diffuse excitation which may originate from different brain nuclei that make wide-spread, unspecific projections to multiple cortical areas. Their function is to keep a general arousal level and to maintain a state of alertness as a state of readiness to perceive and act. Those nuclei may reside in the reticular formation of the brain stem (serotonergic raphe nuclei, norandrenergic locus coeruleus), in the regions of thalamus called unspecific nuclei or matrix, or in cortical regions, like the cholinergic nucleus basalis Meynert from the basal forebrain [Kandel et al., 2000].

The tonic unspecific excitation is accompanied by a neuronal noise [Lücke, 2009], which is generated independently for each unit. It is modeled by the multiplicative gaussian white noise η_t, parameterized by a sufficiently small coupling σ. This noise term is applied to break the symmetry of the initial conditions at the very begin of learning, where the afferent synaptic weights are homogeneous and every external input leads to the equal activity response among the units. The symmetry breaking is necessary to enforce the WTA decisions under this condition and initiate the learning. Due to the bifurcation property of the dynamics, the module is extremely sensitive to even slightest differences among unit activities, so that already a small amount of noise is sufficient to cause the symmetry breaking effect. An alternative to a noise term in the dynamics would be a random initialization of the synaptic weights. Here I make the choice in favor of the noise term because it also provides a technical shortcut for testing the module's function under the different noisy conditions.

In addition to the local competitive mechanism mediated by the lateral inhibition within the module, we use a simple form of forward inhibition (FFI) acting on the incoming afferent inputs [Douglas and Martin, 1991, Swadlow, 2003, Burkhalter, 2008, Spruston, 2008, Pouille et al., 2009]. This kind of inhibitory mechanism was employed in a vast number of classical neuronal modeling studies [Grossberg, 1970, 1980, McClelland, 1981, Cohen et al., 1990] and also utilized in the previous approach [Lücke, 2005]. The afferent input is usually given by the scalar product of the synaptic weights vector and the vector of corresponding presynaptic activities. To model the disynaptic feed-forward inhibition, the incoming presynaptic activities are transformed as following before they make up the raw afferent input via the respective receptive field of a unit (for simplicity, I term here vector of synaptic weights also the receptive field of the unit):

$$\hat{p}_i^{Source} = p_i^{Source} - \frac{1}{K}\sum_k^K p_k^{Source}, \qquad Source \in \{BU, LAT, TD\}, \tag{2.1.16a}$$

$$\tilde{I}^{Source} = \sum_k^K w_k^{Source} \hat{p}_k^{Source}, \qquad Source \in \{BU, LAT, TD\}, \tag{2.1.16b}$$

where p^{Source} stands for raw presynaptic activity, $\hat{p}^{Source}$ is the presynaptic activity transformed by FFI, K is the total number of incoming synapses of a certain origin, w_i^{Source} are the synaptic weights and $\tilde{I}^{Source}$ designates the intermediate value of the afferent input from the respective origin. This

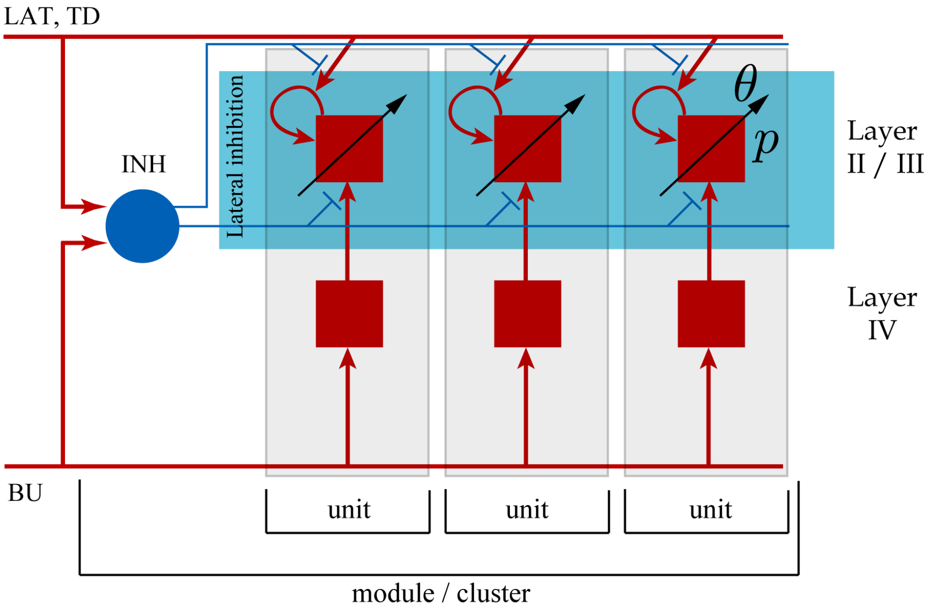


Figure 2.9: Module circuit diagram. As specified by the Eq. 2.1.13, the lateral (LAT) and top-down (TD) afferents modulate the strength of unit self-excitation, while the bottom-up (BU) afferents exert driving influence on unit's activity (see also Eq. 2.1.14). All afferent synapses are plastic and excitatory, exerting in addition feed-forward inhibition (INH) on their targets (Eq. 2.1.16, 2.1.17). The adaptive intrinsic excitability θ determines how easy a unit gets excited by afferent input.

intermediate value is further made mean-free, where the mean is computed across the afferent inputs to different units of the module:

$$I_j^{Source} = I_j^{Source} - \frac{1}{N}\sum_n^N \tilde{I}_n^{Source}, \qquad Source \in \{BU, LAT, TD\} \tag{2.1.17}$$

where N is the number of units in the module and I^{Source} the actual input entering the dynamic equation Eq. 2.1.13.

Although all plastic synaptic connections in the network are taken to be of excitatory nature, FFI allows units to exert inhibitory action across the modules in a network. An important effect of this processing is the selection and amplification of strong incoming signals at the cost of weaker ones, which can be interpreted as presynaptic competition among the local afferent signals [Yuille and Grzywacz, 1989, Spratling and Johnson, 2002, Lücke, 2005]. Here, this mechanism should enhance the effect of competition between assemblies representing different memories, as strong assemblies become able to disrupt cooperative signal exchange between units within the weaker ones. This communication disruption within weak, less coherent unit assemblies makes them become even more labile and susceptible for suppression via the stronger ones. Another advantage of FFI is that it helps to avoid useless computation on the postsynaptic side by canceling the incoming excitation if the activity differences within the transmitting module are too small, indicating only little progress in the decision process [Lücke, 2005]. This functionality has roughly the meaning of "no decision – nothing to react to".

2.2 Homeostatic activity regulation

The activity dynamics equation (Eq. 2.1.13) contains the variable θ, which stands for the intrinsic excitability of the unit. Higher values of θ correspond to higher unit excitability, implying a greater potential to become active given a certain amount of input and vice versa. Now, unit activity homeostasis can be achieved by adapting the intrinsic excitability θ if the average unit activity level $\langle p \rangle$ deviates from a defined target activity level p_{aim}:

$$\frac{d\theta}{dt} = \tau_{\theta}^{-1}(p_{aim} - \langle p \rangle), \tag{2.2.1}$$

where $\langle p \rangle = \frac{1}{T}\int_{t}^{t+T} p(t)dt$ is the average activity of the core unit measured over the period T of a decision cycle, p_{aim} specifies the target activity level and τ_{θ}^{-1} is the inverse time constant. The target activity level p_{aim} is a simple function of the number of core units n in a module, $p_{aim} = \frac{1}{n}$. The initial value of the intrinsic excitability is $\theta(0) = 0$.

The mechanism of intrinsic excitability regulation has a sound neurophysiological basis described in numerous experimental works [Desai et al., 1999, Marder and Prinz, 2002, Debanne et al., 2003, Zhang and Linden, 2003, Davis, 2006, Nelson and Turrigiano, 2008, Maffei and Fontanini, 2009]. It has been widely used in neural modeling to stabilize the neural circuits and to optimize their function [Buhmann et al., 1989, Földiák, 1990, Liu et al., 1998, Gorchetchnikov, 1999, Triesch, 2007]. Here, the homeostatic regulation of unit activity encourages a uniform duty cycle across units in the network to assure their equal participation in memory trace formation during the learning phase. Bearing in mind the strong competitive character of the neuronal dynamics, the regulation of the intrinsic excitability θ changes the *a priori* probability of a core unit to be the winner in a decision cycle. So, if a certain unit happens to take part too frequently in encoding of the memory content, violating the requirement of uniform win probability across the units, its excitability will be downregulated so that the core unit becomes more difficult to activate, giving an opportunity for other core units to participate in the representation. Reversely, an unit being silent for too long is upregulated, so that it can get excited more easily and contribute to memory formation. This functional consideration justifies the choice for the target activity level, $p_{aim} = \frac{1}{n}$. Depending on the number of the units in the module, the target activity level defines the target uniform probability for a unit to win, which decreases with the increasing module size.

2.3 Activity-dependent bidirectional plasticity

Intrinsic plasticity introduced in the previous section is a synapse-unspecific adaptive mechanism, which regulates the excitability of the unit. In order to become selective for a certain specific input pattern, a unit has to be able to adjust its synaptic weights in activity-dependent fashion. Such adaptive mechanism is classically termed synaptic plasticity [Martin et al., 2000, Feldman, 2009]. I choose here a bidirectional form of synaptic plasticity to specify how a synapse connecting one unit to another may undergo a change in its strength w:

$$\frac{dw}{dt} = \underbrace{\varepsilon p^{pre} p^{post}}_{\textit{Hebbian term}} \underbrace{\mathcal{H}(p^{post} - \theta_0^{-})}_{\textit{neutral / LTD gate}} \underbrace{\mathcal{H}_{-}^{+}(p^{post} - \theta_{-}^{+})}_{\textit{LTD / LTP gate}} \underbrace{\mathcal{H}(\chi - A(t))}_{\textit{Total activity gate}} \tag{2.3.1}$$

with the sign switch functions $\mathcal{H}(x)$ and $\mathcal{H}^{+}_{-}(x)$

$$\mathcal{H}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}, \qquad \mathcal{H}^{+}_{-}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \tag{2.3.2}$$

providing the bidirectional form of the synaptic modification. The amplitude of the change is determined by the correlation between the presynaptic activity p^{pre} and the postsynaptic activity p^{post}, both variables being non-negative due to the properties of the unit activity dynamics. The learning rate ε specifies the speed of modification being the inverse time constant. Other variables determine the sign of the modification. The threshold $\theta^{+}_{-} = \max(\vec{\mathbf{p}}^{post}_{t})$ is used to compare the postsynaptic unit activity against current maximum activity in the module. $A(t)$ is the the total activity level in the module at time point t, $A(t) = \sum_{i=1}^{n} p^{post}_{i}(t)$, where n is the number of units in the module and $p^{post}_{i}(t)$ their postsynaptic activities at time point t. $A(t)$ is compared to a variable gating threshold χ [Lücke, 2005], which pursues the average total activity level $\langle A(t) \rangle$ computed over the period T of a decision cycle:

$$\frac{d\chi}{dt} = \tau_{\chi}^{-1}(\langle A(t) \rangle - \chi), \quad \langle A(t) \rangle = \frac{1}{T}\int_{t}^{t+T} A(t)dt \tag{2.3.3}$$

with τ_{χ}^{-1} as inverse time constant. The threshold initial value can be set to an arbitrary value between zero and one, here I set it to $\chi(0) = 0.5$. Furthermore, the postsynaptic activity p^{post} is compared to the sliding threshold θ_{0}^{-} that follows the average postsynaptic activity $\langle p^{post}(t) \rangle$ computed over the period T of a decision cycle:

$$\frac{d\theta_{0}^{-}}{dt} = \tau_{\theta_{0}^{-}}^{-1}(\langle p^{post}(t) \rangle - \theta_{0}^{-}), \quad \langle p^{post}(t) \rangle = \frac{1}{T}\int_{t}^{t+T} p^{post}(t)dt \tag{2.3.4}$$

with the inverse time constant $\tau_{\theta_{0}^{-}}^{-1}$. The initial value of the threshold can be any positive value close to zero. Here I choose $\theta_{0}^{-}(0) = p_{aim}$, being equal to the target postsynaptic activity level (see Eq. 2.2.1).

The rule employed here is a simplified version of bidirectional synaptic plasticity that assumes existence of two sliding thresholds θ_{0}^{-} and θ^{+}_{-} (Fig. 2.10). There are some classical bidirectional plasticity rules employing sliding thresholds, like BCM [Bienenstock et al., 1982] or ABS [Artola and Singer, 1993]. Being confirmed by neurophysiological studies [Ngezahayo et al., 2000, Cho et al., 2001, Abraham et al., 2001, Bear, 2003, Abraham, 2008], the sliding thresholds subdivide the range of postsynaptic activity into zones where no modification, depression or potentiation may occur. The sliding thresholds are adaptive as their name suggests, changing their value according to the previous activation history on the postsynaptic side. There is strong evidence that the sliding nature of the thresholds has its physiological substrate in a Ca^{2+}-dependent intracellular mechanism. This mechanism is thought to be able to measure and integrate activity-dependent entry of Ca^{2+} into the neuron, establishing a protocol of previous average unit activation by low-pass filtering its activity. This measurement establishes the value of the sliding thresholds, being probably stored in form of various bounded Ca^{2+}-dependent protein kinases, like Ca^{2+}/calmodulin-dependent kinase II (CaMKII). The current entry level of Ca^{2+} can be then compared to the average entry level as reflected by the sliding threshold, thus implementing comparison between the current and the average neuron activity [Lisman, 1989, Artola et al., 1990, Lisman, 1994, Abraham and Tate, 1997, Castellani et al., 2001, Cavazzini et al., 2005, Sanhueza et al., 2007]. Such dependence of synaptic modification on the previous activation protocol is also termed *metaplasticity* [Abraham and Bear, 1996, Abraham, 2008]. In detail, if the postsynaptic activity level is too low ($p^{post} < \theta_{0}^{-}$), no synaptic modification can be triggered. A mediocre level of activation

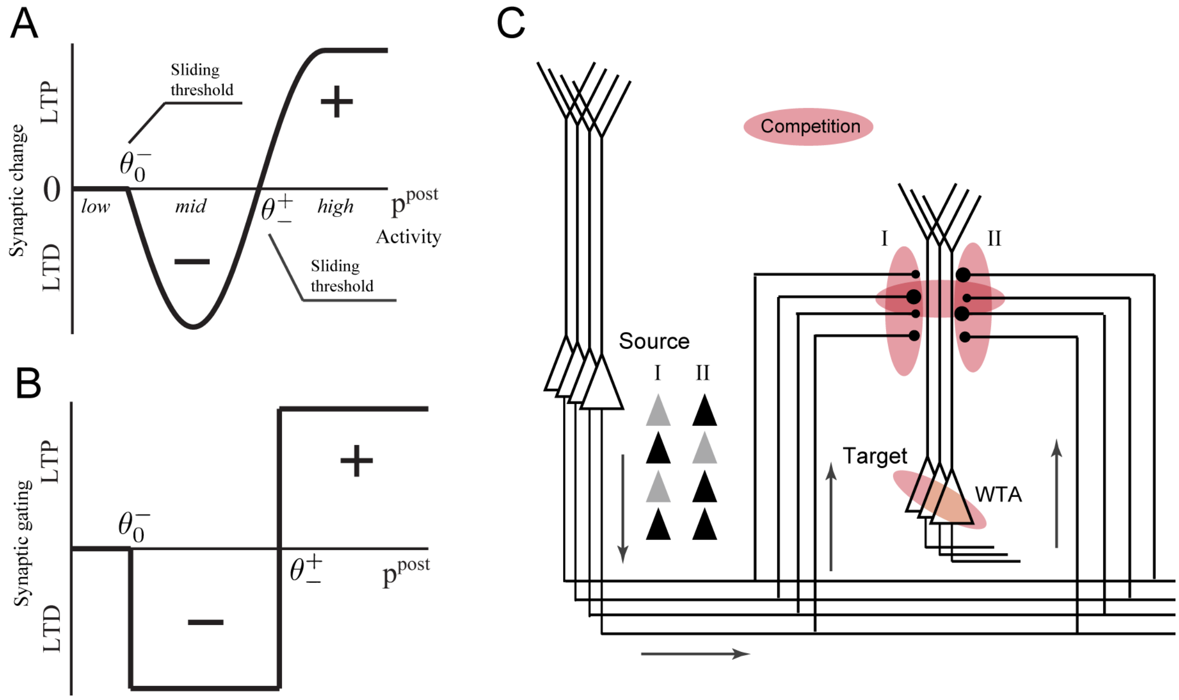


Figure 2.10: Bidirectional plasticity and competitive learning. **(A)** Bidirectional form of synaptic plasticity equipped with two sliding thresholds as suggested by experimental studies. **(B)** Approximating the experimental rule, the gating rule modulates the classical Hebbian plasticity by determining the sign of the synaptic modification. **(C)** Combined with the WTA operation, the bidirectional plasticity introduces competition in synapse formation across the receptive fields of the units. The synaptic competition should encourage strengthening of the synapses storing the discriminative features of the input patterns, while attenuating the synapses conveying features common to the input patterns. Within the receptive fields, the multiplicative synaptic scaling promotes further competition between the synapses converging on the same unit.

($\theta_0^- < p^{post} < \theta_-^+$) promotes long-term depression (LTD, negative sign), and only a high level of activity ($p^{post} > \theta_-^+$) makes long-term potentiation (LTP, positive sign) possible.

Combined with the winner-take-all-like behavior of the module units, the intended effect of the rule is to introduce competition in synaptic formation *across* the receptive fields of the units (Fig. 2.10 **(C)**). This competition should assist separation of input patterns that have to be considered as distinct, despite the eventually high similarity and strong overlap. Imagine two input patterns of that kind that are initially not separated in terms of unit selectivity, so that there are at least two units becoming active if one or another pattern is presented. If the input patterns are presented over and over, this situation will also reoccur. Now, the probability for potentiation which happens on the active synapses will be different depending on the kind of pattern feature conveyed by an active synapse. If the synapse conveys a feature which is common to both patterns, then the probability to win for the unit that is contacted by such a synapse will be rather low, as there may be plenty of other units responding to the same feature and entering the competition. Consequently, the probability of potentiation for such a synapse will be low, too, and the probability for depression increases, as the units that become active and then loose the competition will tend to get depression on their active synapses.

On the contrary, if the synapse conveys a discriminative feature which is unique for one of the patterns, then it will have higher probability to get potentiated on the unit that becomes selective for the particular pattern. This is because once such a synapse supports the unit to become a winner given the input pattern containing the discriminative feature, the amplification effect sets in, potentiating the

Parameter	Number of units per module			
	8	20	40	120
τ	$0.02\,ms$	$0.02\,ms$	$0.02\,ms$	$0.02\,ms$
τ_{θ}^{-1}	$10^{-4}ms^{-1}$	$10^{-4}ms^{-1}$	$5\cdot10^{-5}ms^{-1}$	$1.2\cdot10^{-5}ms^{-1}$
ε	$5\cdot10^{-4}\,ms^{-1}$	$5\cdot10^{-4}\,ms^{-1}$	$5\cdot10^{-4}ms^{-1}$	$5\cdot10^{-4}ms^{-1}$
$\tau_{\theta_0^-}^{-1}$	$2\cdot10^{-3}ms^{-1}$	$2\cdot10^{-3}ms^{-1}$	$2\cdot10^{-3}ms^{-1}$	$2\cdot10^{-3}ms^{-1}$
τ_{χ}^{-1}	$10^{-3}ms^{-1}$	$10^{-3}ms^{-1}$	$10^{-3}ms^{-1}$	$10^{-3}ms^{-1}$
ϵ	0.02	0.01	$3\cdot10^{-3}$	$3\cdot10^{-4}$

Table 2.1: Model parameters overview for modules of different size.

synapse and increasing further the win probability of the unit with respect to the given input pattern, which will then again lead to potentiation of the synapse and so forth. The same kind of synapse contacting other units that are less selective for the particular pattern feature will face strong competition from the more successful synapse of the selective unit. The successful synapse will consequently prevent from potentiating the less successful synapses that converge on other units, becoming the only one sensitive to the discriminative feature that is able to grow strong. So, the synapses signaling for the common, non-distinctive features will get attenuated, but not vanish, whereas the synapses carrying the discriminative features will be able to grow strong on the units selective for the respective input patterns.

In addition, I here use a form of multiplicative synaptic scaling, which is thought to play a role in homeostatic regulation of total synaptic strength in the cortex [Abbott and Nelson, 2000, Turrigiano, 2008]. The scaling is applied to synapses grouped according to their origin (bottom-up, lateral and top-down). This is modeled by L^2-normalization of the receptive field vector, $\tilde{w}_i^{Source} = w_i^{Source} \big/ \|\mathbf{w}^{Source}\|_2$, with w_i^{Source} as a weight of the receptive field comprising the synapses of the respective origin $Source \in \{BU, LAT, TD\}$, and $\tilde{w}_i^{Source}$ its normalized version. The normalizing procedure can be applied after any sufficiently small number of decision cycles, here I choose this number to be 10 cycles. The mechanism of multiplicative scaling also promotes competition between synapses within the receptive field, as the growth of one synapse happens at the cost of the weakening the others [von der Malsburg, 1973, Miller and MacKay, 1994].

2.4 Model parameters and simulation

To instantiate both competitive processing and competitive learning in the module, a number of parameters in the model differential equations has to be set up accordingly to guarantee the intended function. Some of those parameters, like the unspecific excitatory drive ϵ , turn out to be dependent on the number of the units in the module, while others, like the time constant of the fast neural dynamics of the unit or the synaptic learning rate, do not. Dependent means here that adapting such parameters to a given number of units is necessary to achieve a substantial performance in unsupervised learning tasks. The parameters are summarized in Tab. 2.1.

Parameter choice. The choice of the parameters made here follows the qualitative picture of the physiological properties of neuronal dynamics, of the intrinsic and synaptic plasticity. As neuronal dynamics reflects here the activity of a whole population of tightly coupled pyramid neurons receiving common afferents (activity variable p), we may assume a small time constant value, referring to an almost instantaneous response behavior of a sufficiently large (in order of hundred) population of neurons [Gerstner, 2000]. The plasticity mechanisms are known to operate on diversity of time scales that are substantially slower than the time scale of neuronal activity [Nelson and Turrigiano, 2008, Feldman, 2009]. Here, the time scales of synaptic and intrinsic plasticity are of the same order. It turns out to be critical to set up the time constant of the sliding threshold dynamics used in the plasticity rule (see Eq. 2.3.1 and 2.3.4) at least two orders of magnitude faster than the time constant of the intrinsic plasticity. The setup of the time scales corresponds to the computational requirements put on the module. On the one hand, the module has to react fast on the incoming input and generate a rapid response signaling a decision for a candidate unit that represents the current input at best. On the other hand, the units within the module have to learn their synaptic weights on a much longer term from the variety of incoming inputs and slowly build up a basis to represent appropriately the space the inputs originate from.

Simulation. To simulate the module and networks composed of those, I developed a C++ library that supports the described dynamics and plasticity mechanisms, modifying and extending the existing platform CARL, which was used previously for simulation of hard-wired networks [Wolfrum et al., 2008]. All the stochastic differential equations governing the dynamical behavior are solved numerically using the simple Euler method [Kloeden and Platen, 1992], with a sufficiently small fixed time step $\Delta t = 0.02ms$. To save computational time, the slow adaptive variables θ, θ_{-}^{0} and χ are updated once in a decision cycle, adjusting the time constants accordingly within the simulation. For the larger modules comprising more than $N = 100$ units, a reset signal has to be applied at the begin of each cycle, re-setting the activity levels of the units to an equal low level (the level can be an arbitrary value close to zero, here I choose $p_{init} = 0.02$). This signal is only necessary in the early learning phase, it becomes redundant in the later phase and can be dropped there without any functional loss.

2.5 Unsupervised learning with single and distributed modules

Designed for the processing and unsupervised learning in a layered memory network, the module as a single, isolated device can already solve non-trivial tasks performing unsupervised learning on the incoming bottom-up sensory inputs. In this limited setting, two task types involving unsupervised learning can be posed. The first is a form of unsupervised clustering. In this task type it is required to partition the input space such that each partition, or cluster, contains a number of input data points related according to some distance or similarity measure. For each cluster, a codebook vector has to be formed, so that the input samples from the cluster can be projected onto a low-dimensional representative point, or label. In this way, for each input sample there is a corresponding cluster label, and the whole space is quantized by a finite set of discrete codebook vectors. The second task type is the non-linear component extraction. There, the input samples are mixtures of some hidden underlying causes, or components. The task is to deduce these components by observing the incoming inputs, forming a representation of the input space that allows each input sample to be decomposed as combination of the underlying components that generated the input.

These tasks are related in the sense that in both cases the unsupervised learning procedure has to find a proper basis to represent the accessible input space in a compact and faithful way. This ability is certainly also of crucial importance for a system that has to build up a memory domain in an unsupervised fashion. Proving the ability of a single module to be capable of finding a "good" basis, or a vocabulary, for the representation of the locally available input space can be thus considered as preliminary test for the functional role of the module as a node in the full network architecture.

To provide the module with input of natural complexity, I will take use of standard databases containing natural face images. I will show that distributed isolated modules attached to specific landmarks on the face image are already capable of building a distributed representation for the identity and gender of the persons from the images presented during the learning in an unsupervised fashion. For the

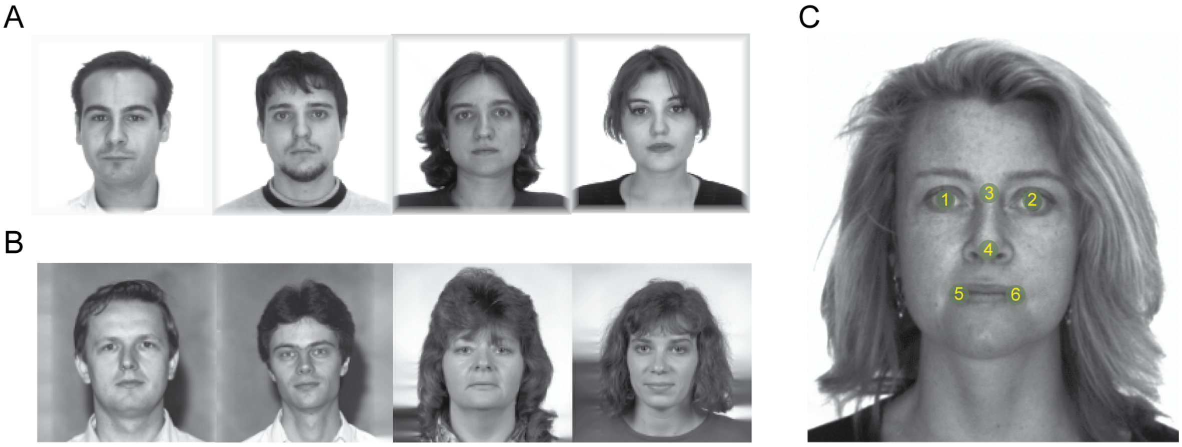


Figure 2.11: Face databases used for learning. **(A)** AR Database (120 persons used). **(B)** FERET Database (1000 persons used) **(C)** Facial landmarks selected for the extraction of local Gabor filter banks.

demonstration of component extraction ability, I will use two scenarios where the images of natural scenes and the artificial black-and-white bar images will be employed to create input data sets. After the learning the units obtain selectivity for certain causes, or components, underlying the input. The emerging coding scheme of the module can be interpreted in terms of probabilistic signaling for the components hidden in the current input. The competitive computation underlying the coding scheme is carried out within the frames of ongoing gamma rhythms. In such a coding scheme, each gamma cycle can be thus interpreted as an atomic fragment of competitive processing and learning.

2.5.1 Unsupervised clustering and learning of facial features

Data format. The AR and FERET databases containing gray-scale human face photographs will be used throughout this work (Fig. 2.11). AR database contains 126 different persons in total, of which at most 120 are used here [Martinez and Benavente, 1998], picked for each task randomly to create an input data set. For each person, there is a number of views taken under different conditions. The original view with neutral facial expression is accompanied by a duplicate view depicting the same person at a later time point (two weeks after the original shot). Furthermore, there are variations in emotional expression such as smiling or sad for both original and duplicate views. The FERET database provides 1000 persons [Phillips et al., 2000]. Here I make use of two different view variations, one containing faces with neutral expression and another containing faces with a smile. For the limited task setting with single isolated modules, only neutral face views are taken (As benchmarking of isolated modules is not in focus of this work, I do not perform any tests on alternative face views in this section, this will be done for the full network architecture presented in the next chapter). Further, each data set contains roughly the same number of males and females.

The images are automatically pre-labeled with a graph structure put upon the face, positioning nodes on consistent landmarks across different individuals with a software (EAGLE) based on the algorithm described in [Wiskott et al., 1997]. After labeling, a subset of $L = 6$ facial landmarks is selected around the eyes, nose and mouth regions (Fig. 2.11 **(C)**). In the task involving only one single module, one of these landmarks located at the tip of the nose is chosen to attach the module to the image. In the extended version of the task, each landmark is subserved by a single module, resulting in $M = 6$ distributed isolated modules being attached to respective landmarks on the image. The single module is provided with a sensory image signal represented by a Gabor filter bank extracted locally (Fig.

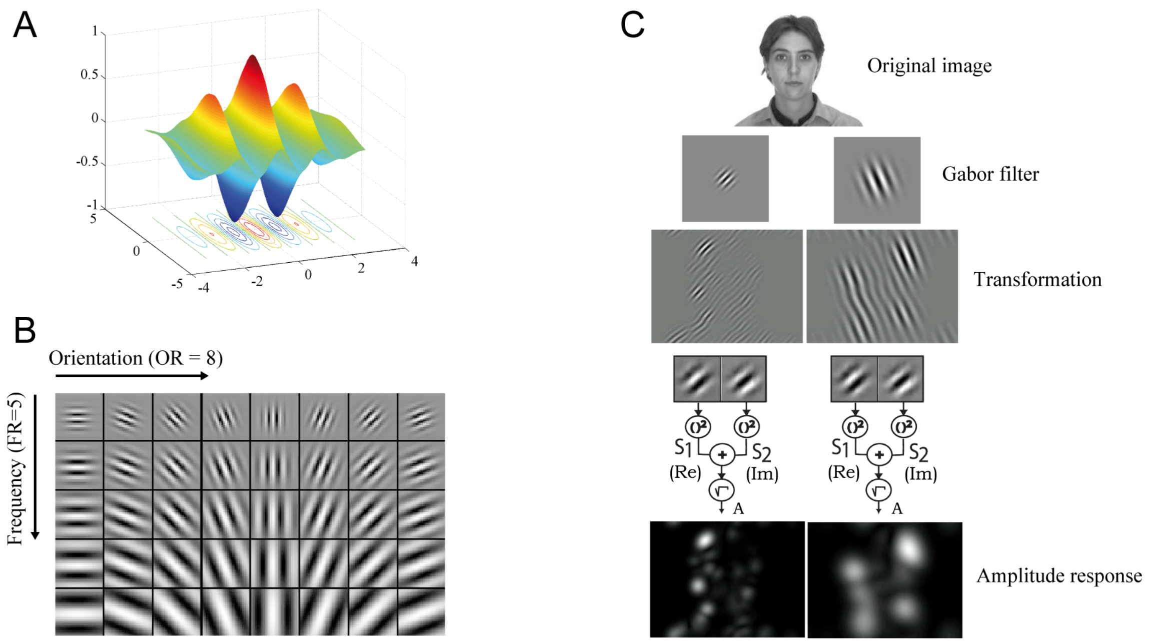


Figure 2.12: Gabor filter bank. **(A)** A Gabor wavelet (real part). **(B)** Gabor filter bank constructed from wavelets of 5 different frequencies and 8 different orientations. **(C)** By combining two phase-shifted wavelets, each filter produces an amplitude response that can be extracted from any point on the original image. The full filter bank extracted locally is thus a vector of local amplitude responses.

2.12). The Gabor wavelet family used for the filter operation is parameterized by the frequency k and orientation φ of the sinusoidal wave and the width of the gaussian envelope σ [Daugman, 1985]. Here, $s = 5$ different frequencies and $r = 8$ different orientations are sampled uniformly to construct the full filter bank according to a standard setup (for more details refer to [Wiskott et al., 1997]). The local filtering of the image produces a complex vector of responses, containing both amplitude and phase information. Only the amplitude part consisting of $s \cdot r = 40$ real coefficients is used to model the responses of complex cells that are tolerant against small local translations. This amplitude vector is further normalized by L^2-Norm to serve as bottom-up input for the single module (Fig. 2.13).

Module configuration. For simple test procedure with only one module, a number of persons from AR database $(P = \{20, 40, 120\})$ is randomly selected, allocating a corresponding number $(N = \{20, 40, 120\})$ of units within the module (Fig. 2.13). For the extended test procedure with M distributed, but isolated modules (6 in total), each module contains only $N = 20$ units (Fig. 2.14). The training data set is then either 120 or 1000 persons from the respective database (AR or FERET). Because in this scenario multiple persons will have to share same units that learn a particular local appearance, the setup is well-suited to test for the clustering capability of the single module as well as for the combinatorial code generated by distributed modules to represent the compositional face identity .

Open-ended unsupervised learning. Before the learning procedure starts, the structure parameters are initialized homogeneously, all intrinsic excitability values and all synaptic weights being undifferentiated (as each receptive field vector is L^2-normed, the initially equal weight values depend on the number of synapses converging on a unit). Before the start, the unit activities can be set to an arbitrary low initial level, here I choose $p(0) = 0.02$. During the iterative learning procedure, for each decision cycle a face image is selected from a training data set randomly and presented to the system, evoking a pattern of activity in the single modules and triggering synaptic and threshold modification mechanisms. The learning procedure is *open-ended* as there is *neither* a stop condition *nor* an explicitly

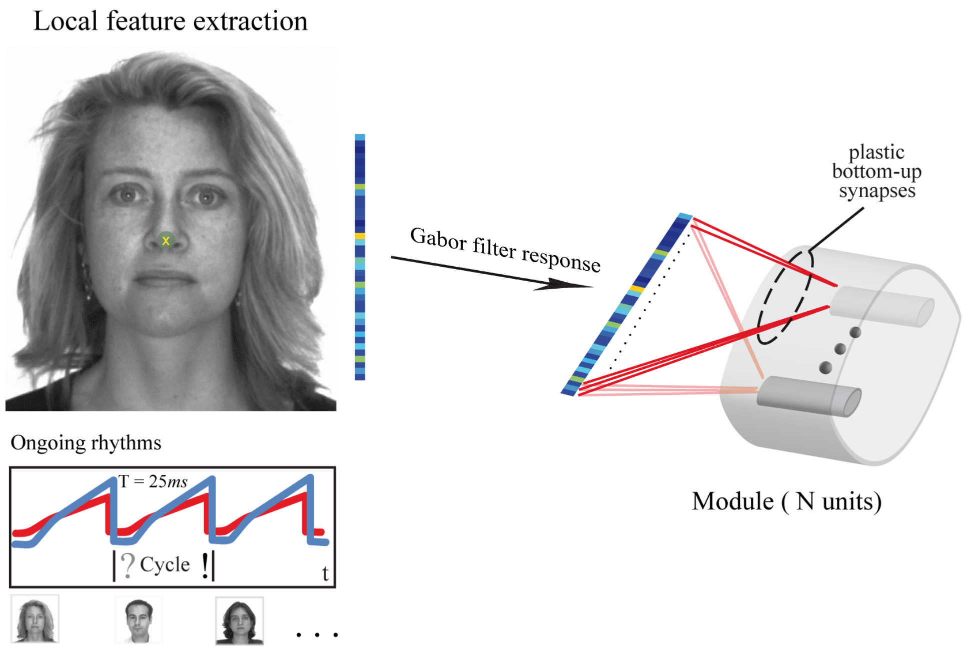


Figure 2.13: Learning procedure for single module. At the begin of each decision cycle, a Gabor filter bank (a vector of 40 components, 5 frequencies × 8 orientations) is extracted from a landmark of a randomly selected image and presented as bottom-up input to the module attached to the respective landmark. The learning is open-ended and self-stabilizing, an explicit stop condition is thus not necessary.

defined time-dependent learning rate variables which would decrease with time progress and freeze modifications at some point.

Evaluation of learning error. The learning error is defined as a rate of wrong responses to person identity from the training data set containing original face views with neutral expression. As the on-line learning is done in a completely unsupervised fashion without providing any data labels, it has to be clarified what is a wrong response given a certain input. Therefore I define here a form of prediction error using the previous history of units responses observed during a period of the learning phase. From these observations, the conditional probabilities for a certain face identity on the input given a particular winner unit $P(person|winner)$ can be computed (as well as the conditional probabilities for a unit to win given a certain person on the input $P(winner|person)$). The learning error can be then evaluated for the short interval following the preceding period used for the computation of the conditional probabilities. Presenting a face image on the input elicits response in form of distributed activity pattern across the modules (see also Fig. **(B)**). In case of only a single module, one unit becomes the winner in course of the decision cycle. This decision is taken as prediction for the face identity on the input by simply determining the most likely person given the particular winner unit from the previously computed probability table : $\tilde{person} = \underset{person}{\operatorname{argmax}} P(person|winner)$. Now the identity error rate can be easily computed as fraction of the number of wrong guesses ($\tilde{person} \neq person_{true}$) over the number of total cycles during the short interval.

In case of M distributed isolated modules, a simple form of voting can be used to make the same kind of prediction about the person identity. Each module signals its decision about the given local input by picking a winner unit in course of decision cycle. The prediction for person identity is then

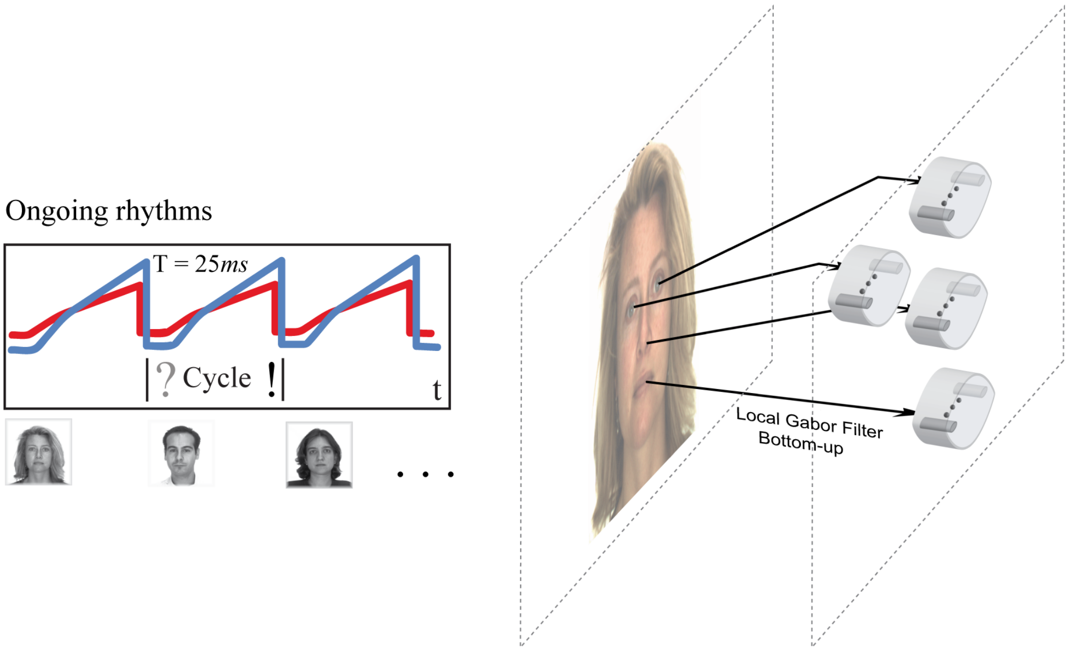


Figure 2.14: Learning procedure for distributed isolated modules. Six modules are attached to pre-specified landmarks on the image ($N = 20$ units per module). The procedure follows the scheme described for a single module, extracting six different Gabor filter banks as bottom-up input for six distributed modules from a randomly selected face image at the begin of each decision cycle.

computed via simple voting as $\tilde{person} = \operatorname*{argmax}_{person} \sum_{winner=1}^{M} P(person|winner)$. The same can be done for determining the gender error rate by taking the conditional probabilities $P(gender|winner)$, with $gender = \{male, female\}$, instead. The gender prediction reads then in case of one single module $\tilde{gender} = \operatorname*{argmax}_{gender} P(gender|winner)$, and $\tilde{gender} = \operatorname*{argmax}_{gender} \sum_{winner=1}^{M} P(gender|winner)$ in case of distributed modules. Note that the probabilities $P(winner)$ do not need to be taken into account here, as the mechanism of homeostatic activity regulation (see Sec. 2.2) renders those to be virtually the same if observed over a sufficiently long time interval.

Results. First, let us take a look on a single module that gets its input from a single landmark in the face image (in this case, the landmark is located on the tip of the nose, see Fig. 2.13). Presenting the persons images ($P = \{20, 40, 120\}$) from the database results in the formation of the bottom-up synaptic structure that captures the local appearance of each presented face in the synaptic weight vectors of units within the module. This bottom-up connectivity matrix can be visualized, and the course of the learning error on the data set can be be plotted by computing the prediction error as defined previously (Fig. 2.15).

The vocabulary the module units form to represent the local input space is in this case simply the set of codebook vectors, each of those representing the local appearance of one particular person shown during the learning (Fig. 2.15). This is the most simplest kind of unsupervised clustering, where each element from the input space is projected in bijective fashion to a point in the representational space. Still, the task is non-trivial, as it is not pre-specified, which unit has to acquire selectivity for which person, so it may happen that multiple persons get clustered on a single unit, while consequently leaving

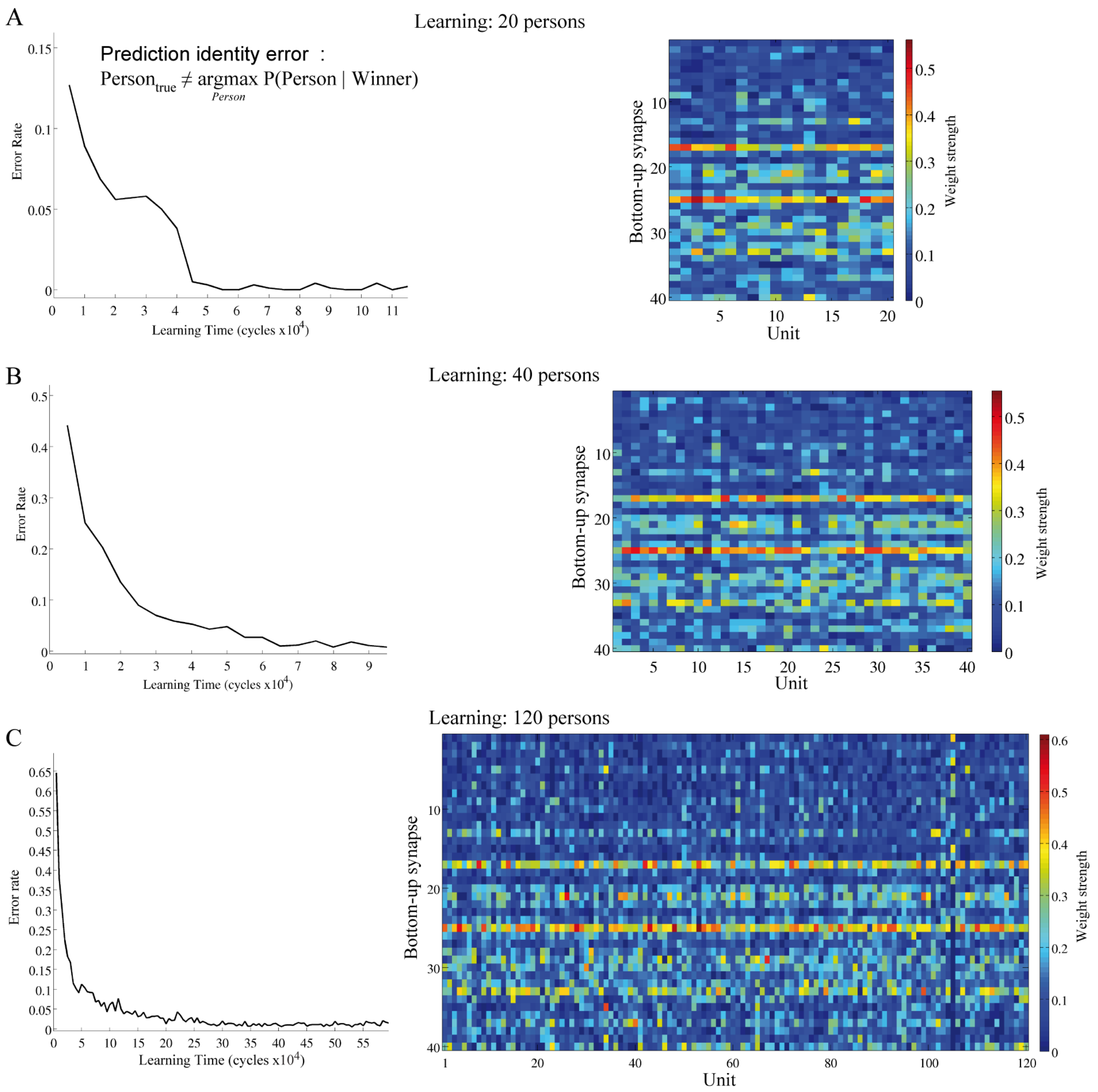


Figure 2.15: Learning error course and synaptic structure formation for single modules of different size attached to a facial landmark on the nose tip. The task was to develop unit selectivity for **(A)** $P = 20$, **(B)** $P = 40$ and **(C)** $P = 120$ different persons shown during the learning. Learning error time course on the left, synaptic weights formed by each unit on the right (one column corresponds to the bottom-up receptive field of one unit). After a certain number of decision cycles, the error rate on the data set goes close to zero, the successful unit differentiation is evident.

some of them unused, or dead (which is known as dead unit problem in the classical competitive learning). Reversely, multiple units may happen to acquire selectivity for the same person, while leaving a number of persons unattended. The single module manages to successfully avoid these difficulties. Crucial for this success is the combination of the proper unit usage load balancing done by the homeostatic activity regulation on the one hand and, on the other hand, of the unit differentiation in terms of developing selectivity for different patterns performed by competitive synaptic modification. The balancing of the unit usage load renders the winner-take-all operation performed by the module "fair" on a long term, which means that each unit gets on average equally often an opportunity to become a winner of a decision cycle and adapt its synaptic connectivity to the incoming inputs. The competitive learning reassures at the same time that the synaptic weight vectors of the units are driven apart, making the units selective for different local appearance for the data set. As shown in the time course of the learning error, the error rate on the data set goes close to zero after a certain number of decision cycles, underpinning the success of unit differentiation (Fig. 2.15).

Now let us consider a configuration containing distributed isolated modules ($M = 6$), each attached to its dedicated facial landmark. Again, a number of persons from the face image data set is presented to the system during the learning phase. However, this time the number of persons $P = \{120, 1000\}$ exceeds the number of units available locally in the module, $N = 20$. To develop a distributive representation for individual faces, the system has to perform unsupervised clustering with each single module, projecting multiple local appearance features on the same unit. The result of the clustering can be visualized by plotting the unit selectivities in respect to person and to gender. The degree of unit selectivity for person identity and gender is taken from the conditional probabilities $P(person|winner)$ and $P(gender|winner)$, that are estimated from the history of unit responses as described earlier in the section. A selectivity plot can then be constructed by stacking for each unit conditional probability values one on another in form of bars corresponding to each value. The person selectivity plot shows that each unit has a number of persons that preferably activate the unit and make it win (Fig. 2.16, 2.18 **(A)**). The preferred persons share units in a balanced way as expressed by roughly equal length of bars for the corresponding conditional probabilities.

The value of conditional person probability corresponds to a degree to which a unit becomes selective for a particular person. This values can be further used to compute how many persons cluster on, or share, a particular unit. To do this, I define a threshold for cumulative probability, $P_\theta = 0.98$ and add up the values until the threshold is reached. The threshold is there to discard the persons for which selectivity degree is very low (here < 0.02). Counting the number of times needed to reach the threshold, the number of persons is obtained for which the unit developed a significant degree of selectivity, or, in other words, we get the number of persons that share the unit. As it is evident from the plot, the unit load is balanced, with about $P/N = \{6, 50\}$ persons sharing a unit on average (Fig. 2.16, 2.18 **(B)**). The load scatters for individual units towards slightly higher and lower usage. The balanced unit usage load is reflected once more in the roughly uniform win probability of the units (Fig. 2.16, 2.18 **(C)**). The individual deviations from the average win probability $P(winner) = 0.05$ correspond then to more or less utilized units.

The same kind of unit selectivity computation can be performed not only for the person identity, but also for the gender, male of female. The unit gender selectivity plot is made by using the conditional probabilities $P(gender|winner)$. For each unit, the gender selectivity score can be obtained. This score delivers a value between 0 and 1, obtained by calculating the difference of the two conditional gender probabilities and taking its the absolute value. The zero value means no selectivity for a particular gender, whereas 1 denotes the maximal selectivity for either male or female. Through the unsupervised learning, some of the units are able to acquire high selectivity for male of female faces, while other units do not differentiate between the genders (Fig. 2.16, 2.18 **(C)**, **(D)**). It is remark-

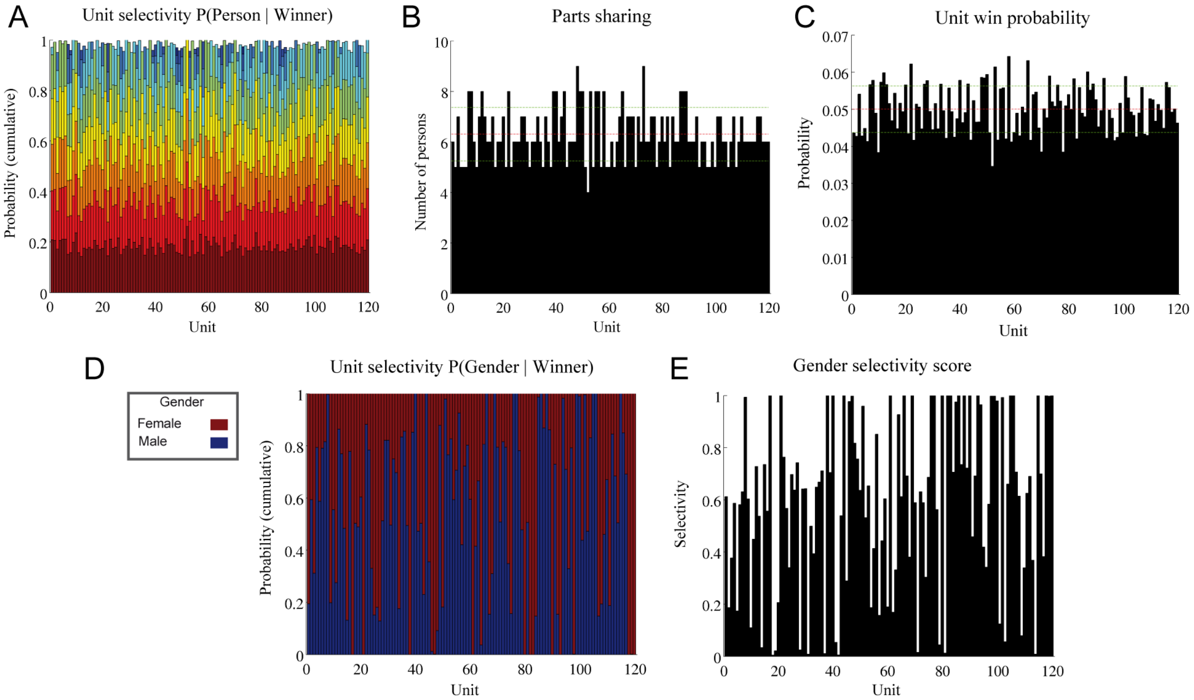


Figure 2.16: Unit selectivity formed by learning with $M = 6$ distributed isolated modules ($N = 20$ units each), $P = 120$ persons in the training set. The plots show the state after $5 \cdot 10^5$ decision cycles.**(A)** Selectivity for different persons. Each colored bar corresponds to a conditional probability $P(person|winner)$ for a respective unit. **(B)** Number of persons sharing a unit, computed from unit selectivity. **(C)** Unit win probability computed over time interval of $2{\cdot}10^4$ cycles. The balanced unit usage load is reflected in roughly uniform win probability, with deviations corresponding to more or less utilized units. **(D)** Gender selectivity. Each colored bar corresponds to a conditional probability $P(person|gender)$ for a respective unit. **(E)** Gender selectivity score. Some units develop very high gender selectivity (score 1), while others only poorly differentiate between male and female faces (score close to 0).

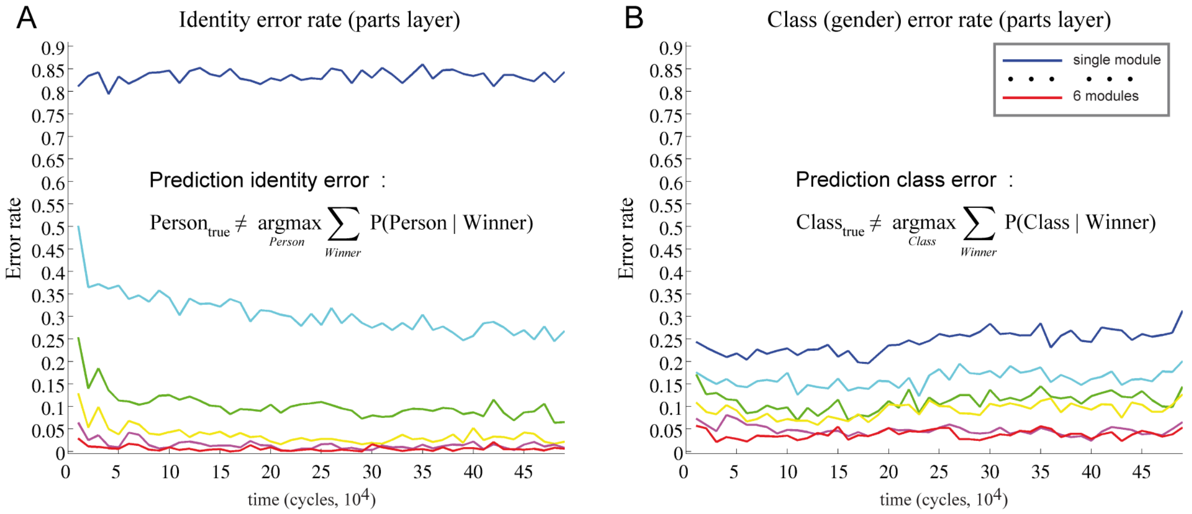


Figure 2.17: Learning error for $M = 6$ distributed isolated modules ($N = 20$ units each), $P = 120$ persons in the training set **(A)** Person identity error. **(B)** Gender error. The more modules are involved in identity or gender estimation, the less is the error rate. For 6 modules, the identity error is almost zero, while the gender error is well below 5%.

able, that this selectivity can be established in unsupervised fashion alone on the basis of local face appearance.

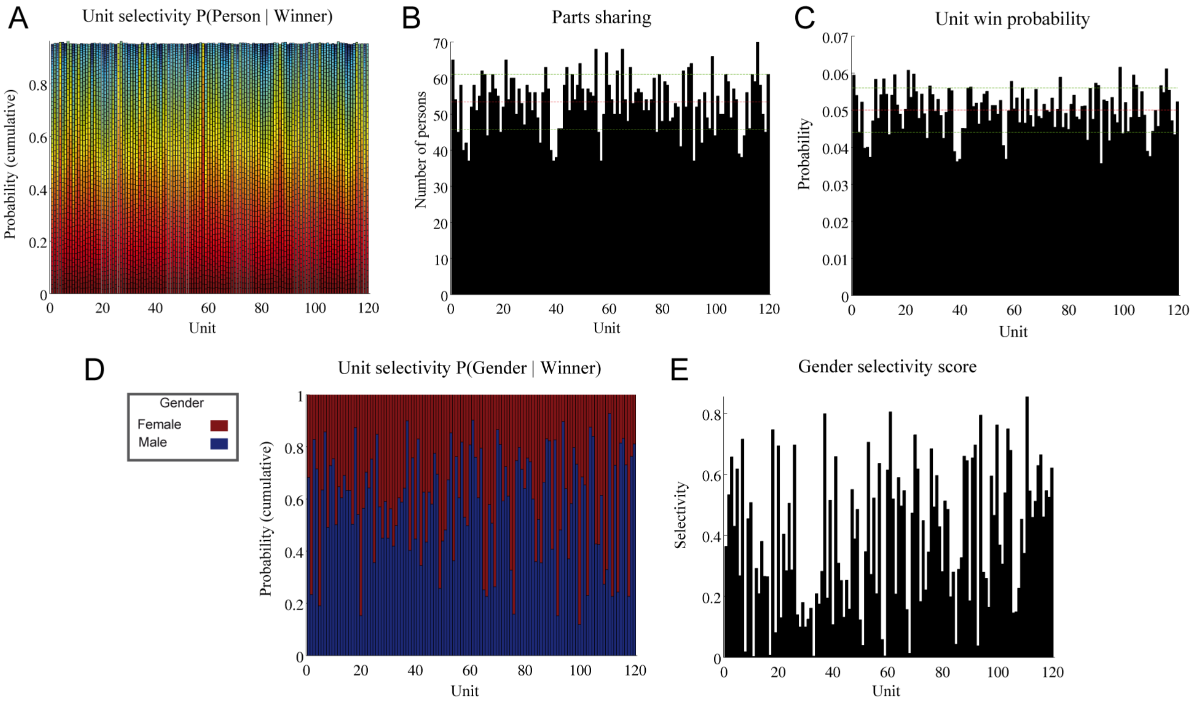


Figure 2.18: Unit selectivity formed by learning with $M = 6$ distributed isolated modules ($N = 20$ units each), $P = 1000$ persons in the training set. The plots show the state after $6 \cdot 10^5$ decision cycles. The qualitative picture of unit selectivity and unit usage load is the same as for learning 120 persons. About $50 - 60$ persons on average get clustered here on one unit and have to share it. The gender selectivity score is still high for many units, although no unit is able to reach the maximum score of one.

As each unit represents an element of local vocabulary that is formed in the course of learning, the individual faces shown during the learning phase are represented in distributive, combinatorial fashion as composition of their local elements, or parts. Combinatorial means here, that for each face a subset of preferred units is picked from the distributed modules by choosing one best fit candidate from multiple alternatives offered by each vocabulary. A unit is shared by multiple persons, so the question arises whether the unit subsets that correspond to memory traces representing individual faces can indeed discriminate good enough between different persons, so that the prediction error rate on the learning data set gets close to zero.

Indeed, looking on the time course of the learning error for both cases of $P = \{120, 1000\}$ persons reveals the drop of the learning error rate, which goes close to zero for $P = 120$ and settles around a low value below 0.05 for $P = 1000$ persons in the data set (Fig. 2.17, 2.19). The low error rate provides evidence for the quality of the distributed combinatorial representation underlying the memory traces formed during the learning. This evidence is further fortified by the balanced form of the unit selectivity (Fig. 2.16, 2.18 **(A)**, **(B)**), which is also reflected in the approximately uniform distribution of the average unit win probability (Fig. 2.16, 2.18 **(C)**). Thus, the individual faces find their way into well-separated memory traces, the traces being appropriately distributed over the available unit resources. Each memory trace allows the recognition of a particular face experienced during the learning phase. The recognized face is represented as the composition of its local elements, or parts, selected in combinatorial fashion from the vocabularies established for local appearance.

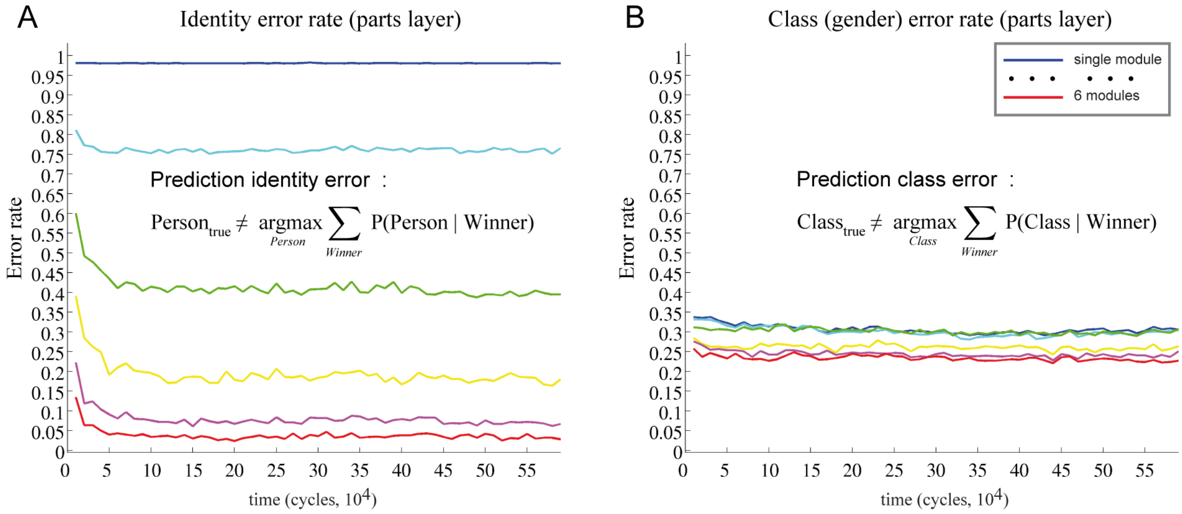


Figure 2.19: Learning error for $M = 6$ distributed isolated modules ($N = 20$ units each), $P = 1000$ persons in the training set. **(A)** Identity error. **(B)** Gender error. Although being much lower than the chance level (50%), the gender error rate gets significantly larger than in case of learning 120 persons. The clusters of local appearance seems to be less consistently dominated by only male or only female features (see also Fig. 2.18 **(E)**), which consequently leads to more errors in gender classification.

The distributed modules are also able to identify the person gender. There is an obvious discrepancy for the gender error rate between learning of 120 and 1000 persons. The gender error rate is below 5% for 120 persons, increasing to 25% for 1000 persons (Fig. 2.17, 2.19 **(B)**). This suggests that unsupervised learning of gender category has its limitations if based purely on local facial appearance. This limitation becomes apparent if a great number of different persons has to be categorized according to their gender without any instruction. It may well happen, that males exhibit locally female appearance characteristics and vice versa. In such cases, additional source of information apart from local appearance (topographic properties, mimic, haircut, voice, etc) would be necessary to enable proper categorization of a particular face.

2.5.2 Unsupervised feature extraction and gamma cycle coding scheme

The ability to perform unsupervised clustering on natural face images does not necessarily imply the ability to perform unsupervised feature, or component, extraction from the input data, which can be an arbitrary combination of those components. Here we will deal with two classical unsupervised learning tasks that demand the functionality of this kind. The first task is the unsupervised learning of an overcomplete basis for the natural image patches extracted from a set of nature scene photographs [van Hateren and van der Schaaf, 1998]. The second task is a version of the so-called bars test [Földiák, 1990]. There, the aim is to extract single vertical and horizontal bars from artificial black-and-white images presented on the input, each image being a non-linear superposition of the respective bar components.

Data format and learning procedure. For the first task, gray-level image patches of 10×10 pixels are extracted from the natural scene images [van Hateren and van der Schaaf, 1998]. The images are transformed using a difference of gaussians (DoG) filter, which coarsely mimics the effect of preceding retinal and specific thalamic (lateral geniculate nucleus, LGN) processing on the sensory visual input. The filter employs standard set of parameters, taking $\sigma_+ = 1.0$ pixel for the positive gaussian kernel and $\sigma_- = 3.0$ pixel for the negative gaussian kernel to reflect the neurophysiological plausible ratio

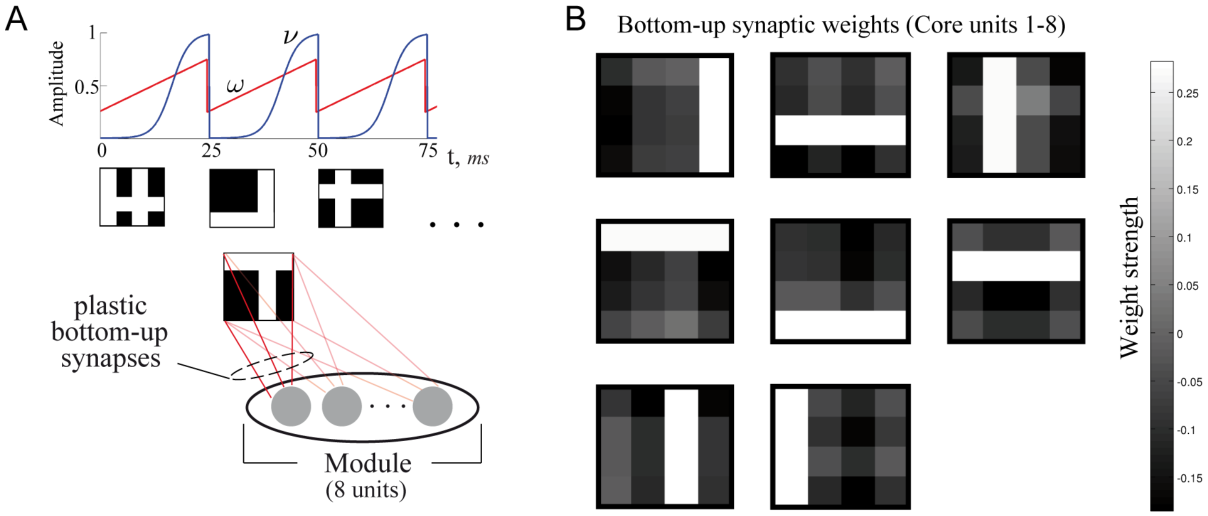


Figure 2.20: Unsupervised component extraction : bars test. **(A)** Presenting different superpositions of vertical and horizontal bars as the bottom-up input. **(B)** Synaptic weights formed after $3 \cdot 10^4$ cycles. The module manages to extract the single bars, finding the hidden components, or causes, that generate the input samples shown during the learning.

$\frac{\sigma_+}{\sigma_-} = \frac{1}{3}$ [Somers et al., 1995, Carandini and Ringach, 1997, Lücke, 2009]. For each patch, the image and the extraction location on the image are chosen randomly. The extracted patches are provided as bottom-up input to the single module of $N = 120$ units, presenting one patch per cycle (Fig. 2.21 **(A)**). For the second task, the input data set is created by superimposing $b = 4$ horizontal and vertical bars of one pixel width for each input sample (Fig. 2.20 **(A)**). The bars are represented by white pixels, while the background is black. The probability for a bar to appear on the input is uniform for all bars, $p = \frac{1}{b} = 0.25$. The input samples are presented randomly as bottom-up inputs to the single module containing $N = 2b$ units, showing one input sample per cycle. The structure parameter initialization is done in the same way as described in the preceding subsection.

Results. After running a number of decision cycles, we can take a glance on the structure of the bottom-up synaptic weights formed during the unsupervised learning. In the first task, the learning procedure ran for 10^6 cycles. Being exposed to the patches of natural images, the single module develops receptive fields that makes its units ($n = 120$) selective for certain properties of the natural input, or in other words, turn the units into filters (Fig. 2.21 **(B)**). Many of those filters have characteristics of oriented band-pass Gabor-like filters or localized blob-like filters as encountered in the primary visual cortex V1 [Jones and Palmer, 1987, Ringach, 2002]. Few of them do not seem to possess a regular structure, which is also in line with the irregular-shaped receptive fields found in the early visual areas [Martinez et al., 2005]. For the second task, about $3 \cdot 10^3$ cycles are already sufficient to develop receptive fields that contain the single horizontal or vertical bars as components of the superimposed input shown during the learning phase (Fig. 2.20 **(B)**). Thus, the single module ($n = 8$ units) is able to solve successfully the posed non-linear component extraction task.

Having developed appropriate synaptic structure to represent the input space, we may ask the question what is the coding scheme the single module now employs to reflect a given input in the activity of its units. As pointed out before, the module performs a winner-take-all-like computation on the incoming signals, the computation being executed in the repetitive frames of the gamma cycle. If new input is presented in each cycle, then the two-phase WTA coding will select the unit candidates and then amplify the strongest one while suppressing the rest on the basis of the input strengths. What happens

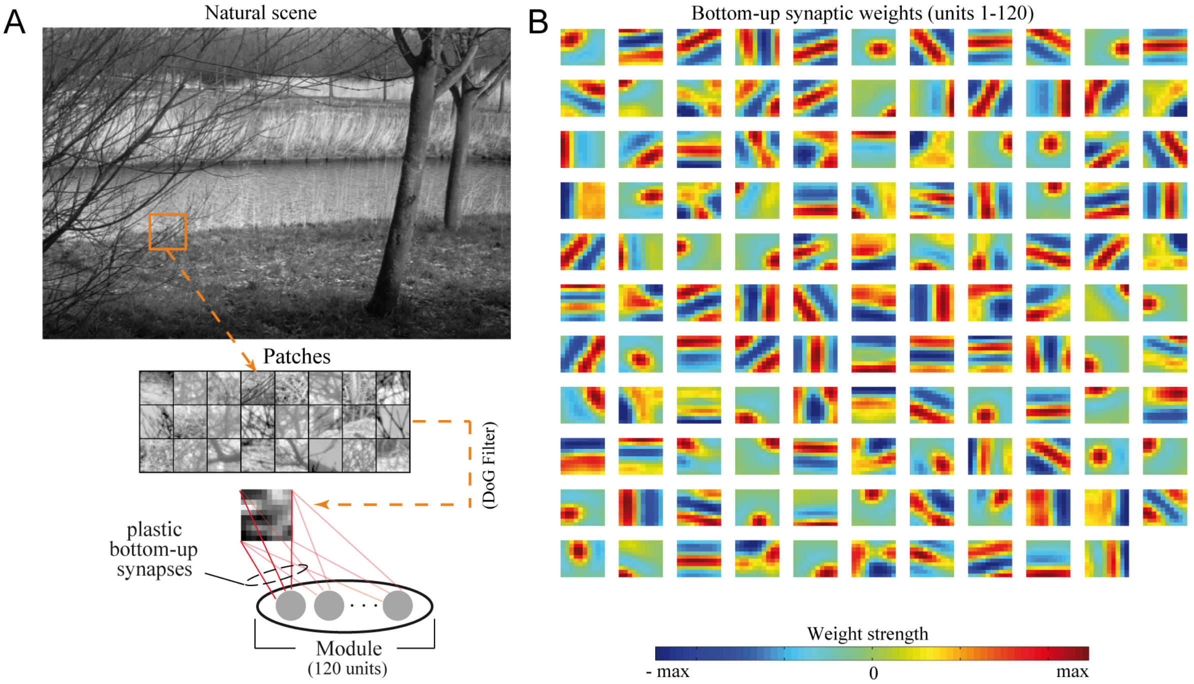


Figure 2.21: Unsupervised formation of an overcomplete basis for natural image patches. **(A)** Learning procedure as described in the text. **(B)** An overcomplete basis (120 filters) formed for the 10×10 image patches, showing synaptic weights after $8 \cdot 10^5$ cycles.

if instead the same input is presented and kept over the duration of multiple cycles?

The result shown in the Fig. 2.22, 2.23 provides an interesting view on the coding scheme which may be hypothetically employed in the cortical processing [Zhang and Ballard, 2002, Fries et al., 2007, de Almeida et al., 2009]. In this scheme, an atomic coding fragment is the single decision cycle in gamma range. The strong competition between the units governs the activity formation within each single cycle, resulting in only one winner unit being highly active at the cycle's end. If the presented input is a composition of multiple components that were captured during the learning by the module units, the hard winner-take-all computation alone would not be appropriate for reflecting the input structure. Within the single cycle, the two-phase WTA coding may relax the hard winner-take-all nature of the computation by representing the alternative candidates in the first, soft-WTA phase of the cycle. However, if the read out is done at the late cycle phase, only the winner unit is visible.

Now, if looking on the distribution of the winner units over multiple successive cycles, it becomes apparent that different candidate units can be chosen to be the winner of a cycle, signaling one of the components that make up the current input sample. In the case of a simplified situation where artificial image of two crossing bars is presented (Fig. 2.22 **(A)**), so that each component making up the input can be clearly identified, the competitive computation picks out only one of the units responsible for the respective bar to be the winner within each cycle. This selection can be interpreted in probabilistic terms, assigning each unit a certain win probability given the current input. The both candidate units obtain in this case equally high probability to win ($P(winner|input) \approx 0.5$). The candidate selection happens in alternation over cycles. If one unit becomes the highly active winner $p > 0.45$, the other has to skip the cycle and wait for the next opportunity (Fig. 2.22 **(B)**). The rest of the units that is not selective for the components on the current input has consequently a very low win probability close to zero. If we compute the average unit activities over the whole duration of the cycle sequence, we

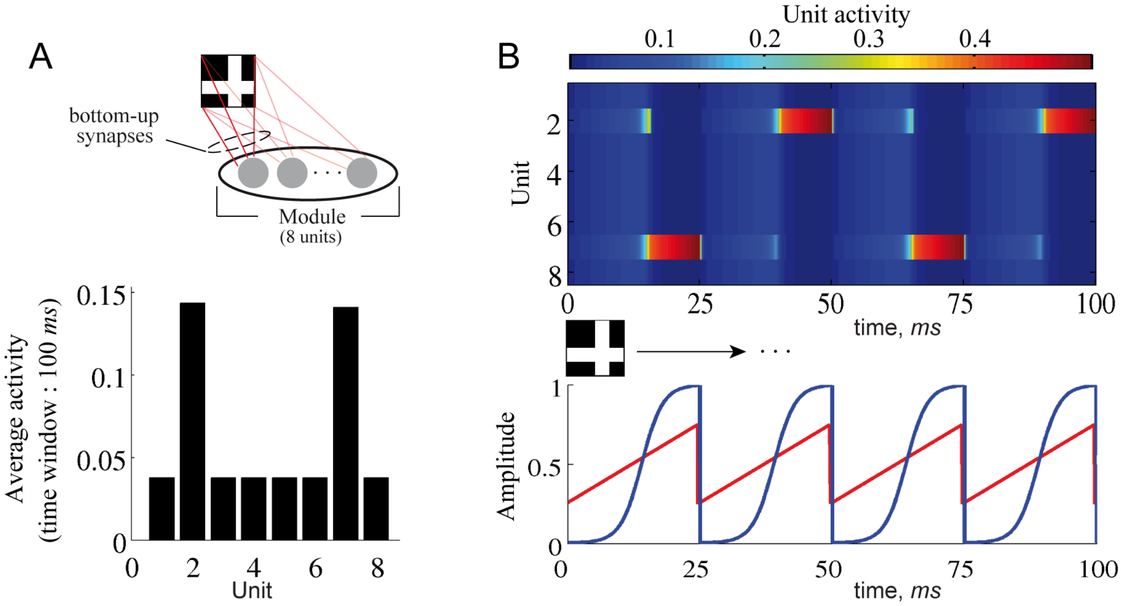


Figure 2.22: Gamma coding scheme, crossed bars example. Two crossed bars are fixed on the input over multiple cycles (here 4). **(A)** The average activity measured over multiple cycles (here 4, $100ms$ time interval) reveals both units responsible for corresponding bars responding with roughly equal low rate ($\bar{p} < 0.15$). **(B)** Without averaging, different coding picture emerges. In each gamma cycle only one of the responsible units becomes a winner, suppressing the rest. The winner gains in the late cycle phase high activity ($p > 0.45$). The selection alternates over cycles, assigning the candidate units a certain win probability that corresponds to their responsibility for the current input. Consequently, the candidate units tend to skip cycles in this probabilistic computation.

obtain a roughly equal low activity for the both candidate units ($\bar{p} < 0.15$) and an even lower baseline activity for the rest (Fig. 2.22 **(A)**). Thus, if the measurement procedure is extended over a long time interval, the fine structure of activity formation during the competitive decision cycle gets lost. The short discrete fragments of high activity corresponding to the win events disappear and we get to see only the low average activity description of the presented stimulus.

The same kind of probabilistic computation also takes place in a more complex situation, where the components making up the input are not any longer clearly identifiable, like it is in the case of an natural image patch (Fig. 2.23 **(A)**). The overcomplete basis formed by the module units is used to represent the given image patch in a sparse fashion. Within the cycle, and particularly close to its end, the sparseness is extreme as only one winner unit corresponding to a suitable basis function is selected to become highly active for a short time ($p > 0.45$). Again, if observing the unit responses over longer time across many cycles, we see few different units that are selected to represent the current input in different cycles. The candidate units are the basis functions that are taken to represent the current stimulus. Within one decision cycle, only one of those units is able to stay highly active at the end of the short processing fragment, while the others have to skip the cycle. Over the longer time interval, each candidate unit get selected multiple times to signal its participation on the encoding of the current input. These selection is again interpretable in terms of assigning to different candidates different probabilities to win within each cycle according to their suitability as a basis function for the given input. The activity sparseness is high not only within but also across the cycles, as only few basis functions from the overcomplete set are picked out to represent the input (Fig. 2.23 **(B)**).

As in the previous scenario, computing the average unit activity over many cycles discards the

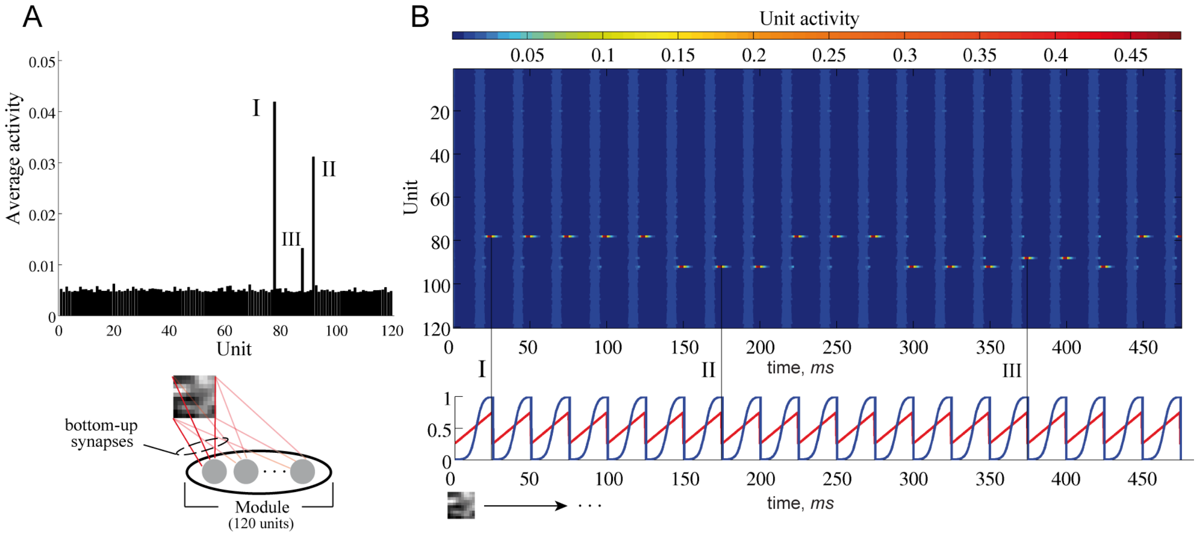


Figure 2.23: Gamma coding scheme, natural image patch example. A natural image patch is fixed on the input over 20 cycles ($500ms$ time interval). **(A)** The average activity measured over 20 cycles ($500ms$ time interval) reveals three candidate units responding to the image patch with very low rate ($\bar{p} < 0.05$). **(B)** Again, different coding scheme is revealed if the activity is not subject to averaging. Only one of three candidate units becomes a winner in each gamma cycle, reaching a highly active state ($p > 0.45$). Alternating probabilistic selection of a winner unit makes the candidate units occasionally skip the cycles. A repetitive winner-take-all computation performed within successive gamma cycle frames produces thus short discrete fragments of high activity. If averaged over time, this signaling behavior may appear as unit activation at a very low rate.

fine structure of activity formation within the cycle. Importantly, the short fragments of high activity ($p > 0.45$) get lost in this averaging procedure. Instead, the input seems to be encoded by the very low average activity rate with $\bar{p} < 0.05$ (Fig. 2.23 **(A)**). Thus, if a similar coding scheme is indeed employed in the cortical processing as suggested here and in [Zhang and Ballard, 2002, Fries et al., 2007, de Almeida et al., 2009], in order to see the fine structure of competitive activity formation one has to be careful in choosing not only the right population of tightly interconnected pyramidal neurons to measure the responses from, but also the right time scale (about the half of the gamma cycle, $t \approx \frac{T}{2} = 12.5ms$ to measure the average population activity, as otherwise the picture of neuronal signaling would appear completely different.

The varying selection of the unit candidates over multiple decision cycles depends here on two mechanisms of the unit dynamics. First, the homeostatic activity regulation makes a unit less susceptible for the excitatory input if it gets activated too often. This resembles an adaptation, or fatigue, effect, which makes the winner unit to loose competition against a candidate that may have less stronger input, but is more excitable. The time course of the fatigue effect depends then on the time constant of the intrinsic plasticity (see Eq. 2.2.1 and Tab. 2.1). Second, the level of neuronal threshold noise (Eq. 2.1.13) determines the probability of the candidate units to win the competition against the unit with the strongest input. The higher the noise level, the more probable is the win event of a candidate unit with less stronger input. Conversely, if the noise level is low, the unit selection becomes more conservative, allowing only the unit with the strongest input to win the competition most of the time. The threshold noise can be thus seen equivalent to the temperature parameter T, sometimes referred to as computational temperature, which is widely used in different settings, like statistical mechanics, stochastic optimization or reinforcement learning to determine the level of system's entropy and enable transition to the less probable states that otherwise cannot be visited [Ackley et al., 1985, Barto and Anandan, 1985, Hertz et al., 1991, Doya, 2002].

The dynamics of the module that governs the selection of the unit candidates from cycle to cycle can be also interpreted in terms of transient visiting different attractors in the phase space of unit activities [Horn and Usher, 1989, Herrmann et al., 1993, Rabinovich et al., 2000, 2008, Durstewitz and Deco, 2008]. Different components, or basis functions, extracted during the learning form attractors of the module's dynamics. Given a certain input, the winner-take-all computation tends to select the unit with the strongest input, so that the active winner unit and the silent rest define the initially visited attractor state. The initial attractor state gets unstable with time as homeostatic activity regulation renders the winner unit less and less sensible to the incoming input. Once the excitability level of the winner unit gets low enough, the attractor state becomes unstable, or ruined, so that a transition to another attractor state with the next suitable candidate occurs. Furthermore, if the noise level is made high enough, a transition from any attractor state to another one, where a different unit candidate becomes a winner becomes possible. So, while the synaptic structure and WTA operation define the attractor landscape, the intrinsic plasticity of unit excitability and the neuronal threshold noise instantiate a type of transient attractor dynamics in the module, shaping the course of the transitions from one attractor to another given a particular input.

2.6 Discussion

Aiming at neural network architecture that can build up a hierarchically organized memory domain in an unsupervised fashion, a model of a cortical module that will act as an elementary computational node in the network was developed and tested in this chapter. There are two essential core functionalities the module was shown to be capable of. First, the module is able to develop an appropriate local vocabulary of elements, or descriptors, for the locally accessible input space, covering it with a compact set of representative units. The units acquire their selectivity by shaping their synaptic structure in the course of unsupervised learning under exposure to external input. Second, the module is able to process the input rapidly within a single gamma cycle and choose a winner unit to represent it in a sparse way. The successful instantiation of the local competitive unsupervised learning was demonstrated in the task of learning local facial features with single and distributed modules. The modules were able to form local vocabularies for facial features and recognize the identity and gender of different persons shown during the learning. In addition, a single module was shown to be capable of unsupervised component extraction, forming an overcomplete basis of oriented filters from natural image patches and solving the classical bars test.

Neuronal mechanisms of unsupervised competitive learning. The model of a cortical module presented here is based on some core assumptions about the form of neuronal interactions within the microcircuit of a small local cortical patch, or cluster. This form is mainly sculptured by the interplay between the strong excitatory coupling among the pyramidal neurons defining the segregated fine-scale functional subnetworks and the global unspecific inhibition acting upon all of them [Yoshimura and Callaway, 2005, Yoshimura et al., 2005, Haider and McCormick, 2009]. The interaction between specific excitation and global inhibition is the neuronal substrate that mediates the competitive nature of processing within the module on the fast time scale. The adaptive mechanisms of bidirectional plasticity and homeostatic activity regulation recruit the competitive processing on the fast time scale to perform an advanced kind of competitive learning on the slow time scale. This kind of functionality supports the hypothesis stating that the competitive WTA-like processing and learning is a basic computation performed by local cortical circuits [Douglas et al., 1995, Douglas and Martin, 2007, Rozell et al., 2008, Litvak and Ullman, 2009], being carried out in discrete fragments, or cycles, of an ongoing gamma frequency rhythm ($40 - 100$Hz) [Zhang and Ballard, 2002, Fries et al., 2007, Börgers et al., 2008, Burwick, 2009, de Almeida et al., 2009]. Within each cycle, selection and graded amplification of the candidate units occur, so that these candidate units get an opportunity to modify their synapses according to the strength of their input and the level of their activity.

The notion of competitive processing and competitive learning has a long tradition in computational neuroscience and machine learning [von der Malsburg, 1973, Grossberg, 1976b, Kohonen, 1982, Rumelhart and Zipser, 1985, O'Reilly, 1998]. The computation the module performs on the incoming inputs can be considered as a flexible WTA operation. In this kind of competitive processing, there are two different phases corresponding to the two different WTA schemes, the soft and the hard one. So, during a large part of the competitive decision cycle, not only the activity state of the winner unit is accessible for read out, but also the activity states of a number of winner opponents. Importantly, the unit activity states are graded, depending explicitly on the input strength. This is in contrast to the standard WTA implementations, where activity states are essentially binary, signaling "on" (winner) or "off" (looser) status only. The flexibility of the operation is also due to the possibility to influence the duration of the different phases and the degree of sparseness within the phases, controlling the number of units that may become active. This can be done by simple tuning of the amplitudes of the ongoing rhythms ν and ω (Fig. 2.8). The tuning can be potentially performed on the fly via a local or global mechanism, adapting the coding scheme to the current processing demands.

Considering unsupervised competitive learning, the adaptive mechanisms employed here impose important constraints on the formation of synaptic structure that defines the unit selectivity and thus the vocabulary for the local input space. The homeostatic regulation of unit activity encourages a uniform duty cycle across units in the module. This assures the equal participation of the units on vocabulary formation during the learning phase. Thus, in the long term each unit gets on average equal opportunity to win and shape its receptive field, becoming selective to a certain aspect of the input space. This is crucial for avoiding the classical dead unit problem, which arises if a learning procedure cannot properly balance the unit usage load [Rumelhart and Zipser, 1985, Grossberg, 1987a]. This failure leads usually to improper partitioning of the input space into large clusters that correspond to few overactive units, while other units are left unused, or dead. Employing an adaptive unit threshold to prevent the dead unit problem on the basis of previous unit activation history is a classical technique in competitive learning [Grossberg, 1976a, Földiák, 1990], also known as fair, frequency-sensitive or conscience winner-take-all mechanism [Desieno, 1988, Ahalt et al., 1990].

The balanced unit usage load alone cannot guarantee the successful partitioning of the input space into well-separated clusters. It may happen that all units learn to respond only to a restricted region of the input space, ignoring the rest, or conversely become selective to a property which is common to all the input patterns, not being able to discriminate between them. For the successful separation, the competition in synaptic formation *across* the receptive fields of the units becomes crucial [Bienenstock et al., 1982]. Here it is instantiated by combining the WTA-like operation performed in the course of a decision cycle with the activity-dependent bidirectional plasticity that can either potentiate or depress a given synapse. The consequence is that the synapses can only grow strong on the expense of weakening others. This competition supports differentiation between input patterns that have to be considered as distinct, despite being highly similar in their raw appearance. This is known to pose great difficulties for the common approaches to competitive learning [Rumelhart and Zipser, 1985, O'Reilly and Munakata, 2000, O'Reilly, 2001, Kohonen, 2001]. The procedure implemented here in neuronal fashion by the WTA-operation combined with mechanisms of synaptic and intrinsic plasticity can be also related to the rival penalized competitive learning algorithm [Xu et al., 1993, Xu, 2002]. There, a similar mechanism that attracts the winner towards and pushing the competitors, or rivals, away from the given input is employed to enable proper cluster formation over the input space.

The unsupervised clustering is one kind of vocabulary formation offered by the module. From the perspective of clustering, classical competitive learning can be interpreted as a discrete approximation of a generic density estimation algorithm that seeks maximum likelihood mixture-of-Gaussians model for the input data [Becker and Zemel, 2002]. This type of basis formation may be inappropriate, if the inputs are superpositions of different independent sources, or components, that are non-Gaussian, or sparse [Hyvärinen and Oja, 2000, Bartlett et al., 2002, Asari et al., 2006]. Confronted with such a source separation task, the single module is able to extract the components from the input samples presented during learning, which corresponds to finding hidden causes responsible for generation of the input. The established vocabulary can be then used to represent the incoming input as combination of the components contributing to the current input sample. Independent of the nature of a particular unsupervised learning task, the learning is in general self-regulating and life-long. As such, it does not have to rely on time-dependent decrease, or freezing, of learning rate or on an explicitly defined stop condition, which are commonly specified by hand in many standard approaches to prevent destabilization of the learning procedure.

Hypothetical role of the gamma rhythm in competitive learning. The coding scheme that emerges from the module's dynamics supports hypothetical role of the gamma cycle as an atomic fragment of the ongoing probabilistic signaling and decision making in the cortex. This coding scheme is a consequence of a repetitive execution of the WTA computation carried out in the successive gamma cycle frames. The WTA computation can be hypothesized to be carried out rapidly within only a single cycle due to very low latencies of fast inhibition [Connors and Gutnick, 1990, Galarreta and Hestrin, 1999, Yoshimura and Callaway, 2005, Xu and Callaway, 2009] and due to the ability of the excitatory population of tightly coupled pyramidal neurons to generate an almost instantaneous robust response [Gerstner, 2000]. The probabilistic selection of a winner unit makes units participating in encoding of the current input to skip cycles occasionally. A repetitive WTA computation performed within gamma cycle frames produces short discrete fragments of high activity. If averaged over a longer time interval, these short isles of high activity get lost. The signaling may then appear as unit activation that happens at a very low rate because of occasional probabilistic cycle skipping (see also Fig. 2.22 **(A)**, 2.23 **(A)**).

Therefore, if a similar coding scheme is indeed employed in cortical processing as suggested here and in [Zhang and Ballard, 2002, Fries et al., 2007, Börgers et al., 2008, de Almeida et al., 2009], in order to see the fine structure of neuronal population signaling one has to be careful in choosing not only the right population of tightly coupled pyramidal neurons to measure the responses from, but also the right time scale to measure the signals. Importantly, the activity variable p is *not* a mean-field-like population activity in the classical sense. To measure p experimentally, the population averaging has to be done *not* simply over an arbitrary spatially neighboring population of neurons. Instead, the functional subnetwork corresponding to the highly specific interconnected population has to be identified first [Song et al., 2005, Yoshimura et al., 2005, Haider and McCormick, 2009]. This can be done by precise focal stimulation of single cells using optogenetic methods [Boyden et al., 2005, Miller, 2006] or local glutamate uncaging ("photostimulation") [Boucsein et al., 2005, Shoham et al., 2005] technique and subsequent whole-cell recording of the pyramidal cell responses from the local cortical patch of approximately $600 - 800\mu m$. After identifying the functional fine-scale subnetwork, which may have a complicated and widely extended layout, p can be measured by simultaneously recording the pyramidal neurons comprising the network and computing the average population rate. Moreover, the averaging has to be performed over a sufficiently small time window (smaller than the half of the gamma cycle, $t \approx \frac{T}{2} = 12.5ms$), as otherwise the discrete, fragmented nature of the coding scheme would be no longer observable (see also Sec. 2.5.2).

Cortical processing seems to make abundant use of oscillatory rhythms in the gamma range [Gray et al., 1989, Steriade et al., 1993, Traub et al., 1996, Buzsáki and Draguhn, 2004], and it has been hypothesized here and in other work that these rhythms define a decision cycle, an atomic fragment of competitive activity formation [Zhang and Ballard, 2002, Fries et al., 2007, Börgers et al., 2008, de Almeida et al., 2009]. Another classical hypothesis assigns to such rhythms a role in the coordination of distributed neuronal signaling and in the formation of coherent neuronal assemblies - a functionality, which is often referred to as binding [von der Malsburg, 1981, Engel et al., 1992, von der Malsburg, 1995b, Buzsáki and Chrobak, 1995, Singer, 1999, von der Malsburg, 1999]. Interestingly, there is evidence that background oscillations in the cell membrane potential modulate synaptic plasticity in cortical neurons [Huerta and Lisman, 1995, Wespatat et al., 2004]. According to this evidence, the potentiation and depression of a synapse occur preferentially at the peak and at the trough of the oscillatory cycle. This modulation can be intuitively explained here by the two-phase WTA operation and the form of the bidirectional plasticity rule. So, the sliding thresholds in the plasticity rule permit potentiation only for the highly active winner unit at the end of the cycle, whereas depression is restricted to the units that are of average activity in the early phase of the cycle and get deactivated later. As there is also support for a phase reset mechanism that locks the ongoing oscillatory activity to the presented stimulus [Makeig et al., 2002, Axmacher et al., 2006], the gamma rhythms can be also hypothesized to coordinate not only activity, but also synaptic modification and thus memory storage across the neuronal units distributed in the network [Harris et al., 2003, Buzsáki and Draguhn, 2004, Axmacher et al., 2006].

Extreme sparseness within the module. Within a singe cycle and across multiple cycles, the activity generated in the module is rendered sparse due to the competitive nature of the computation. There is vast amount of neurophysiological evidence supporting a sparse coding scheme in the cortex [Barnes et al., 1990, Young and Yamane, 1992, Rolls and Tovee, 1995, Olshausen and Field, 1997, Vinje and Gallant, 2000, Weliky et al., 2003, Quiroga et al., 2005, Axmacher et al., 2008a, Quiroga et al., 2008]. If learning an overcomplete basis to cover the local input space, the sparse representation that picks only few basis functions to interpret the incoming inputs is known to have several crucial advantages. On the one hand, it is easy to read out and it optimizes information transmission in general [Field, 1987, Olshausen and Field, 1996, Simoncelli and Olshausen, 2001, Földiák, 2002, Olshausen and Field, 2004], allows faster learning and better generalization by being related to regularization techniques that prevent overfitting [Bishop, 2006, Asari et al., 2006, Rehn and Sommer, 2007], and it also may contribute to greater memory capacity [Palm, 1980, Tsodyks and Feigel'man, 1988, Buhmann et al., 1989, Okada, 1996, Sommer and Palm, 1999, Rehn and Sommer, 2007]. In addition, sparse codes have low metabolic cost, which is important not only for the processing in the brain, but for any system that has to minimize its power consumption [Levy and Baxter, 1996, Attwell and Laughlin, 2001, Laughlin, 2001, Lennie, 2003, Vincent et al., 2005].

The kind of sparseness utilized here within a single decision cycle may be termed hard sparseness, as it puts a hard limit on the number of units within the module that are able to stay active, as opposed to soft sparseness, that limits the average unit activity only [Rehn and Sommer, 2007]. Although in principle tunable via modification of the ongoing rhythms ω and ν, the sparseness instantiated here allows only one unit within the module to survive at the cycle's end. This is still appropriate for the intended memory architecture because of at least two reasons. First, most importantly, the target architecture is a multi-layered network composed of many modules. As the hard WTA operation is restricted locally to act only among units within a module, in the network there will be multiple units, one per module, that stay active at the end of the cycle given a particular input. This coding is sparse, but it is not anymore the extreme, localist representation scheme as given by the hard winner-take-all operation [Földiák, 2002]. Second, it is reasonable to assume that the structure of the local, restricted input space a single module gets to look at is not too complex, such that it may well be sufficient to encode each incoming input sample by a single winner unit. Taken together, the winner units from different modules in the network should then be able to provide an accurate sparse description of the global, complex stimulus as combination of its much less complex constituents.

A transient attractor dynamics implemented by the module. The ongoing competitive computation performed by the module can also be interpreted in terms of transient attractor dynamics [Herrmann et al., 1993, Friston, 1997, Rabinovich et al., 2008, Durstewitz and Deco, 2008]. The selectivity of the units, given by their synaptic weights formed during learning, defines a point attractor landscape in the phase space of unit activities. Each point attractor corresponds to a module state where all units except the winner unit are deactivated. This state can be labeled by an element from the local vocabulary to which a particular unit is sensitive. Given an input, the system gravitates toward a point attractor with the closest label, or in other words, toward the winner unit which prefers most strongly the current input. Arrived in an attractor state, the system does not stay there forever. The mechanism of homeostatic activity regulation updates continuously the intrinsic excitabilities of the module units, so that at some point the previously selective unit fatigues and abandons its preference for the current input, giving the next best candidate the opportunity to respond. In other words, an attractor becomes unstable, or ruined, and a transition to the next one occurs. This situation occurs over and over, so that the system continues visiting different attractor states in a certain order dependent on the input presented to the module. The trajectory of the attractor visit is largely determined by the dynamics of the homeostatic activity regulation. Neurophysiologically, different regulatory mechanisms that act on different time scales may contribute to this type of transient attractor dynamics. To reveal the individual contribution of different regulatory components, a more detailed modeling of distinct synapse-unspecific mechanisms of activity-dependent cell excitability tuning, like cell adaptation, fast and slow intrinsic plasticity [Nelson and Turrigiano, 2008], is required in the perspective.

Heading for the network. The module designed here is prepared for the function in a multi-layered memory network. There it will serve as a container that, depending on the hierarchy level, learns and holds either a collection of local image descriptors (parts) or an explicit set of global identities. The very crucial issue in such a network is appropriate communication between the distributed modules within and across the stages of the hierarchy. The recurrent signaling between the modules should allow the right interpretation of the compositional identity of the stimulus from the local ambiguous information accessible to each module. This communication has to be learned by simultaneous formation of proper lateral and top-down network pathways. The previous approach [Lücke, 2005, 2009] turned out to cause severe difficulties if hosted within a network of multiple layers. There, the information exchange between interconnected modules seems to fail, so that their internal activity states cannot be properly communicated across the network. Two essential properties of the model introduced here should overcome this drawback. First, the two-phase WTA coding scheme provides a graded representation of the incoming input and gives an opportunity to read out the graded internal state of the module during a sufficiently large fraction of the decision cycle. Second, separation of synapses in functionally different groups according to their hierarchical origin should enable proper integration of different kind of evidence into the local decision making performed by the modules. The following chapter will show whether these expectations are justified and the modeling done on the level of a single computational node can be successfully transplanted to the level of the network architecture.

3

A self-organizing hierarchical visual memory: unsupervised learning of a generative compositional object representation

Substantial neurophysiological and neuropsychological evidence suggests that the brain uses parts-based representations to handle arbitrary complex objects [Fujita et al., 1992, Tsunoda et al., 2001, Tanaka, 2003, Biederman and Cooper, 1991, Ullman et al., 2002, Hayworth and Biederman, 2006]. For instance, the primate visual cortex is able to process effortlessly natural visual objects by decomposing them rapidly into constituent components [Riesenhuber and Poggio, 1999, Pasupathy and Connor, 2002, Rousselet et al., 2004, Reddy and Kanwisher, 2006]. These components grow in their complexity along hierarchically organized visual pathways. In this way, the compositional nature of the visual object could be captured in a nested hierarchy of parts and their relations. Within a processing stage, the parts could be associatively linked together if they belong to the same object of higher complexity. Across the stages, part-of-the-part relations could be established. Such relations would then define a generative object model, as they would allow to reconstruct, or generate, full parts-based description of an object given a high-level information about its identity.

It is clear that the neuronal populations and the synaptic connectivity have to provide the necessary neuronal substrate for the representation of the parts and the relations between them. However, it is far from being clear how this substrate gets shaped to become a functional processing structure. In other words, it is largely a mystery how the cortex forms and maintains its memory domain for storing and recalling objects of natural complexity.

In order to develop a memory domain that supports a generative, compositional object representation, two difficult tasks of unsupervised learning have to be addressed simultaneously by the cortex and, in general, by any system that aims to build up such a memory through experience with the objects. First, the system has to develop a vocabulary of universal primitives, or parts, of intermediate complexity that can be re-used again and again for compositional representation of more complex objects. Once learned, this reusable vocabulary, or alphabet, has tremendous advantage of combinatorial expression power. That is, any natural object of arbitrary complexity could be represented by picking a sparse subset of parts from the already existing overcomplete vocabulary [Brincat and Connor, 2004, Connor et al., 2007]. Thus, representing and learning a novel object not seen before becomes tractable with such a vocabulary at hand. There would be no need to insert physically new neuronal units to encode the identity of a novel object. The object identity would be decomposed into description of already existing, reusable constituents of much lower complexity, providing readily the substrate for a memory trace.

Second, to capture explicitly the compositional object identity in generative manner, the different relations between the vocabulary parts have to be learned from experience with different objects. The parts themselves are only descriptors of local appearance and its variation. The local appearance however is often highly ambiguous, so that the correct decision about the global identity of the object can be made only if relying on additional, context-based information formed by previous experience. It is the relations between the parts which have to provide this type of information. Learning these relations would then establish memory traces, each comprising a subset of explicitly linked parts that compose a hierarchically stored object identity. During the recognition, or recall, of an already experienced object the context-based information captured in the memory traces could be used to resolve the local ambiguities and arrive at a coherent correct interpretation of the global stimulus.

How can these tasks of finding local vocabularies and establishing appropriate relations between their elements be solved simultaneously, by learning in unsupervised, incremental fashion from the sensory inputs? To address this question, I set up in this chapter a self-organizing two-layered network of distributed cortical modules, installing full plastic synaptic connectivity between the modules within and across the layers. The network connectivity is completely homogeneous in the initial state, without incorporating any kind of pre-specified structure.

As it was already shown in the previous chapter, it is possible for a number of distributed isolated modules to build up vocabularies for local appearance of faces shown during an unsupervised learning stage (Sec. 2.5.1). However, the local vocabularies were neither linked to each other in terms of building up relations between their elements, nor was there a hierarchy for the explicit representation of global face identity. Now, the two-layered network has to develop fully recurrent memory structure, learning not only bottom-up, but at the same time also lateral and top-down connectivity within and between the layers.

This memory structure will serve as a storehouse for associatively linked facial parts laid within lower layer and for person identities emerging as higher-order symbols on the higher layer. A face presented on the input is then encoded by the network in the form of a winner unit assembly. The winner unit assembly is constructed in the course of a single gamma decision cycle in the gamma range by selecting and amplifying one winner unit per module. The winner assembly represents the current face as composite of parts and a higher-order identity label. These parts-based representations are sparse in terms of both the activity generated during the recall and the synaptic patterns constituting memory traces in the network.

Most crucial, the local decisions performed by the distributed modules are no longer independent in such a memory network. In the mature connectivity state, the network decisions about local facial parts and their composition into a global face identity are the result of signal exchange between the modules. This contextual interaction between the modules is mediated by the recurrent lateral and top-down pathways formed during learning. In addition to the purely local WTA competition, which is restricted to act only among units within the single modules, the processing in such a recurrent hierarchical network involves global cooperation and competition between the units across the modules. The decision in favor of a particular unit assembly can thus be seen as result of integration of local sensory cues with global contextual priors. This integration makes possible an interpretation of the incoming stimulus which is coherent among the modules by using the knowledge gained in the course of previous experience.

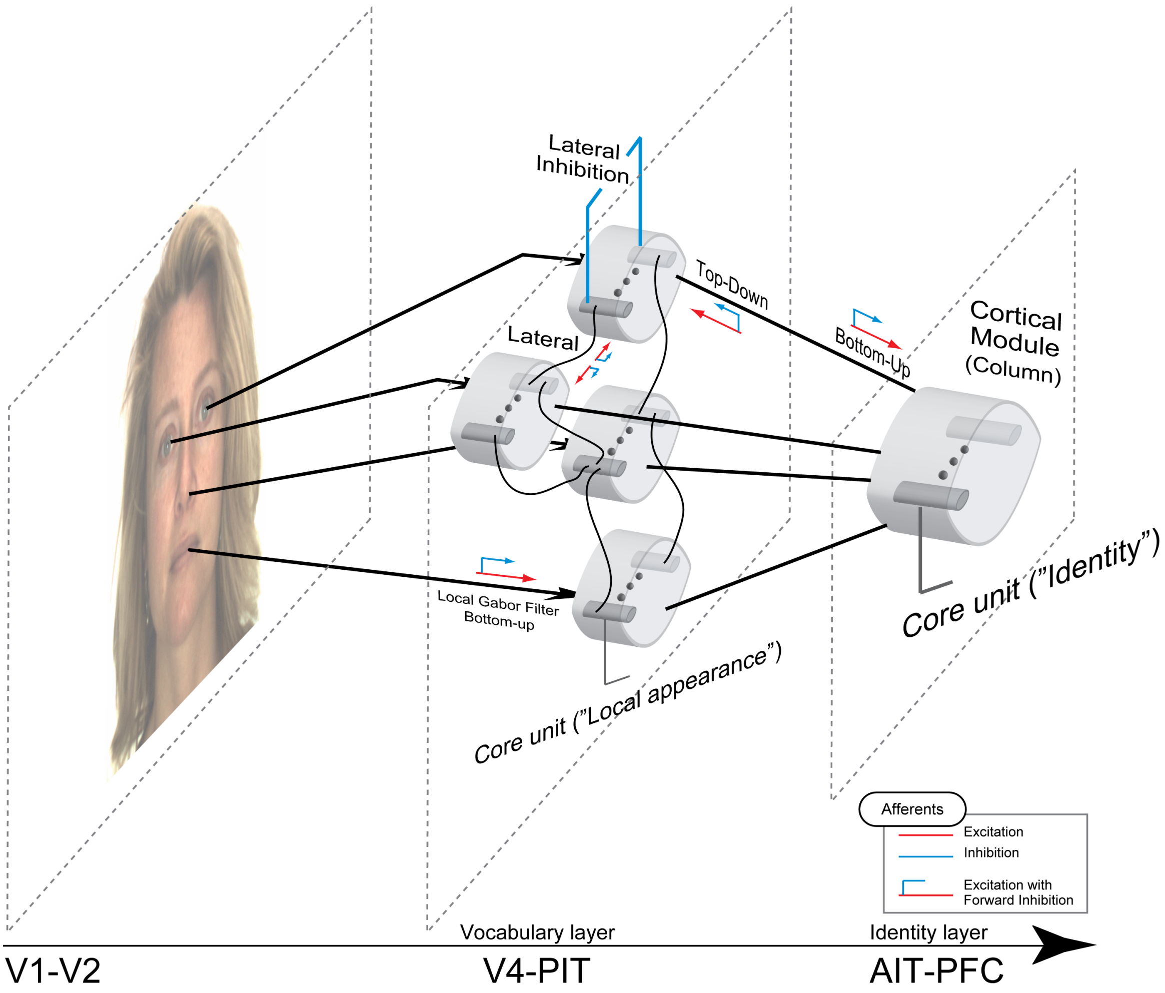


Figure 3.1: Hierarchical visual memory model. A hierarchy of two recurrently interconnected memory layers to store and represent faces in combinatorial, parts-based fashion. The network structure comprising bottom-up, lateral and top-down connections is formed by unsupervised learning from natural images. The lower layer (V4-PIT, resembling a region of the posterior inferotemporal cortex) is composed of modules containing $N = 20$ units each. It learns vocabularies of local image descriptors at landmarks (provided by Gabor filter responses that mimic the processing in early visual cortex V1, V2) and links them associatively with lateral connections. The higher layer, called identity layer (hypothetically residing in a region of anterior inferotemporal or prefrontal cortex, AIT-PFC), contains one module of $N = 40$ units, each developing into a symbol for one person identity. The identity units provide further contextual support for the lower layer by learning top-down projections to the corresponding part-specific units.

The emerging memory network will be tested for its ability to recognize identity and gender of the persons from alternative views not shown during the learning phase. Moreover, two different network configurations will be employed to compare a purely feed-forward architecture with a fully recurrent one in terms of their recognition and generalization capabilities. I will discuss further different properties of the processing in the network during the recall of the stored content, like maintenance of stimulus-induced activity, generative pattern completion, mechanisms of bottom-up and top-down attention and self-generated memory replay in the absence of external input. A rapid, synapse-unspecific learning that improves memory performance without modifying synaptic connectivity will be highlighted. Finally, I will make some remarks on the scalability of the network architecture.

3.1 Unsupervised learning of object identity and category

3.1.1 Network architecture, configurations and experimental setup

Network architecture. The model of a hierarchical visual memory is based on two consecutive, reciprocally interconnected layers of distributed cortical modules introduced in Chap. 2 (Fig. 3.1, 3.2 **(B)**). These layers can be seen in rough correspondence to subregions of the hierarchically organized inferior temporal cortex (IT), like PIT (posterior IT) and AIT (anterior IT). These areas are known to be organized in small clusters or modules, classically termed "columns", of neurons selective to similar shapes or objects [Fujita et al., 1992, Tanaka, 1996, 2003, Sato et al., 2009a]. The cortical regions located even higher in the processing hierarchy, like areas of the prefrontal cortex (PFC), may also contain clusters selective for object categories [Freedman et al., 2001, 2003]. Here I denote the lower memory layer as *parts* or *vocabulary layer* and the higher memory layer as *identity* layer, applying the same terminology to the layer-specific modules. There are $M = 6$ vocabulary modules situated on the lower layer, each containing $N = 20$ units attached to a dedicated landmark (see also Sec. 2.5.1). Each vocabulary module will serve as container for reusable descriptors of the local appearance which it has first to acquire in the course of learning. The units of the identity module on the higher layer, in distinction, have the task to explicitly capture and signal global face identities of different persons. In the main experimental setup, the identity module contains $N = 40$ units for storing $P = 40$ persons shown during the learning phase (in experiments on scalability, the identity module will contain $N = 120$ units, corresponding to the number of persons used there). As each module receives now afferent connections from a number of other modules within the network, the coupling of afferent input is set to a factor $1/M$ for the identity module and $1/(M-1)$ for the vocabulary modules, compensating the increase in total incoming input.

Alternative network configurations. The modules are interconnected within and across the layers via modifiable excitatory synapses. The fully recurrent network configuration, which is the default network setup here, employs all types of plastic synapses - bottom-up, lateral and top-down (for a detailed description of synaptic separation within the module, see Sec. 2.1 and Eq. 2.1.13, 2.1.16b; the synaptic plasticity rule is described in Sec. 2.3). This full connectivity has to be learned from experience with natural face images. An alternative, purely feed-forward network configuration is restricted to use bottom-up connectivity only. The two network configurations are trained independently on the same face image data sets, and their recognition performance is compared later.

Open-ended learning procedure. Before learning, the initial conditions and parameter setup are as described in Sec. 2.4 and Sec. 2.5.1. Again, the network is initially unstructured, intermodular connectivity being all-to-all between the units within and across the layers. A training set containing neutral face views of $P = 40$ persons randomly selected from the AR database Martinez and Benavente [1998] is used for learning (Fig. 3.2 **(C)**). Images are presented incrementally, showing one image per decision cycle (Fig. 3.2 **(A)**). Each image is fed to the lower network layer as a collection of $L = 6$ Gabor filter bank responses extracted locally from the respective landmarks (Fig. 3.2 **(B)**). The network connectivity structure changes in time according to the synaptic plasticity rule described in Sec. 2.3. The learning is open-ended, and the network is stopped for testing as soon as sufficiently mature connectivity has developed.

Performance evaluation. To assess the recognition performance of the memory network, I make a distinction between the learning error, block generalization error and immediate generalization error. The learning error for identity/gender was already defined as the rate of wrong responses to person identity/gender from the training data set used during the learning phase (see Sec. 2.5.1). The generalization error is computed on sets of alternative views not presented before (Fig. 3.2 **(C)**). During

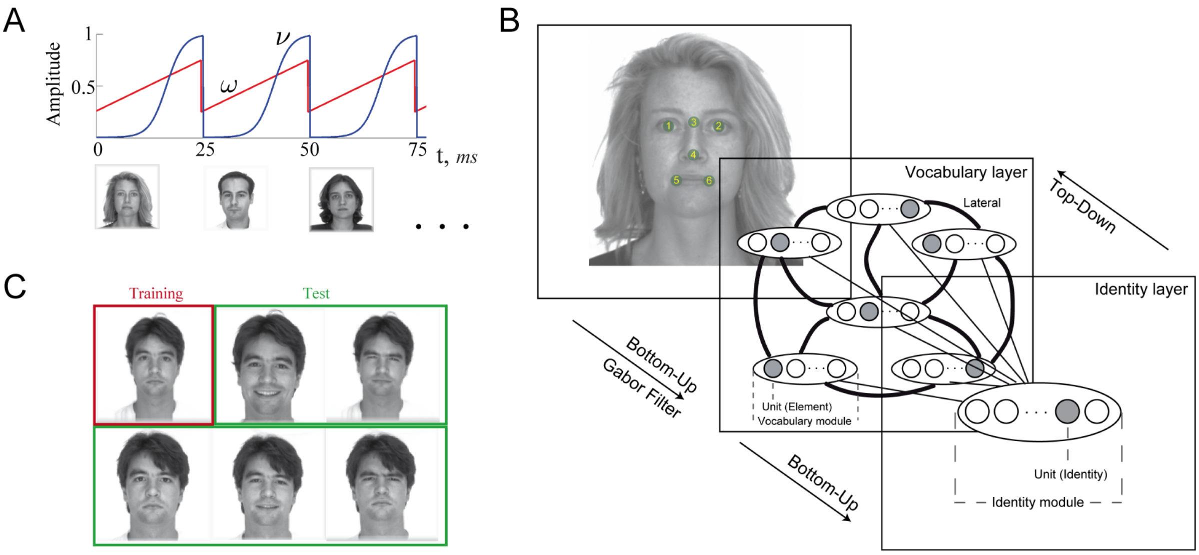


Figure 3.2: Unsupervised on-line learning procedure used to form the network structure of the memory domain. **(A)** During the on-line learning, the face images are presented incrementally showing one image per decision cycle. Decision cycles are defined by excitatory and inhibitory rhythms ω and ν which are taken to oscillate in the gamma range ($T = 25ms$). **(B)** Each face image is fed to the lower layer as a collection of $L = 6$ Gabor filter banks extracted locally from the respective landmarks. This sensory input is the only information the network gets, no identity or other labels are provided. The activity generated during decision cycle guides structure formation in the network. **(C)** Different face views used as input (one person out of total 40 shown). Top left is the original view with neutral expression taken for training. Other views were used for testing (the bottom row shows duplicate views taken two weeks after the original series).

the test for generalization error, all the synapses are frozen to exclude the possibility that recognition improves during the testing phase. The history of network unit responses to faces presented during the learning phase is used again in the same manner for the generalization error rate computation as it was done for the evaluation of the learning error.

Two different kinds of generalization error are motivated by the necessity to take into account the effect of the ongoing homeostatic activity regulation on the processing in the network. Because of this adaptive mechanism, network responses to incoming stimuli depend on the history of recent stimulation. In other words, at any time the network activity state is not only a function of the current input but is also biased by the recent activity states. To avoid the effect of this bias on performance evaluation, the block generalization error is introduced. In order to compute it, the network with frozen synapses and active intrinsic plasticity is tested over many repetitive blocks of novel images. Each block contains a particular alternative view of all 40 persons, presented in random order. The rate of wrong responses is then computed over all blocks, minimizing the influence of bias on the result.

The immediate generalization error computation is performed by disabling all adaptive mechanisms - the synaptic and the intrinsic plasticity. This setup provides an estimation of the system's recognition performance if absolutely no opportunity for of synaptic or non-synaptic adaptation to novel data is possible. This kind of evaluation underestimates the true performance, as the computed error is heavily biased by the stimulation the network experienced during the learning immediately before the testing procedure. Still, it delivers a lower bound for the recognition performance, providing a hint how the network is able to perform in situation, where preceding stimulation interferes strongly with the current network response.

The generalization errors of both kinds can be evaluated for each alternative view type separately.

This is useful in order to be able to discover potential view-specific differences, especially in terms of comparing the fully recurrent and purely feed-forward network configurations. All error types are evaluated separately for vocabulary and identity network layers.

3.1.2 Assessing network connectivity organization

To analyze the progress of network structure formation, I use measures describing different properties of the synaptic connectivity. The distance measure calculates the distance between two synaptic weight vectors $\mathbf{w}_i$ and $\mathbf{w}_j$ [Blais et al., 1998]:

$$\begin{aligned} d(\mathbf{w}_i, \mathbf{w}_j) &= \frac{1}{4}\Big(\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2} - \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|_2}\Big)^2 \\ &= \frac{1}{2}(1 - \cos\phi), \end{aligned} \tag{3.1.1}$$

where ϕ denotes the angle between the two synaptic weight vectors each comprising the receptive field of a unit. The value lies in the interval between zero and one. If the weight vectors are the same, the distance value is zero; if the vectors are orthogonal, the value is 0.5; if the vectors are directed opposite to each other ($\phi = \pi$), the value is one. Utilizing this basic distance measure, I further construct a differentiation measure, which is supposed to reflect the grade of differentiation between the receptive fields of the same type across the units in the whole network. The differentiation grade $\mathcal{D}_k^{Source}$ is computed for the receptive field from a given $Source \in \{BU, LAT, TD\}$ for each module k, $k = 1 \ldots M$, and then an average differentiation value $\mathcal{D}^{Source}$ is built from the values of all M modules:

$$\begin{aligned} \mathcal{D}_k^{Source} =& \frac{1}{N(N-1)} \sum_i^N \sum_{j \neq i}^N d(\mathbf{w}_i^{Source}, \mathbf{w}_j^{Source}) \\ \mathcal{D}^{Source} =& \frac{1}{M} \sum_k^M D_k^{Source}, \end{aligned} \tag{3.1.2}$$

where N is the number of units in the module. The differentiation grade measure is evaluated separately for vocabulary modules on the lower memory layer and for the identity module on the higher memory layer.

Further I employ a measure ζ of sparseness of the inner structure of a receptive field, which is high if there are only few strong synapses and many weak synapses within the receptive field. To assess the same property not only within, but also across receptive fields, an overlap measure ξ is utilized. If the receptive fields of different units have many synapses that have similar weights, the value will be high; if there are only few such synapses, the value will be low. The overlap measure is thus closely related to the differentiation grade between the receptive fields as assessed using the distance measure. Both measures ζ and ξ have the same scheme behind their computation, with the only difference that the former is computed within while the latter across the receptive field vectors using a common sparseness measure $\mathcal{A}^{Source}(s)$ as defined in [Rolls and Tovee, 1995]. Again, the computation is done for each module on receptive fields of the same type $Source \in \{BU, LAT, TD\}$, building then type-specific

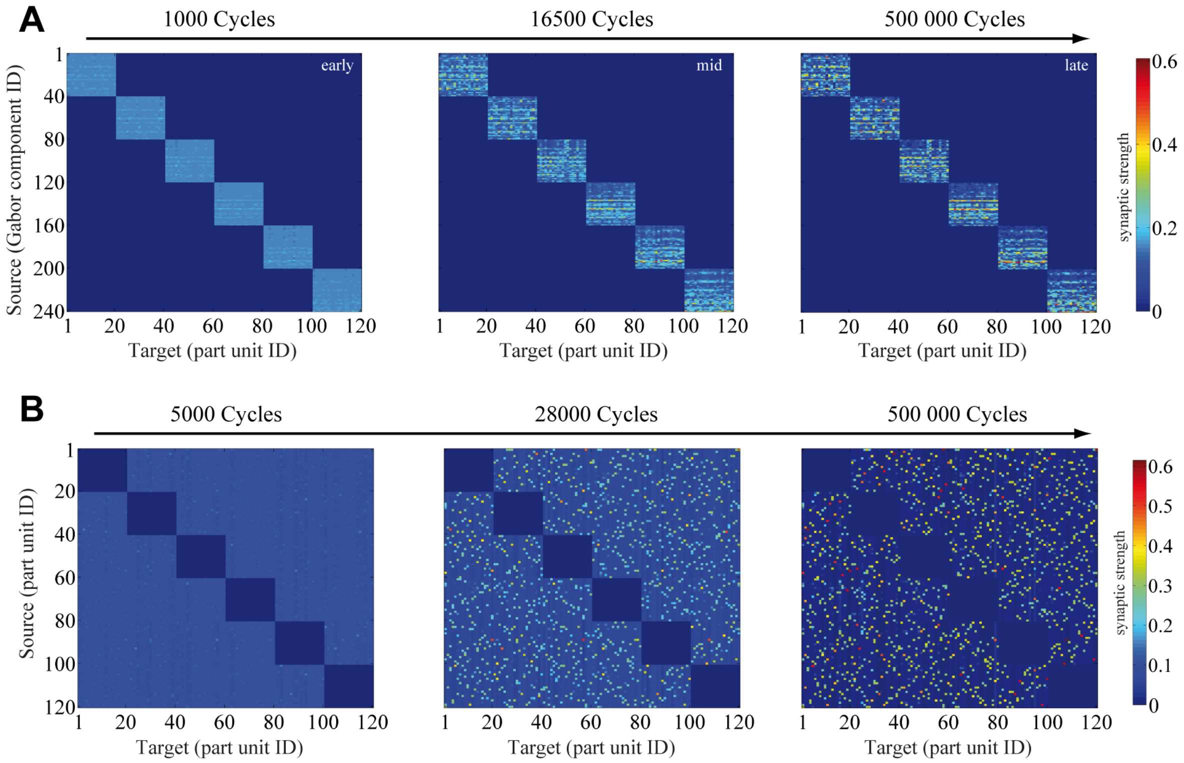


Figure 3.3: Time snapshots of structure formation during the unsupervised learning. From left to right, snapshots from early, middle and late formation phases of **(A)** lower layer bottom-up connectivity, **(B)** lower layer associative lateral connectivity.

average values $\mathcal{C}^{Source}$ and $\mathcal{E}^{Source}$ over all M modules:

$$\mathcal{A}^{Source}(s) = \left(\tfrac{1}{s}\sum_i^s w_i^{Source}\right)^2 \Big/ \left(\tfrac{1}{s}\sum_i^s (w_i^{Source})^2\right)$$

$$\zeta_k^{Source} = \frac{1}{N}\sum_i^N \mathcal{A}_i^{Source}(r), \qquad \xi_k^{Source} = \frac{1}{r}\sum_i^r (1 - \mathcal{A}_i^{Source}(N)) \tag{3.1.3}$$

$$\mathcal{C}^{Source} = \frac{1}{M}\sum_k^M \zeta_k^{Source}, \qquad \mathcal{E}^{Source} = \frac{1}{M}\sum_k^M \xi_k^{Source},$$

where N is the number of units in a module, M the total number of assessed modules and s, r the total number of synapses in a receptive field of type $Source \in \{BU, LAT, TD\}$. The evaluation is done separately for the vocabulary modules and the identity module.

3.1.3 Results

Structure formation. Facing the task of unsupervised learning, the network develops a structural basis for storing the faces of individual persons shown during the learning phase. The distributed vocabularies for local appearance are created on the lower memory layer to represent facial parts. These vocabularies are formed by the bottom-up synaptic connections of the lower layer modules attached to the respective facial landmarks (Fig. 3.3 **(A)**). Each unit of the vocabulary modules becomes sensitive to a particular local facial appearance due to the established structure of its bottom-up receptive

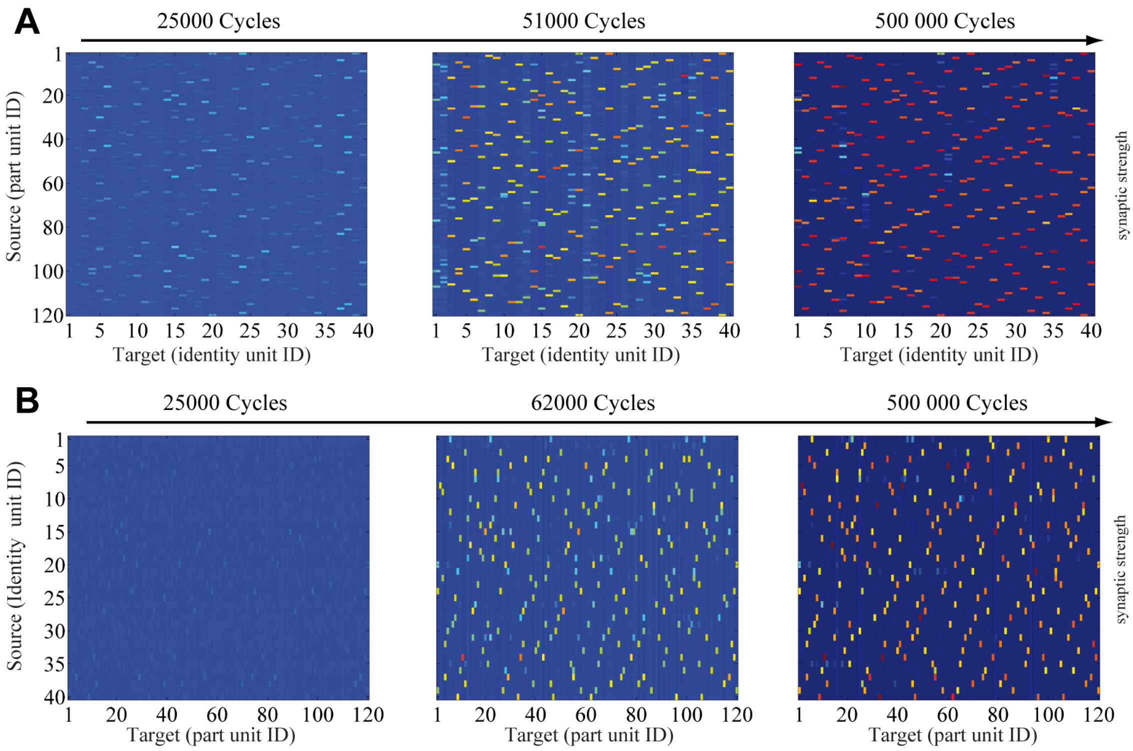


Figure 3.4: Snapshots of structure formation during the unsupervised learning. From left to right, snapshots from early, middle and late formation phases of **(A)** the higher layer bottom-up connectivity holding face identities composed from the parts on the vocabulary layer. **(B)** Top-down connectivity, which is projecting this compositional information back to the vocabulary layer. Consequently, it is roughly the transposed version of the bottom-up connectivity.

field. At the same time, the lateral connectivity between the vocabulary modules gets shaped capturing the associative relations between individual facial parts (Fig. 3.3 **(B)**). These relations are represented by associative links between the units which are regularly part of an individual face presented during learning. The same compositional information enters into the structure of bottom-up connectivity converging on the identity units (Fig. 3.4 **(A)**), being also represented in the top-down connections projecting from the identity module back on the lower layer (Fig. 3.4 (B)).

Each person repeatedly presented to the system during the learning phase leaves a memory trace in the network structure. Memory traces comprise different types of synaptic connectivity, linking a subset of units together into a hierarchical assembly that forms an explicit compositional representation of person's face. (Fig. 3.3, 3.4). On the lower layer, the lateral connectivity links the parts in associative fashion, while on the higher layer the bottom-up and top-down connectivity capture the compositional identity of the face in explicit, generative fashion. The course of gradual differentiation of bottom-up, lateral and top-down connectivity reveals the ongoing process of memory consolidation, where memory traces induced by the face images become more stable and get the opportunity to amplify their structure. A common developmental pattern seems to underlie the time course of structure organization (as defined in Sec. 3.1.2). There is an initial resting phase, where no structural changes appear, followed by a maturation phase, where massive reorganization occurs and the rate of change goes through a maximum (Fig. 3.5). Finally a saturation phase is reached, where the structure stabilizes at a certain level of organization and the rate of change goes down close to zero.

Different connectivity types get organized preferentially within a specific time window (Fig. 3.5).

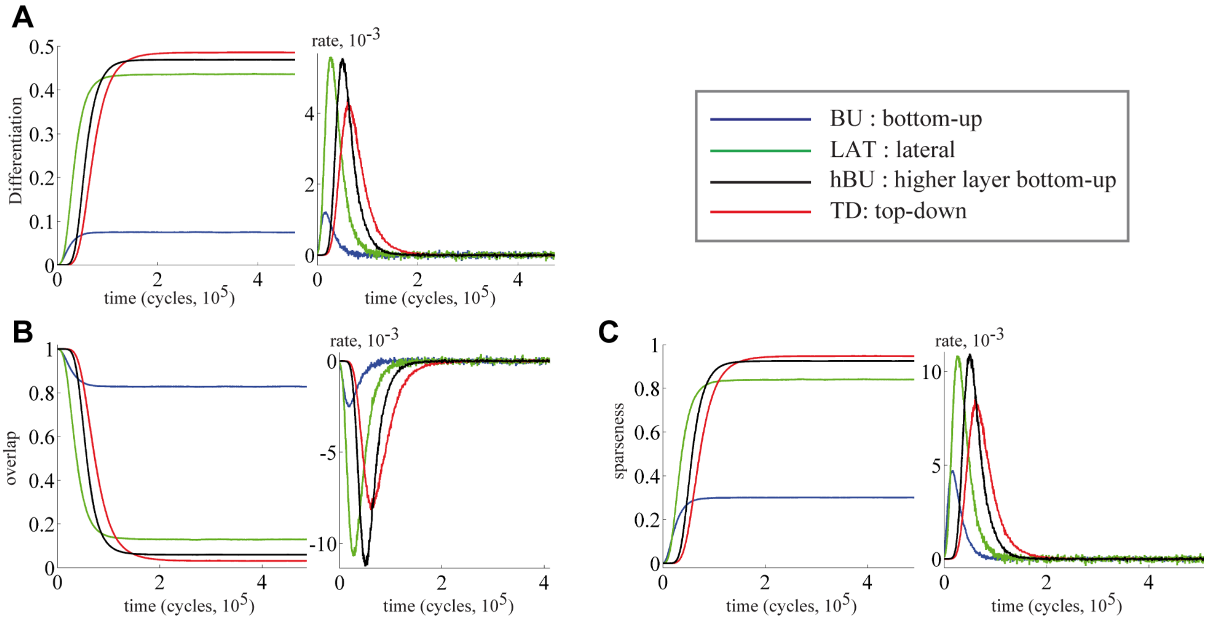


Figure 3.5: Time course of network structure organization. **(A)** Time course of differentiation $\mathcal{D}$ over $5 \cdot 10^5$ decision cycles for different connectivity types; on the left the grade of differentiation, on the right its rate of change. Clear is the general tendency to greater connectivity differentiation with the learning progress as well as the temporal sequence of connectivity maturation. Time course of overlap $\mathcal{E}$ **(B)** and sparseness $\mathcal{C}$ **(C)** over $5 \cdot 10^5$ decision cycles for different connectivity types. As the learning progresses, the overlap between the receptive fields is continuously reduced, the connectivity sparseness increases. Again, the temporal sequence of connectivity development is clearly visible (see the text). BU, LAT, hBU, TD denote respectively lower layer bottom-up, lateral, higher layer bottom-up and top-down connectivity types.

There is a clear temporal sequence of connectivity development, starting with maturation of lower layer bottom up connections, followed by maturation of lateral connections between the vocabulary modules and by the maturation of bottom-up connectivity to the identity module, ending with the formation of top-down connectivity. Because the development of different connectivity types is highly interdependent, their developmental phases overlap substantially. In parallel, there is a gradual increase in sparseness within the receptive fields and progressive reduction of the overlap between them. (Fig. 3.5 **(B)**, **(C)**) The remaining overlap in associative lateral and compositional bottom-up connectivity reflects the extent to which the parts are shared among different faces stored in the memory network.

In the late learning phase, the state of the synaptic structure stabilizes until no substantial changes in the established network structure can be observed (Fig. 3.5). The bottom-up connectivity of the vocabulary modules stays well behind other connectivity types in terms of differentiation grade, of sparseness within the receptive fields and of their overlap reduction achieved in the stable connectivity state (Fig. 3.5). While being the latest to initiate its maturation, the top-down connectivity reaches the highest grades of differentiation and sparseness, also being most successful in reducing the overlap. The lateral connectivity between the vocabulary modules and the bottom-up connectivity of the identity module also show comparably high level of organization. These relationships reflect the distinct functional roles the different connectivity types play in their contribution to the memory traces. The lower layer bottom-up connectivity holds strongly similar prototypes of local facial appearance, while lateral and top-down connectivity stores well-separable information about compositional identity of different faces.

The changes in the synaptic structure are accompanied by the use-dependent regulation of the ex-

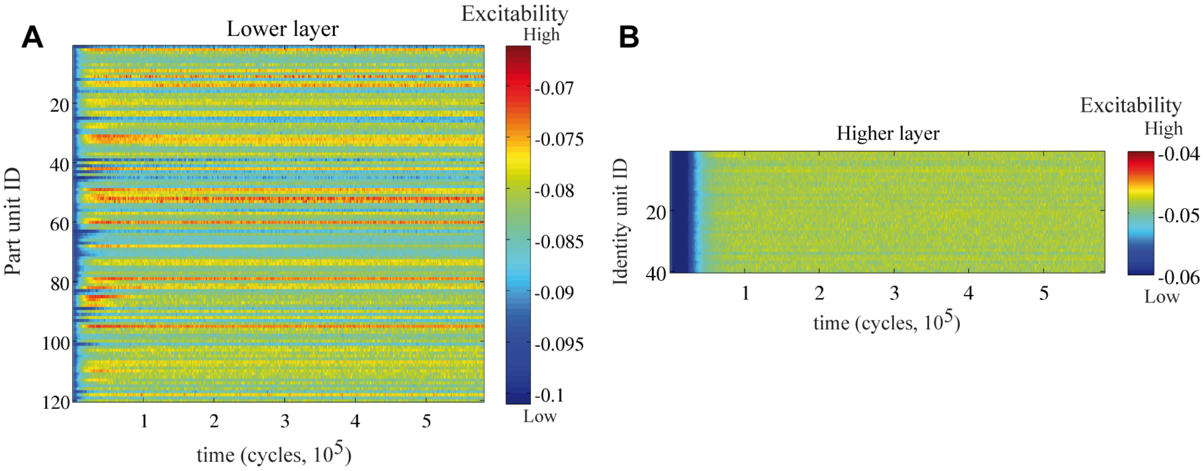


Figure 3.6: Time course of excitability regulation. **(A)** the lower, **(B)** the higher memory layer. Obvious are the much stronger pronounced differences in excitability between the units on the vocabulary layer.

citabilities of the network units. Three developmental phases can be distinguished in the time course of excitability modifications (Fig. 3.6). The first phase is characterized by strong and rapid excitability downregulation in the network. This downregulation settles down the unit activity toward the range of the targeted average activity level p_{aim} (Eq. 2.2.1). In this phase, almost no differences between the individual excitability levels are present (Fig. 3.7). After downregulation crosses its peak, a common upregulation sets in and the differences between the excitability levels become much more pronounced. The upregulation phase leads to a slight increase of the average excitability in the network and is followed by a saturation phase where the excitability stabilizes around a certain level.

Excitability regulation runs differently on different memory layers. On the lower layer the down- and upregulation phases are shorter and occur earlier than the corresponding phases on the higher layer. Moreover, the differences in excitability between the units on the lower layer are much stronger pronounced compared to the rather equalized excitability levels of the higher layer units (Fig. 3.6 and 3.7).

These differences reflect the distinct functional roles the lower and higher layer play in the memory organization. The lower layer serves as a storehouse for associatively linked distributed facial parts that can be shared by multiple faces, while the identity units are conjunction-sensitive units representing the compositional identity of an individual face. Because each memorized person is equally likely to appear on the input, the long-term usage load of the identity units is essentially the same, so no need for a systematic differentiation of excitability levels arises there. Part sharing on the other hand imposes different usage frequency on vocabulary units sensitive to different parts, leading to pronounced use-dependent differences in excitability between them.

The quality of established memory traces is determined by the selectivity the units develop for particular persons shown during the learning phase. The unit selectivity also determines the usage load of a unit. Analogously to the procedure described in Sec. 2.5.1, unit selectivities for person and gender can be computed and visualized, doing it separately for units on the vocabulary and identity layers (Fig. 3.8, 3.9). On the vocabulary layer, the majority of units are shared by two persons, representing the balanced unit usage load that is also reflected in the roughly uniform unit win probability (Fig. 3.8 **(B)**, **(C)**). There are also a few under-utilized units highly selective for only one person, and a few over-utilized units being shared by more than two persons (up to four). Many units are highly selective for the specific gender, preferring only male or only female faces. On the identity layer, the units show

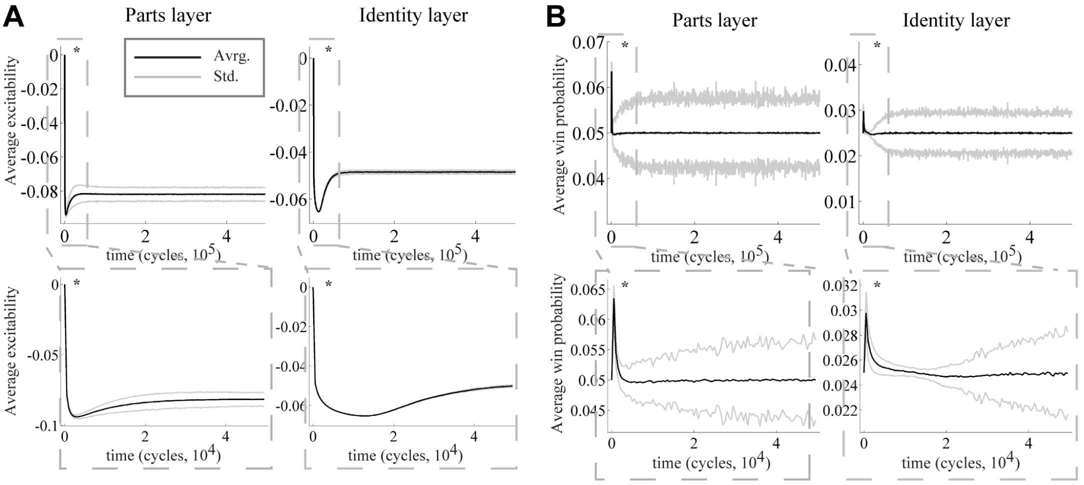


Figure 3.7: **(A)** Time course of average excitability regulation. Above the whole course, below the zoom into down- and upregulation phases. On the left for the part units, on the right for the identity units. Black solid curve is the average value, gray curves mark the standard deviation range. The same nomenclature applies for the time course of the average unit win probability visualized in **(B)**. As visible in **(A)**, the differences in excitability between the units are more pronounced on the parts layer compared to the identity layer. This is reflected again in the greater dispersion of the unit win probabilities around the average on the parts layer, as shown in **(B)**.

very high selectivity for person identity ($P(winner|person) > 0.9$), indicating that an identity unit prefers only one particular person from the training set (Fig. 3.9). As identity units develop such a high degree of identity preference, they also acquire trivially high selectivity for the respective gender.

To further substantiate the evidence for the memory trace quality, I checked whether the compositional representation of face identities captured in the network structure is consistent across different connectivity types. I first analyze whether the part units connected to an identity unit are selective for the same person identity as the identity unit they converge on. The procedure delivers a consistency score between zero and one for bottom-up connectivity of each identity unit. Zero corresponds to a situation where none of the connected units match the selectivity of the identity unit. The maximal score of one corresponds to the highest possible consistency, where all connected part units conform to the selectivity of the identity unit. The consistency score plots taken at early and mature states of connectivity structure show that the consistency grows to the highest level of one in the mature state (Fig. 3.10 **(A)**).

Second, a similar kind of procedure is performed to check the connectivity consistency between bottom-up and lateral synaptic structure. The procedure tests whether the part units converging on an identity unit are consistently interconnected via lateral synapses. The consistency score is again between zero and one. Zero stands for no connectivity between any of the part units, one stands for maximal consistency where all part units converging on the identity unit are indeed laterally interconnected. In the mature connectivity state, the consistency score is close to one for most of the identity units (Fig. 3.10 **(B)**). As the top-down connectivity matrix can be shown to be roughly the transposed version of the bottom-up connectivity matrix (Fig. (Fig. 3.10 **(C)**)), the same analysis is valid for the established top-down structure.

Taken together, the evaluation of structure differentiation, sparseness, overlap and consistency provide strong evidence for the quality of the memory traces formed during the learning. The individual faces are stored in a memory domain with a structure which appropriately captures their compositional identity in hierarchical, generative manner. The contribution of the network units to the memory traces

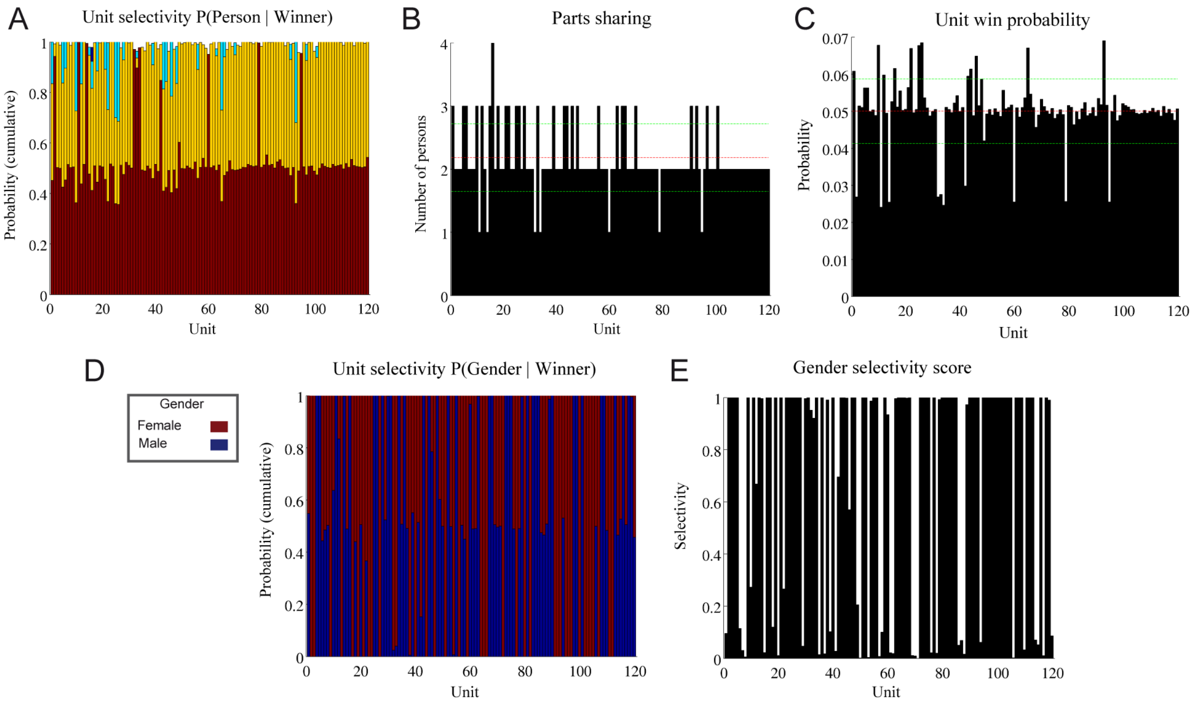


Figure 3.8: Unit selectivity on the vocabulary layer of the memory network ($M = 6$ vocabulary modules ($N = 20$ units each), $P = 40$ persons in the training set). The plots show the state after $5 \cdot 10^5$ decision cycles. **(A)** Selectivity for different persons. Each colored bar corresponds to a conditional probability $P(person|winner)$ for a respective unit. **(B)** Number of persons sharing a unit, computed from unit selectivity. **(C)** Unit win probability computed over a time interval of $2 \cdot 10^4$ cycles. The balanced unit usage load is reflected in roughly uniform win probability, with deviations corresponding to more or less utilized units. **(D)** Gender selectivity. Each colored bar corresponds to a conditional probability $P(gender|winner)$ for a respective unit. **(E)** Gender selectivity score. Some units develop very high gender selectivity (score 1), while others only poorly differentiate between male and female faces (score close to 0).

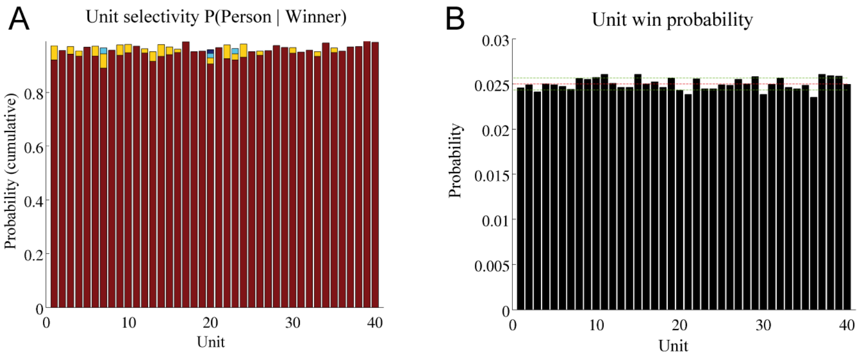


Figure 3.9: Unit selectivity on the identity layer of the memory network (one identity module ($N = 40$ units), $P = 40$ persons in the training set). The plots show the state after $5 \cdot 10^5$ decision cycles. **(A)** Selectivity for different persons. Each unit develops very high selectivity for a particular person from the training set. **(B)** Unit win probability computed over time interval of $2 \cdot 10^4$ cycles. The distribution is close to uniform, reflecting the balanced usage load which is due to the preferences the identity units have developed for memorized face identities.

is well balanced, none of units are strongly over- or under-utilized.

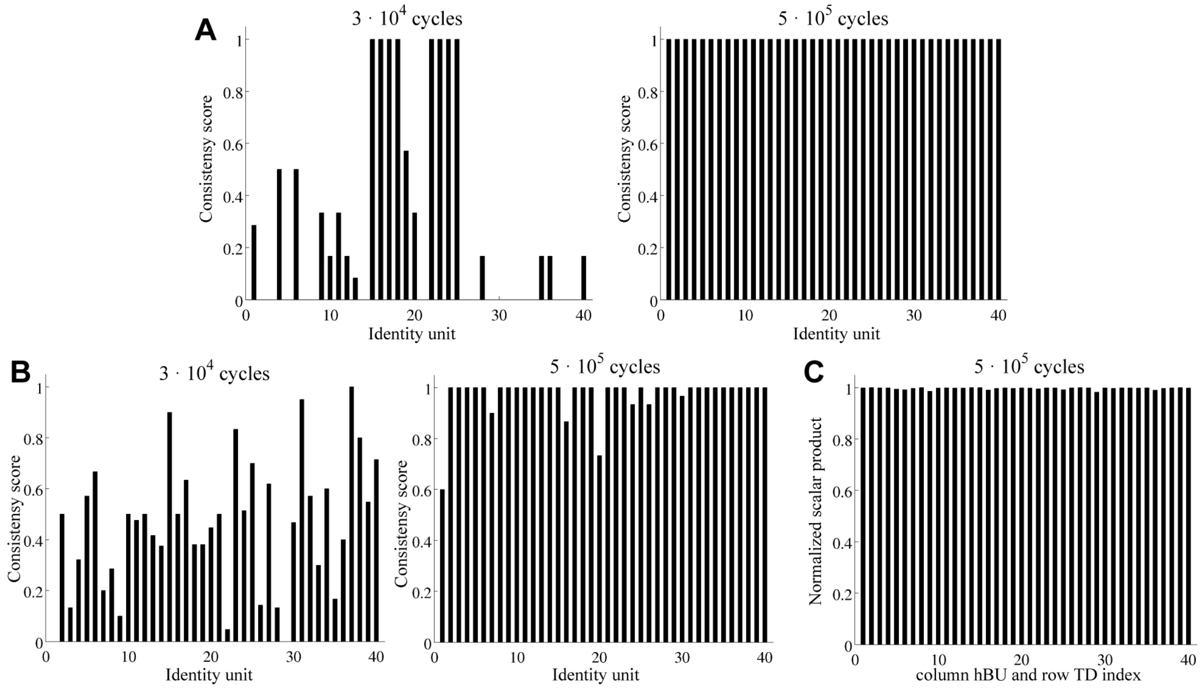


Figure 3.10: Connectivity consistency scores in immature and mature network connectivity states. **(A)** The score is the degree of match between the selectivity of identity units and the selectivity of part units converging on them. This match is perfect in the mature state, showing that part and identity units form consistent assemblies holding the memorized faces. **(B)** The score checks for explicit lateral connections between the part units converging on the same identity unit. Again, in the mature state the consistency score is close to one for most identity units. **(C)** The top-down connectivity matrix is a transposed version of the bottom-up connectivity matrix of the identity units, as revealed by computing the normalized scalar product between the rows and columns of the respective matrices.

Activity formation and coordination. The established synaptic structure supports the compositional representation scheme in which the relations between parts are encoded in two alternative ways. First, the relations are explicitly signaled by the responses of the identity units on the higher layer. These units are sensitive to particular combinations of part units. Each identity unit is therefore a symbolic label for one of the compositional face identities stored in the memory. Second, the relations are represented by assemblies of co-activated part-specific vocabulary units. These assemblies can be constructed on demand to encode a novel face or to recall an already stored one as a composition of its constituent parts. The selection of the part-specific and identity-specific units into a coherent assembly that encodes an individual face is done in the course of the decision cycle spanned by the ongoing gamma rhythm ($T = 25ms$, Fig. 3.11, 3.12).

Within the gamma cycle, the global decision process, which may be interpreted as selection and amplification by competition and cooperation, is responsible for assembly formation. This decision process leads to clear and unambiguous temporal correlations between the selected units, emphasizing those against the rest by amplification of their activity and suppression of the activity of the others. (Fig. 3.11). The initial phase of the decision cycle, where oscillatory inhibition and excitation are low, is characterized by low undifferentiated activities of the network units. With both inhibition and excitation rising, only some of the units are able to resist the inhibition pressure and continue increasing their activity, being selected as candidates for assembly formation in the selection phase.

This selection is determined not only by local competition within the modules, but also by competi-

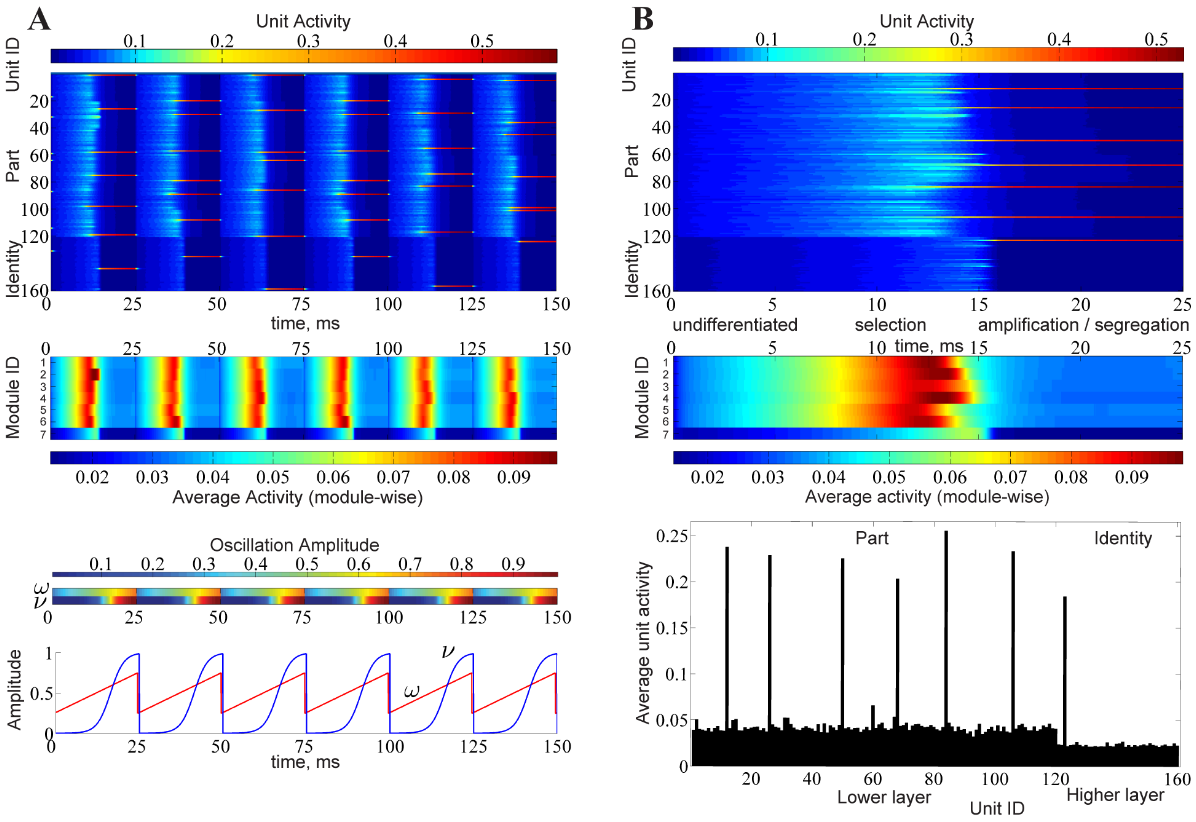


Figure 3.11: Activity formation during memory recall. **(A)** A sequence of six successive cycles, each representing a successful recall of a stored individual face. On the top, the activity course is shown. In each cycle, the winner assembly represents the current input face as composition of its parts and its higher-order identity label. The second and forth cycles show recall of the same face identity. Below is the mean activity course for each module and the oscillation rhythms defining the decision cycle. **(B)** A zoom into a single decision cycle (on the top) to visualize the phases of activity formation, which starts from rather undifferentiated state and progresses toward the highly organized state with an amplified sparse winner assembly at the end. Below the course of average activity within each module and distribution of average unit activities over the decision cycle are shown. The highly competitive nature of activity formation is visible, where winner units get amplified at the cost of suppressing the others.

tion and cooperation between the units across the modules, mediated by the signaling over lateral and top-down connections formed during the learning (Fig. 3.12 **(B)**). Ultimately, growing inhibition leads to a series of local winner-take-all decisions across the modules. The candidates that were not able to get sufficient contextual support from the other units loose the competition. Activity in the network is further sparsened by strong amplification of a small winner unit subset that was most successful in cooperating at the cost of strong suppression of the less successful rest. In the late phase of a decision cycle, this amplified winner units assembly can be then clearly interpreted as an individual face composed of the local features from respective landmarks and labeled with the person's identity as a high-level symbol, solving the assembly binding problem [von der Malsburg, 1999, Singer, 1999].

The competitive nature of activity formation in the network is clearly revealed by comparing the average activity within the modules and activities of the single units. (Fig. 3.11 **(B)**). While the winner unit assembly concentrates increasingly high activity, the average network activity gets progressively reduced after crossing its peak in the selection phase towards the end of the decision cycle, indicating that the amplification of the winner assembly occurs at the cost of suppressing the rest. During the

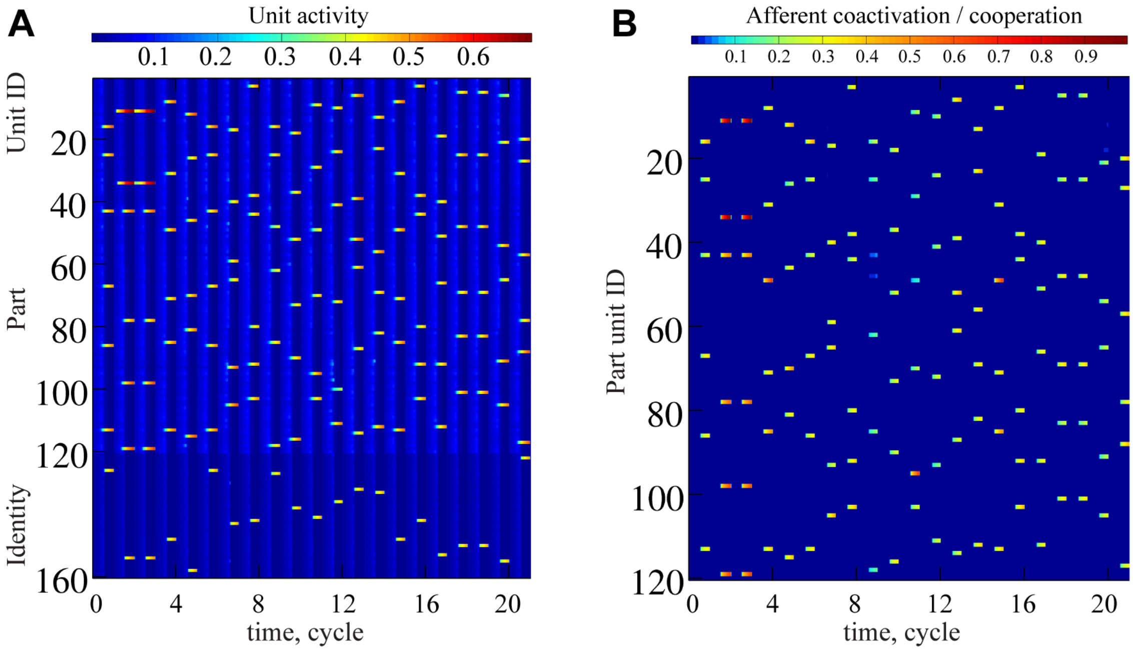


Figure 3.12: Activity formation and signal cooperation. 20 successive decision cycles are shown. **(A)** In each cycle, a presented face induces memory recall in form of the winner assembly. **(B)** The high degree of coactivation of afferent bottom-up, lateral and top-down signals converging on winner part units indicates strong cooperation within the assembly. The activated winner assemblies representing memorized faces are thus highly coherent in terms of their internal synaptic coupling, and the local decisions are characterized by strong agreement of sensory and contextual cues.

whole decision cycle the average network activity level stays low ($p = 0.08 - 0.1$), far below the activity level reached by the winner units at the end of the cycle ($p = 0.4 - 0.6$).

Earlier in this section, I discussed consistency of structure formation in terms of appropriately representing the compositional identity of stored faces in synaptic memory traces. One may also ask a related question about network activity formation, namely to what extent it becomes more organized, or coherent, with learning progress in terms of representing the memorized faces by assemblies of active units. In other words, we are interested in the level of coherence, or agreement, between the local decisions made in the distributed modules and how this level changes on course of learning. One indicator of such coherent decision making is the agreement achieved at the end of the gamma cycle between the afferent signals that arrive at network units from different sources (bottom-up, lateral or top-down). The measure of this agreement can be obtained by first computing the standard correlation coefficient ρ [DeGroot and Schervish, 2001] for each pair of different afferent signals (BU-LAT, BU-TD, LAT-TD) arriving at the part units over a time period of network stimulation, and then making a pair-specific average over all coefficients for the lower layer. Extending this evaluation to different time points during the learning provides a plot showing the development of signal coherency, or signal coordination, over the learning course (Fig. 3.13).

The coordination level between the bottom-up, lateral and top-down signals increases gradually from the initially very low value close to zero toward higher and higher grade (Fig. 3.13). The low coherence value in the early learning phase reveals inability of the signals converging on network units to be in consensus with each other about the local decision outcome. This points to a deranged decision making in the early learning phase, where contextual signaling is deteriorated because of immature lateral and top-down connectivity.

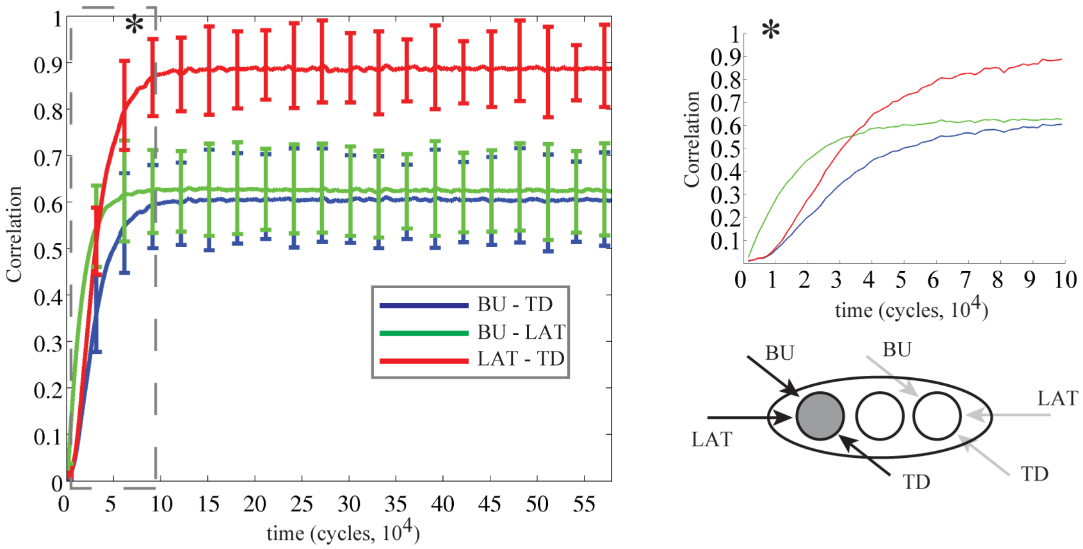


Figure 3.13: Improvement of afferent signal coordination in the course of learning. Standard correlation coefficients ρ were computed for each signal pair. BU, LAT, TD denote respectively bottom-up, lateral and top-down signals.

As learning progresses, the signal pathway structure is gradually improved for the storage and representation of the experienced faces, leading to stronger and stronger consistency in local signaling. The bottom-up and lateral signals are the first to develop a significant grade of coherence. Slightly later, the lateral and top-down signals reach a substantial coherence level. The latest to establish a coordinated cross-talk are the signals from bottom-up and top-down sources. The lateral and top-down signaling manages to establish the strongest final grade of coherence. This level is significantly higher than the coherence between bottom-up and lateral as well as bottom-up and top-down signals. Their coherence still reaches substantial values though, the former being slightly above the latter. The high degree of agreement between the signals exchanged over lateral and top-down pathways provides strong evidence for the successful communication and cooperation between the units representing the different components and the identity label of the same face stored in the memory network.

During the course of a single decision cycle, a related co-activation measure can be used to check whether the incoming signals are coordinated properly to make up the decisions at cycle's end. The relationship between the afferent signal coordination and the memory function becomes particularly clear if the coordination level in a successful recall is compared to the coordination level shown during a failed recall, where the recalled person identity label does not match its compositional parts configuration (Fig. 3.14). In a successful recall, where this match is correctly accomplished, a well-established coordination can be observed between the co-active afferent signals converging on the winner units. In a failed recall, the identity module making a wrong decision sends top-down signals that are not in agreement with the bottom-up and lateral signals conveyed by the vocabulary modules. Consequently, the signal coordination breaks down. This coordination breakdown caused by disagreement between local decisions can thus serve as a reliable indicator of a recall failure (Fig. 3.14 (**B**)). The disagreement between the sensory and contextual signaling can also be interpreted as an error signal, indicating a deviation between the bottom-up signal and the top-down prediction. Although this signal is not represented here explicitly by activity of a dedicated unit, in further perspective it could be of great use for determining the state of the recognition process and for guiding local learning in form of an explicit reward signal.

Another indicator that can help in differentiating a successful from a partially or completely failed

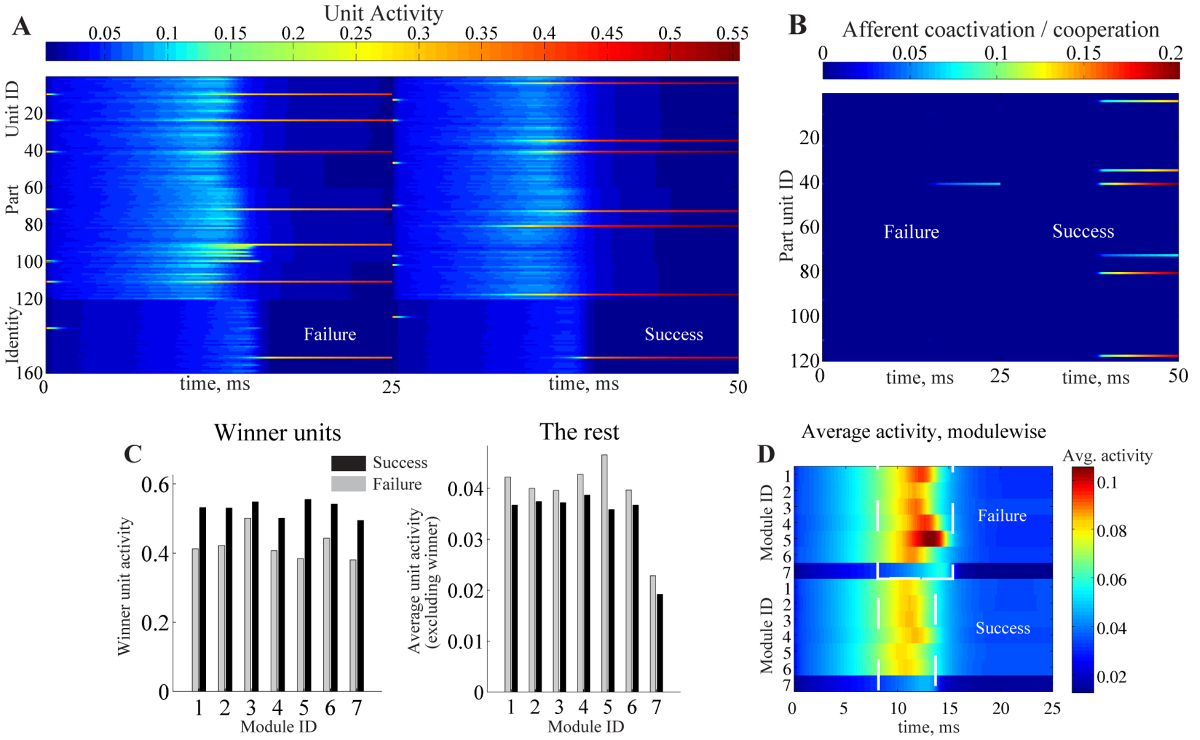


Figure 3.14: Signatures of successful and failed recall. Two decision cycles showing failed and successful recall. **(A)** Activity formation. **(B)** Afferent signal coordination assessed by measuring the co-activation of bottom-up, lateral and top-down signals converging on the network units. In the failed recall, there is a clear break-down of signal coordination in afferents converging on the winner units. **(C)** Winner unit activities at the end of the decision cycle on the left and average unit activities (excluding the winners) over the whole cycle on the right for each module. In the failed recall, the activity of winner units is lower, while the average activity of the rest is higher than in the successful recall. **(D)** Course of average activity in the modules. In the failed recall, a substantial increase in overall activation is clearly seen as well as the shift of its broader peak to a later time point.

recall is simply activity level of the winner units at the end of the decision cycle. A successful recall is accompanied by a high degree of cooperation between the participating winner units, so that the level of their final activation is high. At the same time, the competitive action of the winner units assembly suppresses strongly activity of the rest, so that the overall network activity is substantially diminished. Reversely, a failed recall has something to do with disagreement between some local decisions, resulting in decreased afferent signal coherence, which in turn leads to a much lower level of final winner units activity. The competitive influence of winner assembly is also weakened in this case, leading to a higher overall network activity (Fig. 3.14 **(C)**). Thus, a simple comparison of the winner activities to their win activity level achieved on average can already provide enough information to conclude about the quality of recall. The recall quality can be assessed on the global level of the identity as well as on the component level, where either identity recognition failure or part assignment failure might be stated.

Recognition performance. To assess the recognition capability of the memory network, the learning and block generalization error of two alternative network configurations are evaluated (see Sec. 3.1.1). These different configurations, the fully recurrent and purely feed-forward one, were set up to substantiate the hypothesis about the functional advantage of the context-sensitive recurrent memory network structure over the structure with purely feed-forward connectivity. Both configurations were

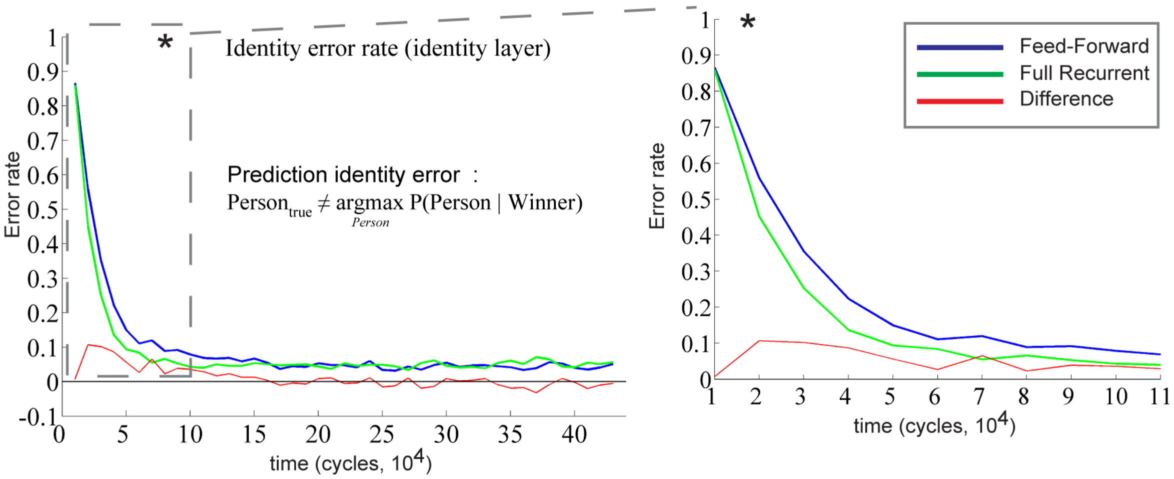


Figure 3.15: Learning error for the identity layer, comparing feed-forward and recurrent network configurations. The identity error rate drops below 5% for both configurations. As seen on the zoom in the early learning phase (right plot), the recurrent network configuration learns faster, reaching the low error level well before the feed-forward configuration is able to do so.

trained under equal conditions and then tested to compare their performance against each other (refer to Sec. 3.1.1).

Both the purely feed-forward and fully recurrent configurations are able to successfully store the face identities of the persons ($P = 40$ in total) in the memory. Both network configurations are also able to classify person gender. Strong decay of the learning error over the time is clearly evident for both network configurations. On the identity layer, the error rate for person identity falls rapidly in the early learning phase (first $5 \cdot 10^4$ decision cycles) until it saturates at the values slightly below 5% in the later phase beyond 10^5 cycles (Fig. 3.15).

There is no significant difference in the error rate on the training set between the both configurations after the saturation level is reached. The time needed to reach the saturation level is significantly shorter for the fully recurrent configuration (saturates around 10^5 cycles) than for the purely feed-forward one (saturates around $1.5 \cdot 10^5$ cycles, Fig. 3.15). Thus, the learning of the training data set progresses about 33% faster for the fully recurrent system than for the purely feed-forward one. The fully recurrent configuration seems to speed up the learning progress in the critical early learning phase. This is most probably due to additional support provided by lateral and top-down connectivity for the organization, amplification and stabilization of the memory traces.

The learning error for person identity and gender can be also evaluated on the parts layer in analogy to Sec. 2.5.1. If all 6 vocabulary modules are involved in person recognition, the identity and gender error rates drop almost to zero already in early learning phase (Fig. 3.16). Again, on the training data set no significant difference can be stated for the two network configurations.

At first glance, analysis of the learning error time course suggests that the only functional advantage of the fully recurrent configuration is the learning speed-up observed in the early phase. However, the learning error is evaluated on the training data set only. It does not tell us anything about the capability of the memory network to generalize over data not shown during the learning. An important functional advantage of the fully recurrent network configuration is revealed if the generalization error rates are compared. The block generalization error is measured on the blocks of alternative face views not shown during the learning phase (see Tab. 3.1, 3.2, 3.3). The identity error can be assessed separately for the parts layer and for the identity layer. For parts layer, the gender error is evaluated in addition to the

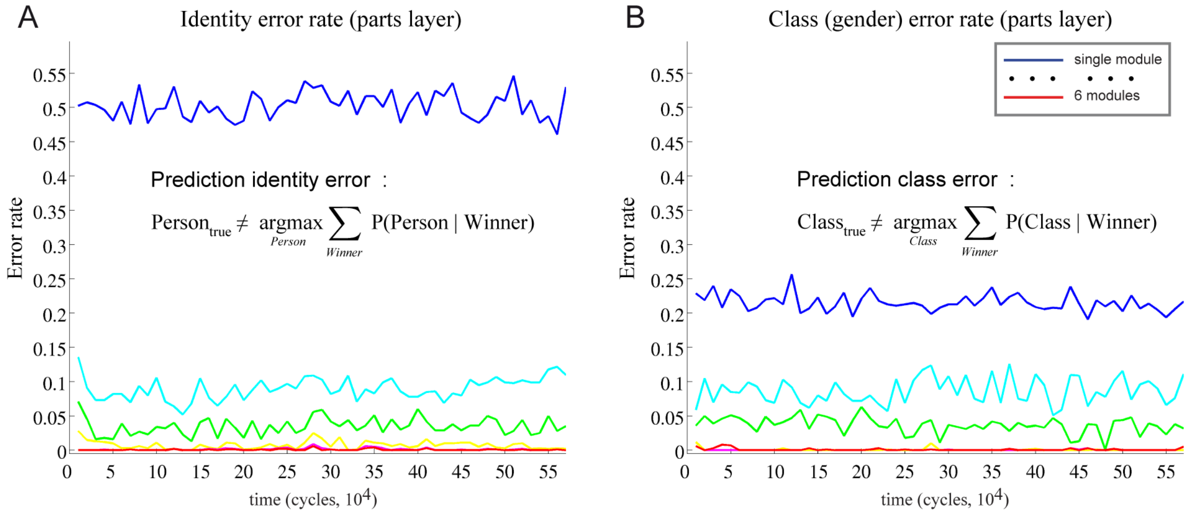


Figure 3.16: Learning error for the parts layer ($P = 40$ persons in the training set). **(A)** Person identity error. **(B)** Gender error. The more modules are involved in identity or gender estimation, the less is the error rate. For 6 modules, both the identity error and gender error drop close to zero.

Configuration	Views, identity error rate (identity layer)		
	Original	Smiling	Sad
Fully recurrent	$0.1\% \pm 0.07\%$	$6.06\% \pm 0.58\%$	$4.02\% \pm 0.42\%$
Purely feed-forward	$0.067\% \pm 0.0528\%$	$5.72\% \pm 0.92\%$	$3.75\% \pm 0.38\%$

Configuration	Views, identity error rate (identity layer)		
	Duplicate	Duplicate, smiling	Duplicate, sad
Fully recurrent	$1.64\% \pm 0.16\%$	$\mathbf{13.41\%} \pm 0.94\%$	$\mathbf{8.74\%} \pm 0.38\%$
Purely feed-forward	$1.75\% \pm 0.13\%$	$\mathbf{18.42\%} \pm 0.93\%$	$\mathbf{13.68\%} \pm 0.64\%$

Table 3.1: Comparison of generalization block error between the purely feed-forward and fully recurrent memory network configuration on identity layer. The configurations were tested after learning time of $5 \cdot 10^5$ cycles. Fully recurrent configuration shows a significantly better performance on the duplicate views with emotional expression (either smiling or sad), while comparable performance is shown on the other views.

identity error.

A striking result is the significant discrepancy in recognition performance between the two configurations manifested on the views that deviate stronger from the original view used for training. Those views are the duplicate views with emotional expression, sad or smiling. There, the error rate for identity and for gender is significantly lower for the fully recurrent memory network configuration. The difference in error rate is particularly large on the parts layer, getting smaller on the identity layer. The identity layer shows in general smaller identity error rates across all views. This may indicate that units on the identity layer are capable of capturing the compositional identity of the faces from their lower layer parts-based representation in favorable way.

Remarkably, for the alternative views that do not deviate much from the original training view, no difference in error rate can be observed between the alternative network configurations. Apparently, the stronger the deviation of the alternative view from the original view showed during the learning, the more evident is the benefit of the fully recurrent configuration over the purely feed-forward one,

Configuration	Views, identity error rate (parts layer)		
	Original	Smiling	Sad
Fully recurrent	0%	14.29% ± 0.34%	13.94% ± 0.32%
Purely feed-forward	0%	15.1% ± 0.31%	13.69% ± 0.29%

Configuration	Views, identity error rate (parts layer)		
	Duplicate	Duplicate, smiling	Duplicate, sad
Fully recurrent	**3.01%** ± 0.07%	**32.05%** ± 0.62%	**16.02%** ± 0.23%
Purely feed-forward	**9.82%** ± 0.13%	**41.39%** ± 0.27%	**29.17%** ± 0.44%

Table 3.2: Comparison of generalization block error between the purely feed-forward and fully recurrent memory network configuration on the parts layer. The configurations were tested after learning time of $5 \cdot 10^5$ cycles. Again, fully recurrent configuration outperforms significantly the purely feed-forward version on the duplicate views with emotional expression, smiling or sad, while showing comparable performance on the other views.

Configuration	Views, gender error rate (parts layer)		
	Original	Smiling	Sad
Fully recurrent	0%	6.7% ± 0.25%	3.84% ± 0.21%
Purely feed-forward	2.5%	7.08% ± 0.24%	3.52% ± 0.2%

Configuration	Views, gender error rate (parts layer)		
	Duplicate	Duplicate, smiling	Duplicate, sad
Fully recurrent	0.64% ± 0.15%	**7.86%** ± 0.27%	**6.2%** ± 0.13%
Purely feed-forward	1.85% ± 0.16%	**12.44%** ± 0.34%	**9.25%** ± 0.29%

Table 3.3: Comparison of generalization block error for gender between the purely feed-forward and fully recurrent memory network configuration on the parts layer. The configurations were tested after learning time of $5 \cdot 10^5$ cycles. Again, fully recurrent configuration outperforms significantly the purely feed-forward version on the duplicate views with emotional expression, while showing comparable performance on the other views that do not deviate strongly from the original one used during the learning.

whereas this advantage is no longer seen if the test view is similar to the original. This means that lateral and top-down connectivity specifically improves the ability of the memory network to generalize over the novel data.

Given only a short time of a single decision cycle, the recurrent connectivity is thus already able to assist and improve the recognition process. However, this improvement is not evident on the face views similar to the original training view. There is an intuitive explanation for this phenomenon. The contextual lateral and top-down connectivity is particularly useful for supporting the decision making in situations where the local sensory information provided by the input is ambiguous and therefore cannot be interpreted correctly without further contextual cues. If, for some reason, there is no such ambiguity in local sensory signal and the local interpretation is clear without any need for further information exchange and disambiguation, a purely feed-forward processing should perform as well. This may occur in overlearned situations, like it is for instance the case for the training data or input data closely resembling it. In such situations, the network can rely on the bottom-up connectivity formed in the course of learning to deliver the correct interpretation of the face identity based purely on local facial appearance.

This overfitting to the training data is punished as soon as novel data arrives, where local appearance

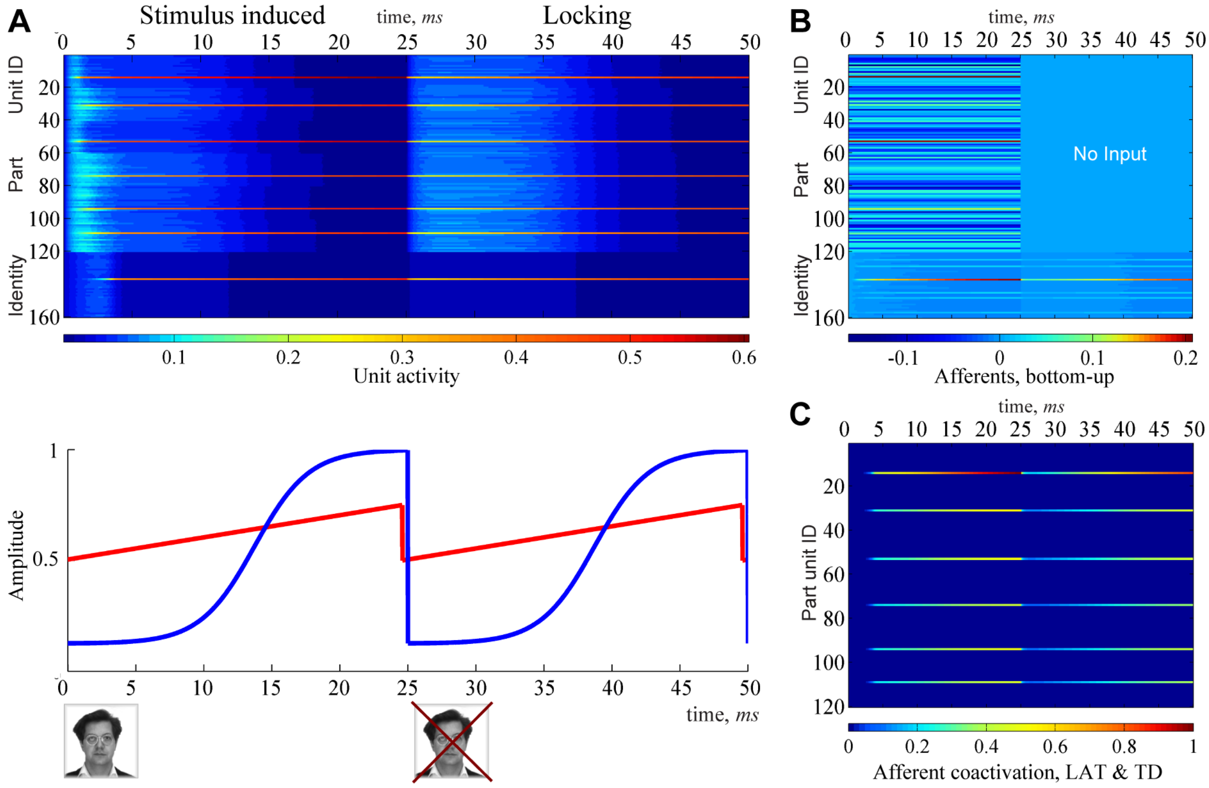


Figure 3.17: Locking of stimulus-induced activity after stimulus removal. **(A)** Using oscillation tuning ($\nu_{min} = 0.15$, $\omega_{min} = 0.5$), the activity induced by a face stimulus in the first cycle can be locked, persisting in the successive cycle if the stimulus is removed. **(B)** Bottom-up afferents arriving on parts layer. **(C)** Afferent coactivation is still high after stimulus removal, suggesting strong cooperation between the units of the persisting assembly.

alone cannot provide such a strong clear cue for the right interpretation. The purely feed-forward configuration, relying only on the similarity computation upon local appearance, is much more probable to suffer from wrong interpretation of the local parts and consequently, of the whole identity. The fully recurrent configuration gains there its benefit, being able to use contextual information to resolve the local ambiguities and arrive at a coherent correct interpretation of the face stimulus despite less familiar local appearance of the novel face view.

3.2 Processing properties during memory recall

3.2.1 Locking : persistent activity after stimulus removal

As shown in the previous section, the memory network structure formed in the course of unsupervised learning from natural face images provides the functionality to recall a compositional representation of a memorized face by presenting its image on the input. This recall is done within a single decision cycle. The activity pattern formed during the recall holds at the end of the cycle the representation of the face shown to the memory network. This activity pattern can be locked, or sustained, over the successive cycles even if the stimulus is removed from the input (Fig. 3.17). This can be achieved by simple alteration of the ongoing excitatory and inhibitory rhythms for all modules in the network. The amplitude of the excitatory rhythm ω is increased, in parallel increasing the amplitude of the inhibitory rhythm ν to keep the activity sparse (in following, I will refer to this procedure as to *oscillation tuning*).

This manipulation of ongoing rhythms has two consequences. First, the self-excitatory coupling of units in the network gets stronger. Second, the influence of contextual signal exchange between the modules over lateral and top-down connections also increases. Now, if a face is presented on the input, the winner assembly formed during the decision cycle has a very strong support by both increased self-excitatory coupling within the units and by increased cooperation between them due to the stronger influence of lateral and top-down signaling. This intrinsic network support allows the winner assembly to stay active in the successive cycles without any external stimulation. The duration of persisting activity in this locked state depends on the time constant of homeostatic activity regulation. If this mechanism were disabled, the locked state could be maintained technically for an infinite period of time, which is of course not plausible in biological terms, where adaptation and fatigue would set in at some point. The maintenance of persisting activity over time is particularly important in delayed-match-to-sample tasks, where the representation of the initially shown sample stimulus has to be kept for a longer delay time period in the working memory in order to be able to match it to the correct alternative from the choice stimuli shown later [Fuster and Alexander, 1971, Miller et al., 1993].

Moreover, increasing the amplitude of the excitatory rhythm even higher would result in a locking state which is completely insensitive to any external input, keeping once generated activity pattern safe from disturbance by external distractors. Such a mechanism could be useful for protecting content of working memory from being rewritten by incoming irrelevant sensory input. This kind of functionality is required for instance in delayed-match-to-sample tasks with distractors presented during the delay time period [Fuster, 1973, Miller et al., 1996].

3.2.2 Generative pattern completion and attentional mechanisms

Established memory traces for the faces shown during learning allow to recover the full compositional face representation by providing only partial cues for the recall. The cues can be provided to a subset of parts layer units (*bottom-up* cue) as well as to single identity layer units (*top-down* cue) to recall the parts-based description of a memorized face identity. Different ways to provide a cue signal are possible and lead to the same effect. A cue can be an additional external input to the unit, a slight elevation of the unit activity level at the cycle's begin, or different other manipulations like increasing the self-excitatory coupling, etc. For simplicity, I use direct manipulation of unit's activity level to demonstrate the functionality.

First, let us look on memory recall induced by partial bottom-up cues. Two part units that belong to a memory trace of a particular person (same person as used in Fig. 3.17) are made slightly more active ($p = 0.1$ for better visibility on the plot, lower activity $p = 0.05$ will do as well) then the rest ($p = 0.02$) at the begin of the decision cycle (Fig. 3.18 **(A)**). This priming leads to rapid (about $5ms$ after giving the cue) activation of the remaining part units that belong to the same trace as well as of the corresponding identity unit. Note that no external input is provided, the only cue was priming of the two part units. If the network is run in the locking regime by oscillation tuning like described previously, the reactivated memory is kept further in the successive cycles. Strong cooperation between the units of the reactivated memory trace is evident from the high co-activation of the lateral and top-down afferents converging on the units. This strong cooperation indicates that the reactivated units indeed correspond to a memorized assembly, which is linked together via existing lateral and top-down connectivity.

The same priming procedure can be applied to a single identity unit, inducing memory recall via a top-down cue (Fig. 3.19 **(A)**). Again, the full compositional description of the respective face identity is reactivated in form of a unit assembly corresponding to the memory trace of the face. The recall time is slightly shorter (less then $5ms$ after providing the top-down cue) than for the bottom-up cue induced

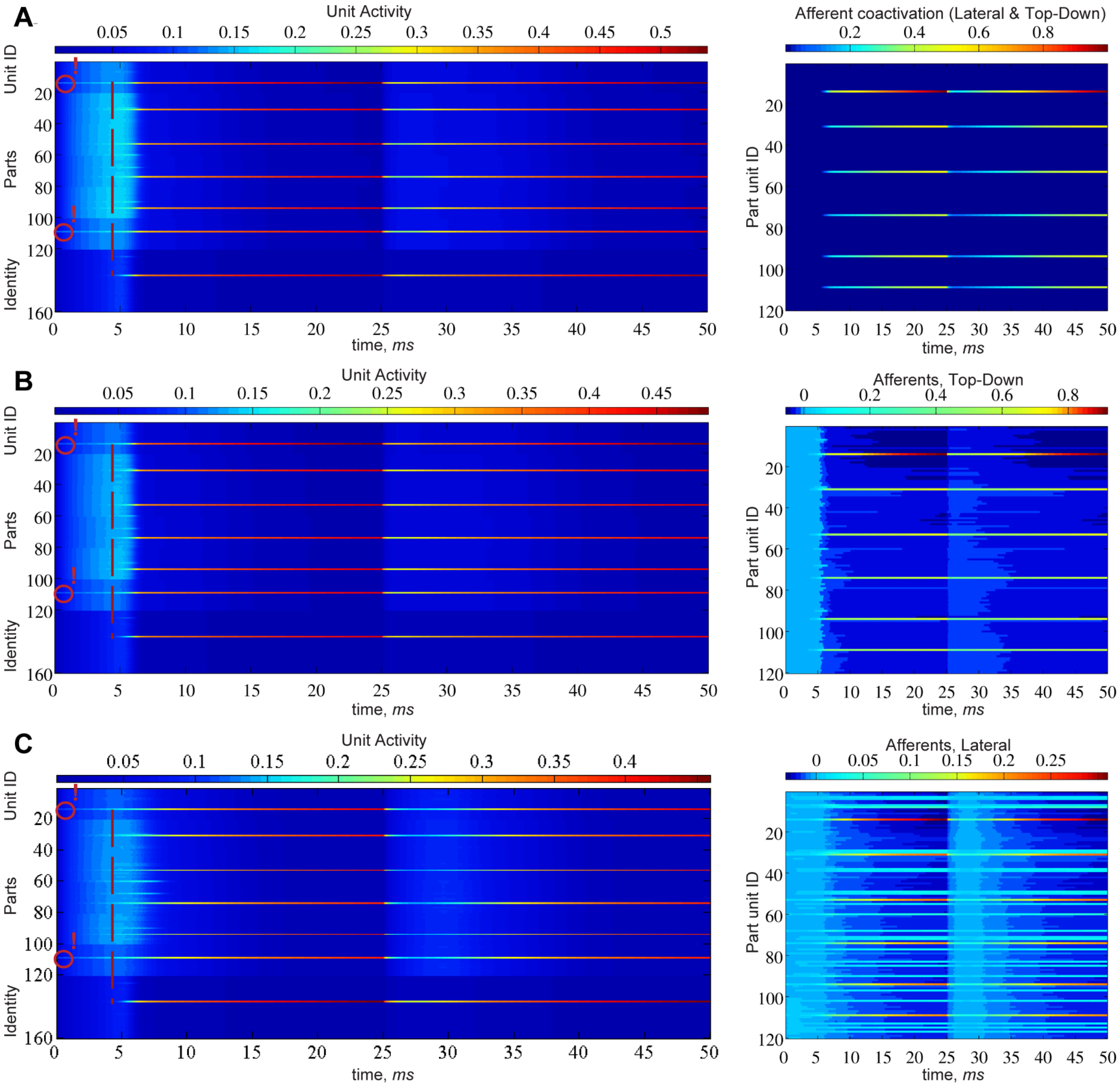


Figure 3.18: Memory recall induced by partial bottom-up cues in absence of sensory stimulus (no face image is presented to the network). Priming a small fraction (two part units) of the memory trace on the lower layer leads to a successful recall of the full compositional pattern of a memorized face. Two successive decision cycles are shown. **(A)** Memory activity during the recall induced by a bottom-up cue, retrieving the parts-based representation of an individual face (initial activity of the two primed part units $p = 0.1$, the rest $p = 0.02$). The completed activity pattern is sustained after the recall, persisting in the successive cycle. High level of afferent coactivation shows that the reactivated assembly corresponds to a memorized content. The cue-induced recall and pattern completion still work if lateral **(B)** or top-down connectivity **(C)** is deactivated. This provides strong evidence that each connectivity type can independently mediate contextual support.

recall.

Memory trace reactivation induced by parts-based bottom-up or object-based top-down cues demonstrates clearly the generative nature of the established memory network. The pattern completion can occur bottom-up, using the partial information about some parts, or it can work by activating a higher-order symbol for the object identity, resulting in both cases in reactivation of the full object description

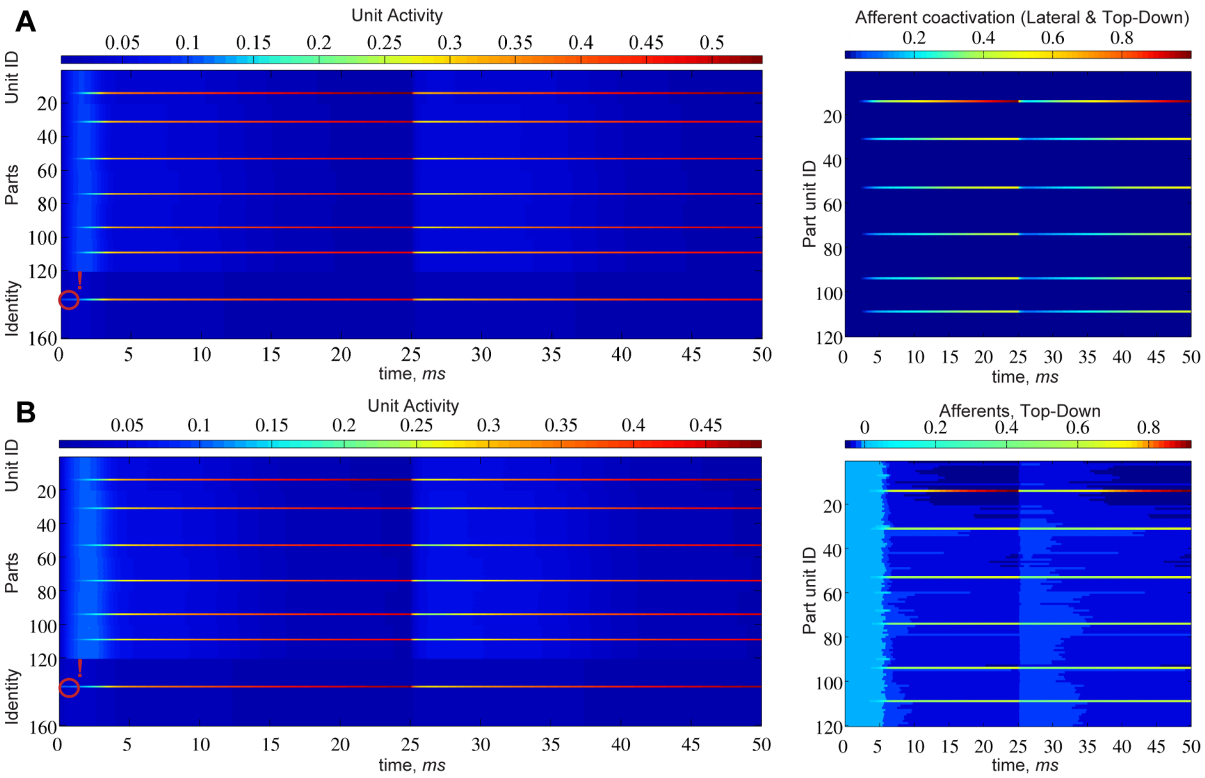


Figure 3.19: Memory recall induced by top-down identity cue in absence of sensory stimulus (no face image is presented to the network). Priming an identity unit on the higher layer leads to a successful recall of the full compositional pattern of a memorized face. Two successive decision cycles are shown. **(A)** Memory activity during the recall induced by a top-down identity cue, retrieving the parts-based representation of an individual face (initial activity of the primed identity units $p = 0.1$, the rest $p = 0.02$). The completed activity pattern is sustained after the recall, persisting in the successive cycle. High level of afferent coactivation shows that the reactivated assembly corresponds to a memorized content. The cue-induced recall and pattern completion still work if lateral connectivity is deactivated **(B)**.

as composition of all its parts. One could ask whether this functionality is supported by both lateral and top-down connectivity or whether it relies on one particular connectivity type only. To answer this, three different setups are tested for the ability to perform the generative pattern completion after priming, deactivating one type of contextual connectivity. For the bottom-up completion, either only lateral or only top-down connectivity remains active. For the top-down completion, top-down connectivity is obviously essential, as otherwise no information about the primed identity can ever get to the vocabulary layer, so only lateral connectivity can be disabled for testing.

The outcome of these experiments shows that each connectivity type can alone mediate the generative pattern completion. Deactivating lateral (Fig. 3.18 **(B)**) or top-down connectivity (Fig. 3.18 **(C)**) in bottom-up priming does not prevent the proper memory reactivation, neither does the deactivation of lateral connectivity in top-down priming paradigm (Fig. 3.19 **(B)**). This provides strong evidence that each connectivity type indeed mediates proper contextual signaling, contributing to the memory trace linkage.

The functionality of generative pattern completion by priming can also be interpreted as an attentional mechanism, which selectively enhances processing of particular memorized content at expense of the rest [Tsotsos et al., 1995, Desimone, 1996, Reynolds et al., 1999]. In this view, parts priming can be interpreted as bottom-up feature-based attention, while identity priming would be a form of

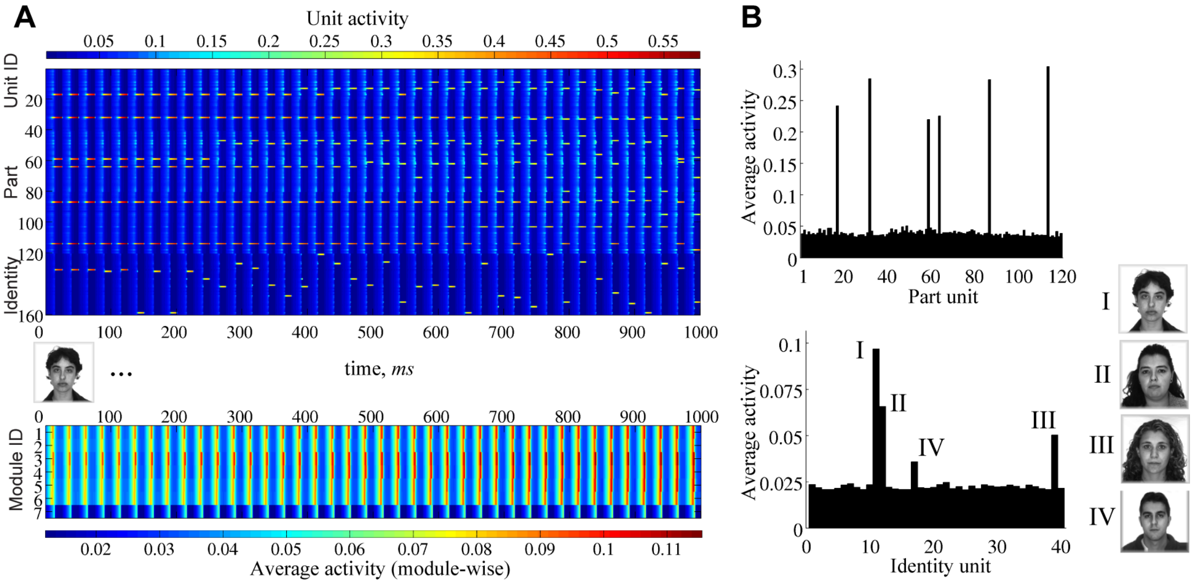


Figure 3.20: Memory recall and encoding over multiple gamma cycles. Fixing a face on the input for many cycles, a transient attractor dynamics can be observed in the network. **(A)** Activity formation in time interval of $1000ms = 1s$ (corresponding to 25 cycles). The attractor states are visited transiently, with part and identity units leaving and joining the assembly. **(B)** Average unit activity measured for the first $250ms$ after stimulus onset. Part units assembly were stable during this time, which is reflected in the high average activity of the 6 winner part units. Multiple identity units participated on assembly in different cycles. Their average activity over multiple cycles is consequently lower. Potentially, the face identity could be represented by weighting the evidence from each identity unit according to its average activity. If the activation levels are ordered, the units may also provide a kind of ranking code for the face on the input.

top-down, object-based attention. The mechanism of attention does not require any specific dedicated subsystem here, it is implemented in the network naturally by competitive-cooperative effects between the distributed modules mediated by their dynamics and the connectivity structure established during the learning.

3.2.3 Recall and encoding over multiple cycles

As pointed out in Sec. 2.6, the dynamics of a single module, which is employed in the network as a basic computational node, can be interpreted in terms of visiting transient attractors if observed over many successive decision cycles. In the network, the attractor structure is much complexer. The phase space contains now distributed unit activities of strongly coupled modules. The attractor landscape is defined by synaptic connectivity linking distributed units into assemblies that correspond to stored memories for the faces shown during learning. An attractor state can be reached within one single cycle by presenting an image of a memorized face, which activates the corresponding winner unit assembly.

Again (see Sec. 2.5.2), if the face image is held fixed on the input the network does not stay in attractor state forever in successive cycles (Fig 3.20). This is due to ongoing homeostatic regulation of unit activity, which makes the winner units progressively insensitive to the incoming afferent input by downregulating their intrinsic excitability. The lateral and top-down connectivity linking the units of the assembly to a memory trace works against the destabilization effect of intrinsic plasticity. On the vocabulary layer, the parts-based unit assembly can remain active for a longer time before it gets destabilized, which depends on the strength of cooperation within the assembly, on the time constant of

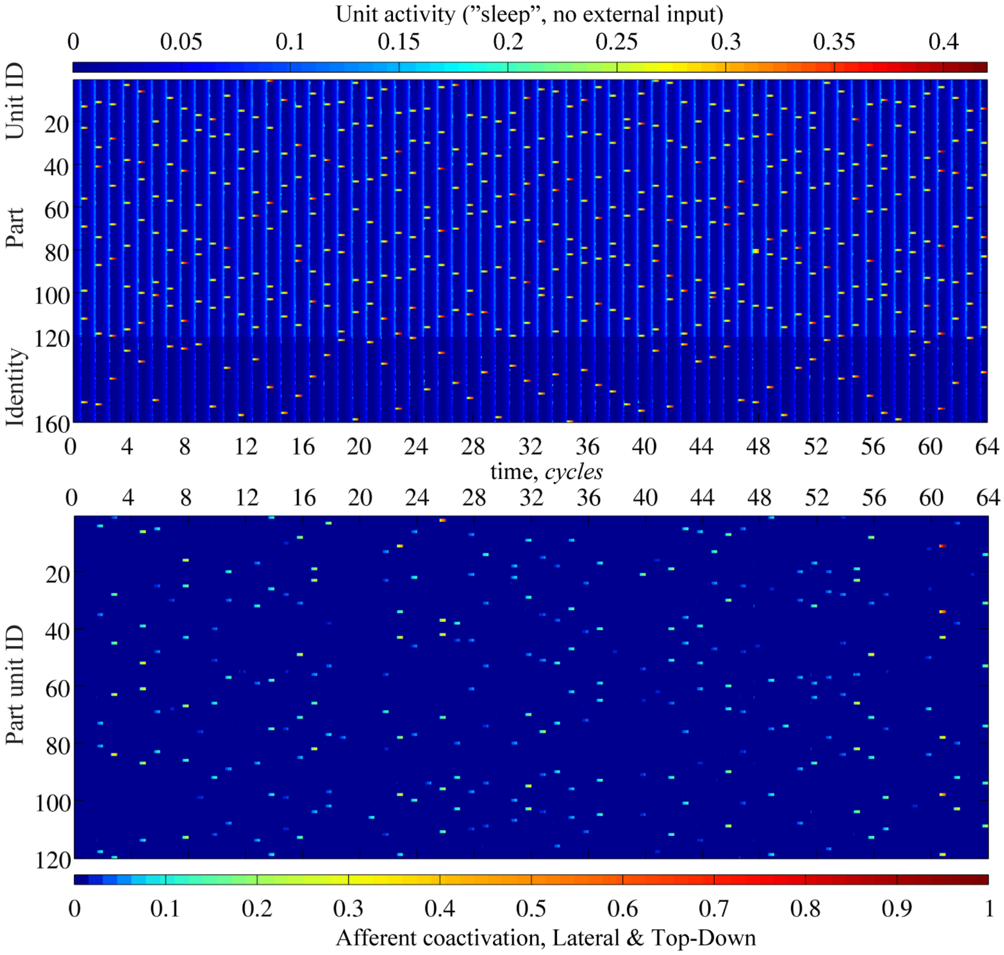


Figure 3.21: Memory replay in absence of external stimuli. Shown above is the activity spontaneously generated by the network in the off-line, "sleep" regime (64 successive cycles shown, $T = 25ms$ each). Below is the afferent coactivation plot. Some of the generated activity patterns have a high degree of cooperation among the assembly units, indicating the replay of a memorized face.

intrinsic plasticity and on the level of neuronal noise. The transition to another identity label signaled by an identity unit is easier, because identity units have to share common part units on the lower layer. While the part unit assembly may still be stable, the identity units may change, signaling for identities that are similar in their composition to the face shown on the input.

This type of representation can be seen as a probabilistic coding scheme (see also Sec. 2.5.2) signaling for compositional face identity. The alteration of identity units over multiple cycles can be interpreted as rank coding for the face identity by counting the win events for a particular unit or by measuring the average unit activity over a number of cycles (Fig. 3.20 **(B)**). Doing so, one would obtain an interpretation of face identity weighted by different prototypes stored in the memory. The same scheme can apply for the part units. Although I use here only the representation established in the course of a single cycle to read out the identity and gender of a person, in perspective a more elaborate decoding is possible, if the activity formation over multiple successive cycles is taken into account.

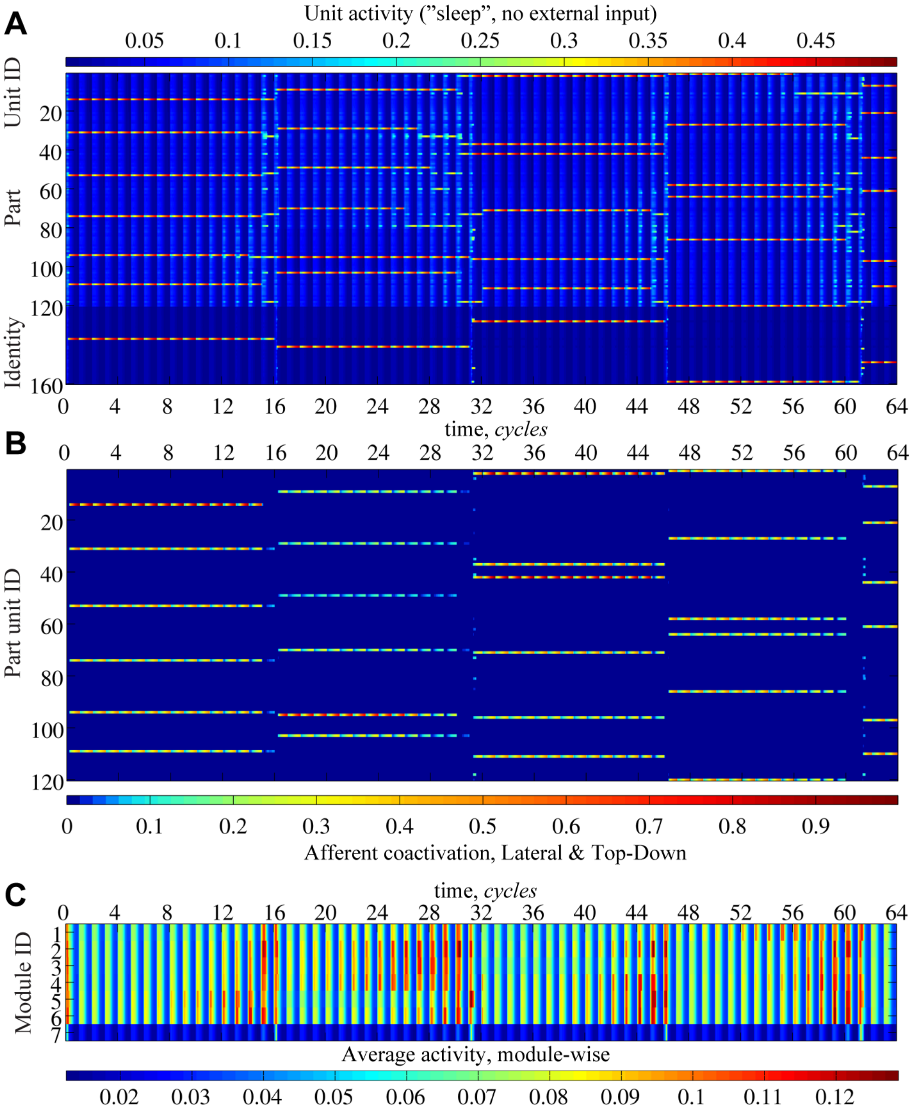


Figure 3.22: Coherent storage-based replay in off-line regime. Using oscillation tuning ($\nu_{min} = 0.15$, $\omega_{min} = 0.5$), the self-generated activity can be made strongly coherent, corresponding most of the time to the patterns already stored in the memory. **(A)** The memory replays the content stored during the learning, retrieving the parts-based representation of individual faces. Once recalled, the activity pattern representing a particular face persists for around 16 cycles ($400ms$). The afferent coactivation is high within the activated assembly, indicating its correspondence to an already stored content **(B)**. Then, a spontaneous transition occurs changing the activity pattern to a new one. Shortly before the transition, the average activity level within the modules shown in **(C)** begin to raise, until it reaches a critical level, where the transition sets in. Immediately after transition the average activity drops back to its normal level, and the generative procedure repeats again.

3.2.4 Self-generated memory replay in absence of external stimuli

As we already saw, the explicit compositional, generative nature of memory traces formed during the learning allows the reactivation of stored faces by priming of either part or identity units, without applying an external image input. Now, how would the network behave if there is absolutely no intervention to its dynamics from outside, neither via image input nor via priming? To see what happens in the network under these conditions, I take the network with mature connectivity and let it run without providing any external input.

Decoupled from external input, the memory network shows spontaneously generated activity patterns (Fig. 3.21). The self-generated activity emerges within the fragments of the ongoing decision cycles. In each decision cycle, selection and amplification of a winner unit assembly takes place, analog to the activity formation in the stimulus-driven, "wake" regime. The evoked assemblies are different from one cycle to the next. For each evoked assembly, the corresponding degree of the co-activation of lateral and top-down afferents provides a hint whether the winner assembly reflects a memorized face identity or not. Some of the self-generated activity patterns are accompanied by a high degree of afferent cooperation, thus corresponding exactly to the stimulus-induced response that would be generated if an already familiar face were presented on the input.

Other patterns are in turn not a replay of the faces stored in the memory, but can be more or less arbitrary combinations of part and identity units. The degree of afferent co-activation for these patterns is consequently low. Such activity patterns can be interpreted as "phantasized" faces never experienced before. The amount of self-generated patterns corresponding to stored or "phantasized" faces can be influenced by tuning the ongoing rhythms. The fact that increasing the amplitude of the excitatory rhythm ω enhanced the strength of contextual binding was already used here for locking the stimulus-induced activity over successive cycles. Applying the same oscillation tuning procedure in absence of any input leads to a more coherent, storage-based memory replay.

If the amplitude of ω is increased further, the qualitative dynamical behavior of the network changes, switching to a regime which is analog to the locking mode described previously in this section. Once evoked, the winner unit assembly is not replaced by a new one in the next cycle, but persists for a longer time period of multiple cycles (Fig. 3.22). Then, a spontaneous transition occurs, and the network selects another assembly to continue with. The consistently high level of afferent cooperation suggests that the recalled assemblies correspond to the representation of faces already stored in the memory.

The memory replay in this coherent regime reveals the attractor states of the network, visiting the attractors transiently by reactivating one assembly after another. Again, the duration of visit depends on the established network connectivity and the time constant of unit intrinsic plasticity. A nice property of this off-line replay is that it provides another opportunity (in addition to analysis of the connectivity structure) to label the units as being members of the same memory trace. This labeling can be potentially performed in unsupervised fashion without necessity to know the particular stimuli used to induce the memories. Further advantages of this kind of off-line memory reprocessing for learning are subject of the following chapter.

3.3 Rapid, non-synaptic learning via excitability regulation

In previous section, the block generalization error was used to test recognition performance of the memory network (Sec. 3.1.3). This type of generalization error was employed to free the performance evaluation from the bias imposed by the recent history of network activation, which affects the excitability of the units and alters their probability to win given an input. A deeper investigation of the recognition performance shown during the test for block generalization error reveals an interesting phenomenon. The recognition rate turns out to be much lower for the very first blocks of presented novel data, improving strongly and rapidly in the immediate successive blocks (Tab. 3.4, Fig. 3.23). Synaptic modifications can be excluded as the improvement cause, as synaptic plasticity was completely disabled during the whole test phase.

This is a puzzling phenomenon, as it appears that the initial difficulties the system has with the correct

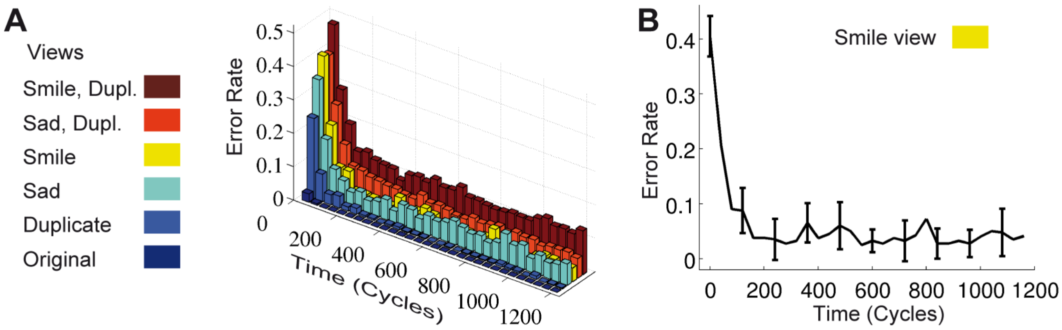


Figure 3.23: Improving recognition performance on blocks of novel data (40 cycles correspond to one presented block), only homeostatic activity regulation is active in the network. **(A)** Time course of the improvement effect shown for different views. Strong and rapid error rate drop is evident across all views. **(B)** Improvement time course for the view with smile expression, with standard error deviation over multiple trials (10 in total).

View	First Block (40 cycles)	After 3rd Block (120 cycles passed)
Original	$2.5\% \pm 2.04\%$	0%
Duplicate	$24\% \pm 2.69\%$	$0.61\% \pm 0.16\%$
Sad	$34.5\% \pm 4.53\%$	$3.7\% \pm 0.4\%$
Smiling	$40.5\% \pm 3.69\%$	$4.3\% \pm 0.63\%$
Dupl. Sad	$39.75\% \pm 4.16\%$	$7.02\% \pm 0.36\%$
Dupl. Smiling	$47.75\% \pm 2.19\%$	$11.62\% \pm 0.96\%$

Table 3.4: Performance improvement on blocks of novel data in absence of synaptic plasticity, only homeostatic activity regulation being active. Drop in identity error rate on the identity layer is particularly strong for the alternative face views deviating strongly from the original view shown during the learning.

recognition of the novel data are abolished without any obvious modification of the synaptic structure in the network. Usually, it is the synapse-specific plasticity which is taken to be the main force behind the optimization of the network function. Here, the only candidate for the observed improvement is the mechanism of the intrinsic plasticity. Intrinsic plasticity was the only adaptive mechanism active during the test phase, performing homeostatic activity regulation across the network units.

Obviously, the homeostatic activity regulation adjusts the excitability of the units in the network during the test phase in a way that is beneficial for the memory function. Two questions arise here. First, what is the specific change in excitability levels that makes the network perform better? Second, is there a shortcut to produce the same effect without showing blocks of novel data in advance? To address the first question, the time course of excitability regulation during the test phase is analyzed (Fig. 3.24). There is a slight downregulation of excitability levels occurring on both part and identity layers. Remarkably, on the identity layer there is clear tendency of excitability levels to move closer together, regularizing the differences in excitability between the identity units (Fig. 3.24 **(B)**). This equalization of the excitability levels makes consequently the identity unit win events to become more uniformly distributed.

As this observation is consistent across testing on different alternative views, it becomes suggestive to hypothesize that the essential change produced by homeostatic activity regulation in the network, leading to strong improvement of recognition performance, is the regularization, or equalization, of the unit excitability levels. This hypothesis is easily verified by equalizing the excitability levels per hand and then comparing the immediate generalization error of the network before and after the manual regularization. Immediate generalization error is evaluated in the network with both synaptic and

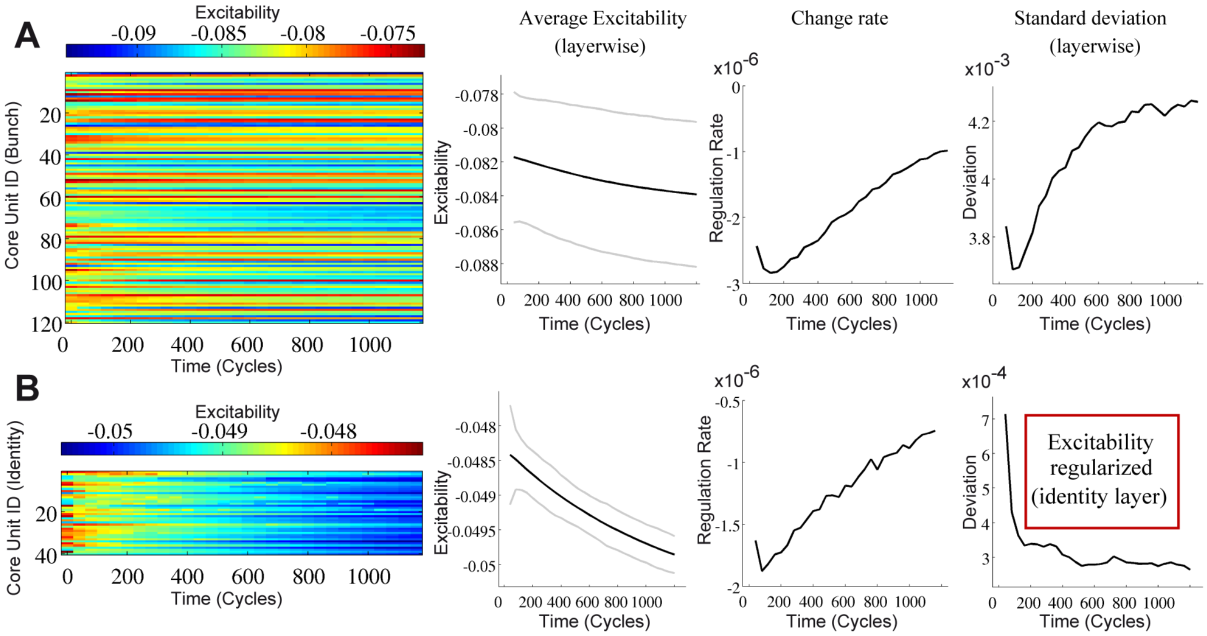


Figure 3.24: Time course of excitability regulation during repetitive exposure to novel data blocks (presenting face views with smile expression), on the lower vocabulary layer **(A)** and higher identity layer **(B)**. Slight excitability downregulation is visible on both layers. Excitability levels of identity layer units get regularized by moving closer together (deviation becomes less than a half of the initial value, **(B)** on the right). On the vocabulary layer, no such regularization is evident (though a slight tendency is there within the first $100 - 200$ decision cycles).

intrinsic plasticity deactivated (Sec. 3.1.1). The immediate generalization error measured in the original network state is high, as expected from the previous evaluation of the time course of the block generalization error. An improvement can be made by the oscillation tuning procedure, enhancing the influence of contextual lateral and top-down connectivity. However, even much stronger improvement sets in if the excitability levels of the units are set manually to the same level (Fig. 3.25).

The tremendous drop in error rate across all views confirms the stated hypothesis about the beneficial effect of excitability level regularization on the recognition performance. Importantly, the manual excitability regularization provides a simple technical shortcut that produces the positive effect without showing any novel data to the network in advance, thus answering the second question. As no presentation of alternative views is required for the improvement of the recognition performance on novel data, the manual excitability regularization can be considered as a method to boost the generalization capability of the memory network. The positive effect does not depend critically on the particular direction of regularization. The excitability levels can be set either to maximum, average, or minimum value, computed either layer- or modulewise. The effect stays basically the same, showing better performance if maximum value is chosen to equalize the levels (Fig. 3.26) This suggests that what matters for the induction of the positive effect is indeed the flattening of the excitability levels, and not a specific direction they move to.

An important observation is made by comparing the fully recurrent network configuration with the purely feed-forward one after applying the manual excitability regularization procedure. Although both configuration take benefit in recognition performance from the excitability regularization (Fig. 3.25, 3.27), the positive effect is articulated significantly stronger for the fully recurrent configuration (Fig. 3.28). This difference gets even stronger pronounced if in addition the oscillation tuning is performed,

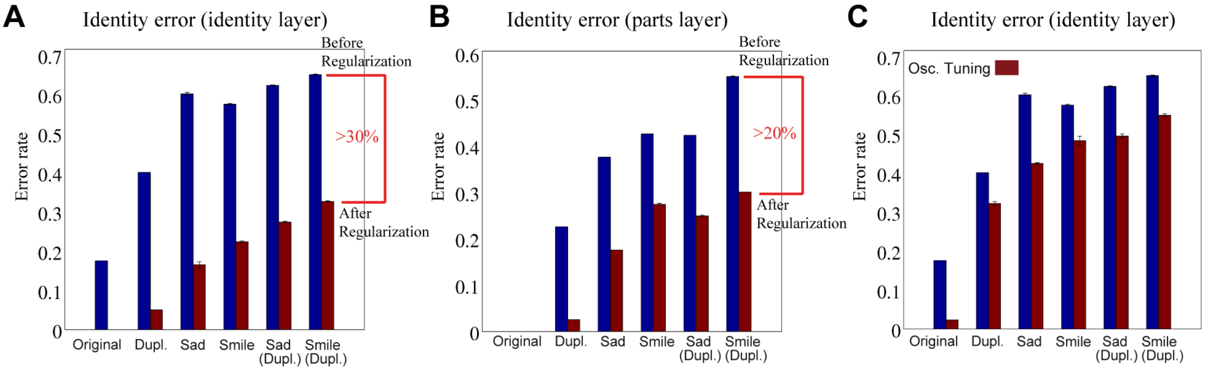


Figure 3.25: Comparison of recognition performance before and after manual regularization of excitability levels for the recurrent network configuration. The regularization is done module-wise by setting the levels to the maximum within the respective modules. Identity error rate for identity **(A)** and parts **(B)** layer is shown. Excitability regularization procedure leads to a dramatic drop in error rate across all views. The positive effect resembles the one encountered during the presentation of the novel data blocks, providing evidence for the causal role of the excitability regularization in the observed functional improvement. In comparison to this effect, oscillation tuning oscillation tuning ($\nu_{min} = 0.1$, $\omega_{min} = 0.5$) can only slightly improve the recognition performance **(C)**.

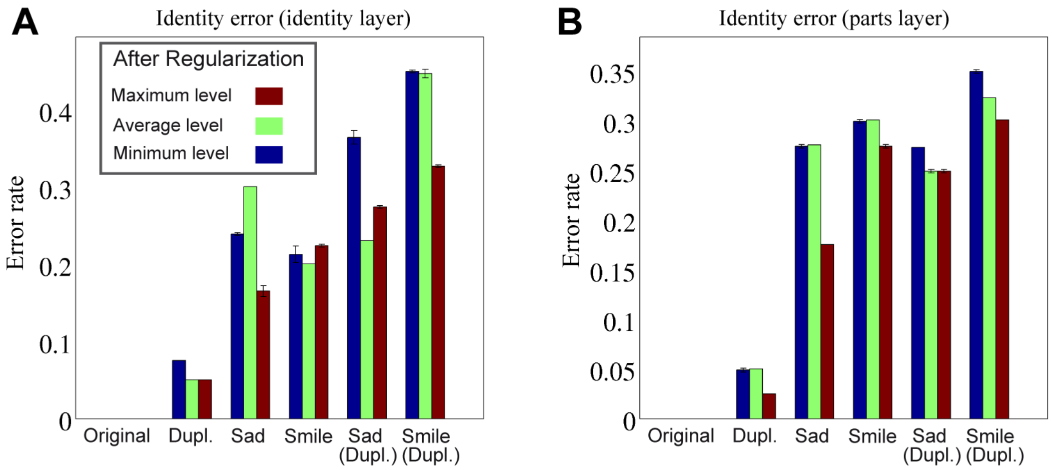


Figure 3.26: Effect of different regularization procedures on recognition performance. Manual regularization procedures were performed by setting excitability levels module-wise either to minimum, average or maximum value within the modules. Identity error after regularization shown for the identity **(A)** and the parts layer **(B)**. The positive effect is qualitatively independent of the particular regularization direction. In the upregulated condition (dark red), the effect is slightly stronger.

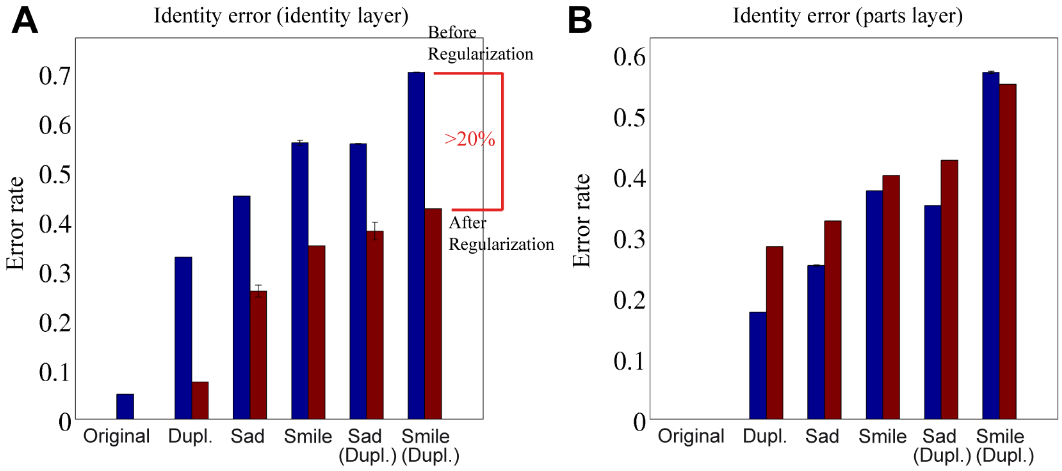


Figure 3.27: Analog to the positive effect observed for the recurrent network configuration, the purely feed-forward version also benefits from the manual regularization procedure. Identity error rate for identity **(A)** and parts **(B)** layer is shown.

increasing the contextual influence via lateral and top-down connectivity. Oscillation tuning has no additional effect on the purely feed-forward configuration as expected. The difference in recognition performance between the two network configurations is of the same quality for the parts and identity layer. This form of rapid, synapse-unspecific learning mediated by excitability regulation is thus in particular beneficial for the fully recurrent network architecture, which clearly outperforms its purely feed-forward counterpart in both identity and gender classification after applying the procedure.

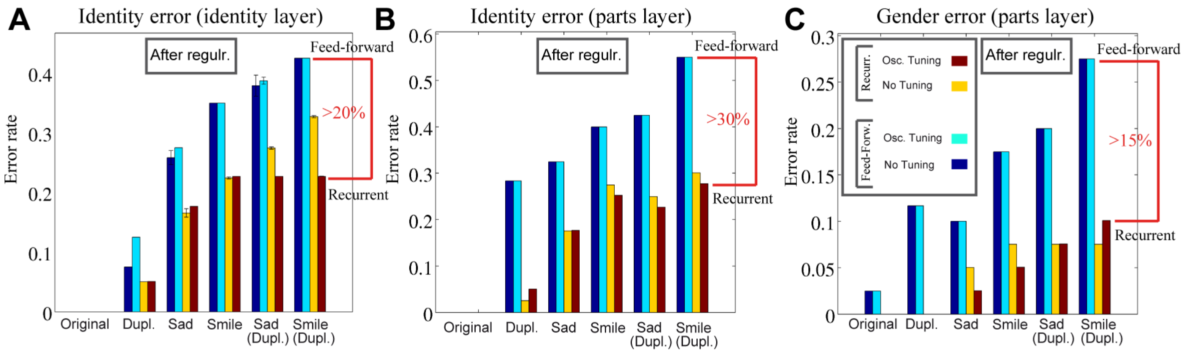


Figure 3.28: Comparison between recognition performance of feed-forward and recurrent configuration after the excitability regularization. Both configurations were tested with and without additional oscillation tuning (($\nu_{min} = 0.15$, $\omega_{min} = 0.5$)) after the regularization procedure (setting the excitability levels to the maximum within the respective modules). Identity error for the identity layer **(A)** and the parts layer **(B)** and gender error for the parts layer **(B)** are shown. The recurrent configuration clearly outperforms the feed-forward configuration in every scenario, the advantage is further amplified by additional oscillation tuning.

3.4 Remarks on scalability

In the previous chapter, it was demonstrated that a large training data set containing $P = \{120, 1000\}$ persons can be put into a memory in form of collections of local appearance descriptors learned by distributed, isolated modules ($M = 6$, $N = 20$ units each) attached to the dedicated landmarks (Sec.

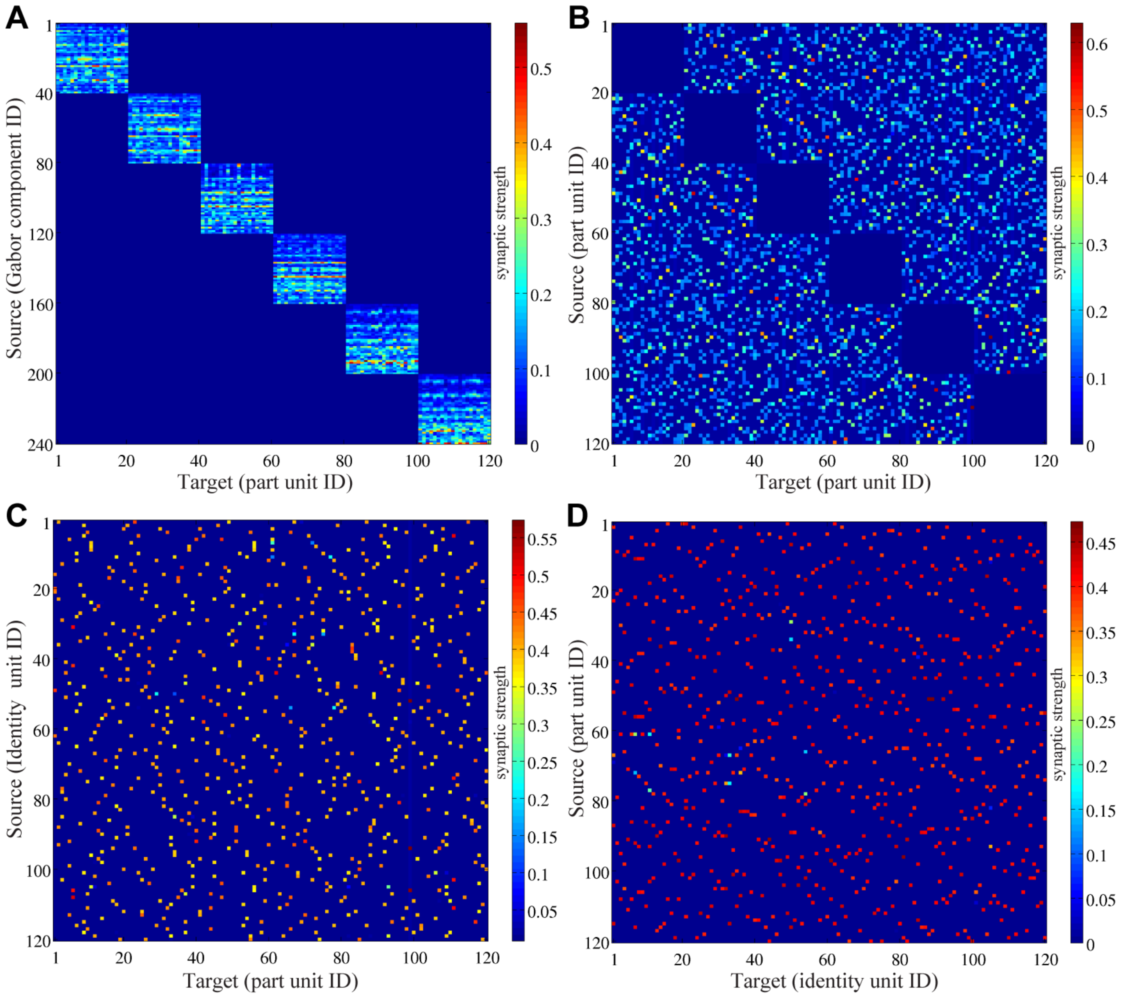


Figure 3.29: Network structure formed by learning 120 persons (connectivity state after $5 \cdot 10^5$ cycles) **(A)** Bottom-up connectivity on vocabulary layer containing appearance of local facial parts. **(B)** Associative lateral connectivity linking the parts to memorized face identities. **(C)** The bottom-up connectivity of identity layer holding face identities composed from the parts on the vocabulary layer. **(D)** The top-down connectivity projecting the compositional information back to the vocabulary layer. The connectivity matrix is roughly the transposed version of bottom-up connectivity shown in **(C)**.

2.5.1). Here, the modules are not longer isolated, so the question arises whether the system is able to handle this large number of persons while developing the proper lateral and top-down connectivity between densely occupied local appearance descriptors. Using only 6 landmarks to sample the local face appearance, we cannot expect particularly good generalization over novel data for that many persons. Still, we can demand that the network is able to form memory traces from the available data, capturing the compositional identity of faces from the large training set in a way that all network units participate to equal extent in memory formation, developing balanced selectivity for the persons shown during the learning. It can be also expected that the identity learning error on the training set should go close to zero even for this large number of persons, given that the lateral and top-down connectivity structure is established correctly without corrupting the memory traces.

Two setups are tested here to probe for the network scalability. The first setup uses the standard setting. It contains the parts and identity layer, with $M = 6$ modules on the parts layer containing

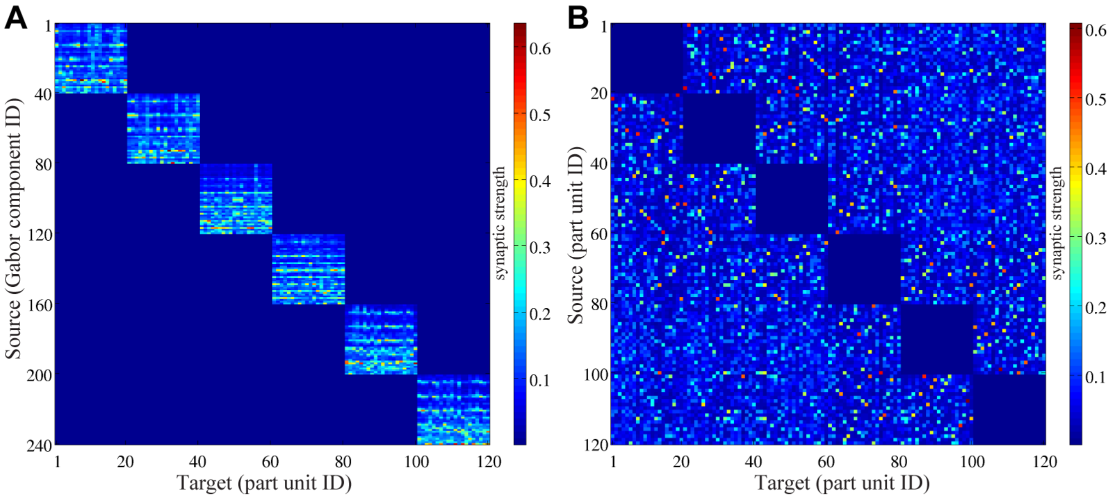


Figure 3.30: Network structure formed by learning 1000 persons ((connectivity state after $6 \cdot 10^5$ cycles)). **(A)** Bottom-up connectivity on vocabulary layer containing appearance of local facial parts. **(B)** Associative lateral connectivity linking the parts to memorized face identities. As part units have to be shared by a large number of memorized faces, the lateral connectivity does not seem to be sparse anymore.

$N = 20$ units each, and one identity module containing $N = 120$ units for learning on a training data set of $P = 120$ persons. The second setup contains only the parts layer of the same size for learning on a large training data set of $P = 1000$ persons. The learning procedure is the same as applied throughout this work, with exception of two modification applied to speed-up the convergence of learning. The maximum amplitude of the excitatory rhythm ω is increased to $\omega_{max} = 1.0$. The persons are presented blockwise, showing each identity one time per block in random order. This reassures that each person will be experienced by the network sufficiently often in reasonable amount of time after the learning begins.

Both setups are able to develop the full structural basis comprising bottom-up, lateral and, in case of the two layered setup, top-down synaptic connectivity for compositional face representation (Fig. 3.29, 3.30). The unit selectivity and unit usage load are well balanced, which is reflected by the appropriate number of persons sharing a part unit on average and by roughly uniform unit win probability observed on the vocabulary layer (Fig. 3.31, 3.32). Analog to previous experiments, there are also units that develop high selectivity for a particular gender.

Further, different persons from the training set (120 or 1000 persons depending on the setup) are successfully put into the memory domain after a certain amount of time spent in the learning phase. This is evident from the time course of the learning error, which resembles closely the findings made in Sec. 2.5.1. For the identity layer of 120 persons setup, the identity error goes to zero and vanishes completely (Fig. 3.33 **(A)**). For the parts layer, the identity error also drops to zero in case of memorizing 120 persons (Fig. 3.33 **(B)**) and saturates at a low level of about 2.5% in case of storing 1000 persons (Fig. 3.34 **(A)**). The gender error is below 5% for 120 persons (Fig. 3.33 **(C)**), for 1000 persons it is around 20% which is still well below the chance level (Fig. 3.34 **(B)**). As already pointed out in Sec. 2.5.1, the unsupervised learning of gender category based on purely local facial appearance from only few landmarks has apparently its limits if confronted with a large number of different persons.

Evaluation of block generalization error shows as expected only a mediocre performance on alternative views (Tab. 3.5, 3.6). The identity error rate increases for alternative views deviating stronger from the original. The same applies for gender error rate. Although gender recognition is substantially

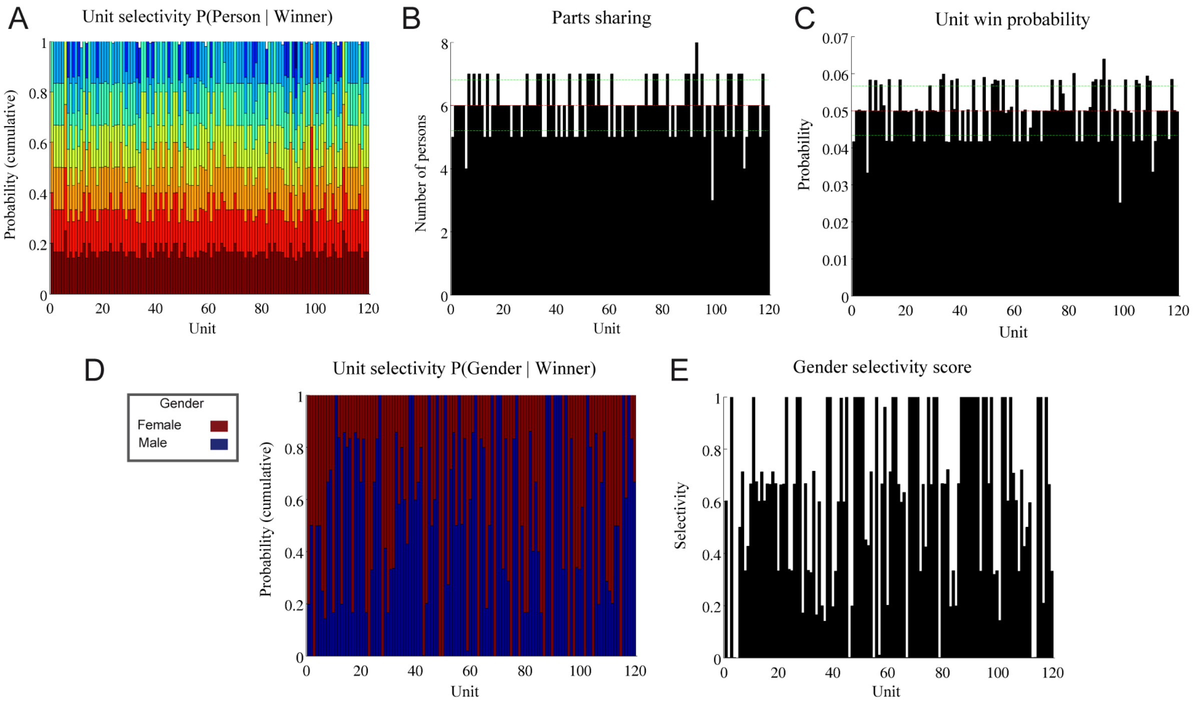


Figure 3.31: Unit selectivity on the vocabulary layer of the memory network ($M = 6$ vocabulary modules ($N = 20$ units each), $P = 120$ persons in the training set). The plots show the state after $5 \cdot 10^5$ decision cycles. **(A)** Selectivity for different persons. **(B)** Number of persons sharing a unit (6 on average), computed from unit selectivity. **(C)** Unit win probability computed over time interval of $2 \cdot 10^4$ cycles. The balanced unit usage load is reflected in roughly uniform win probability, with deviations corresponding to more or less utilized units. **(D)** Gender selectivity **(E)** Gender selectivity score. Some units develop very high gender selectivity (score 1), while others only poorly differentiate between male and female faces (score close to 0).

Error type	Views, Error Rate (after learning 120 persons)		
	Original	Sad	Smile
Identity layer, Identity	0%	$29.78\% \pm 0.32\%$	$37.62\% \pm 0.53\%$
Parts layer, Identity	0%	$35.03\% \pm 0.57\%$	$48.76\% \pm 0.65\%$
Parts layer, Gender	$4.17\% \pm 0.14\%$	$6.9\% \pm 0.22\%$	$16.54\% \pm 0.48\%$

Error type	Views, Error Rate (after learning 120 persons)		
	Duplicate	Duplicate, sad	Duplicate, smile
Identity layer, Identity	$33.89\% \pm 0.57\%$	$55.28\% \pm 0.74\%$	$60.01\% \pm 0.67\%$
Parts layer, Identity	$37.2\% \pm 0.23\%$	$58.34\% \pm 0.71\%$	$63.57\% \pm 0.58\%$
Parts layer, Gender	$5.76\% \pm 0.3\%$	$9.56\% \pm 0.29\%$	$17.28\% \pm 0.54\%$

Table 3.5: Generalization block error for the identity and parts layer after learning 120 persons. The identity and gender error get substantially larger for the alternative views deviating stronger from the original.

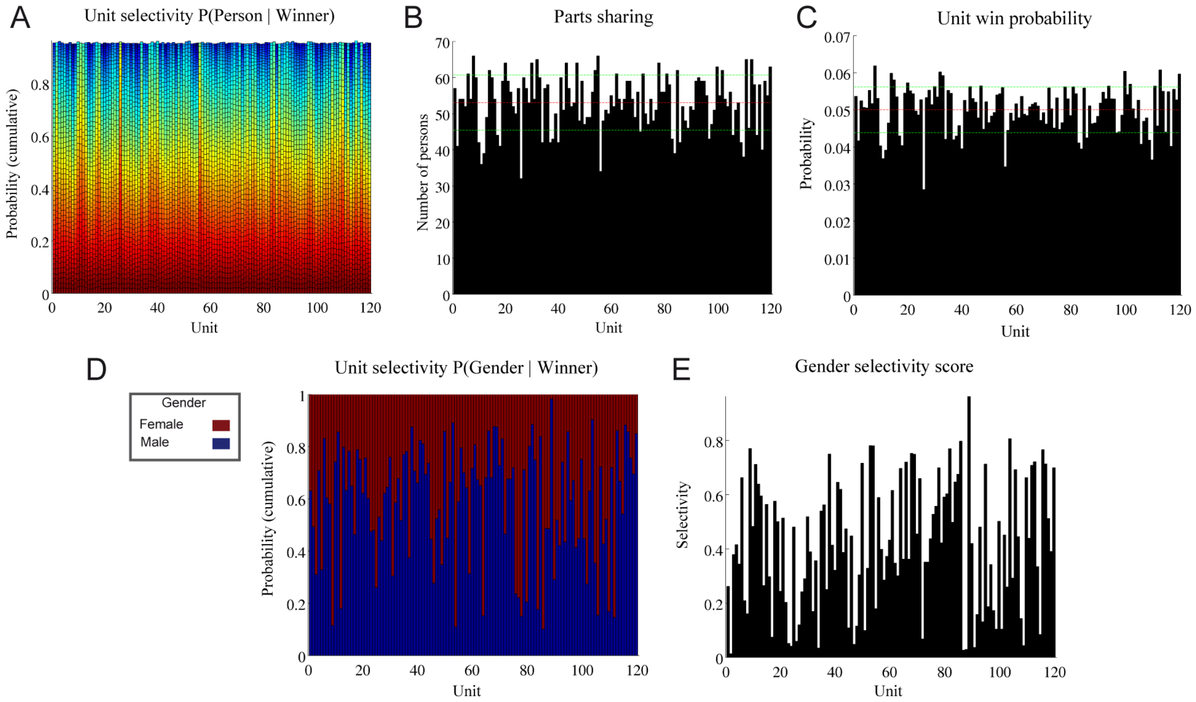


Figure 3.32: Unit selectivity on the vocabulary layer of the memory network ($M = 6$ vocabulary modules ($N = 20$ units each), $P = 1000$ persons in the training set). The plots show the state after $5 \cdot 10^5$ decision cycles. **(A)** Selectivity for different persons. **(B)** Number of persons sharing a unit (around 50 on average), computed from unit selectivity. **(C)** Unit win probability computed over time interval of $2 \cdot 10^4$ cycles. The balanced unit usage load is reflected in roughly uniform win probability, with deviations corresponding to more or less utilized units. **(D)** Gender selectivity **(E)** Gender selectivity score. The gender selectivity score is still high for many units, although no unit is able to reach the maximum score of one.

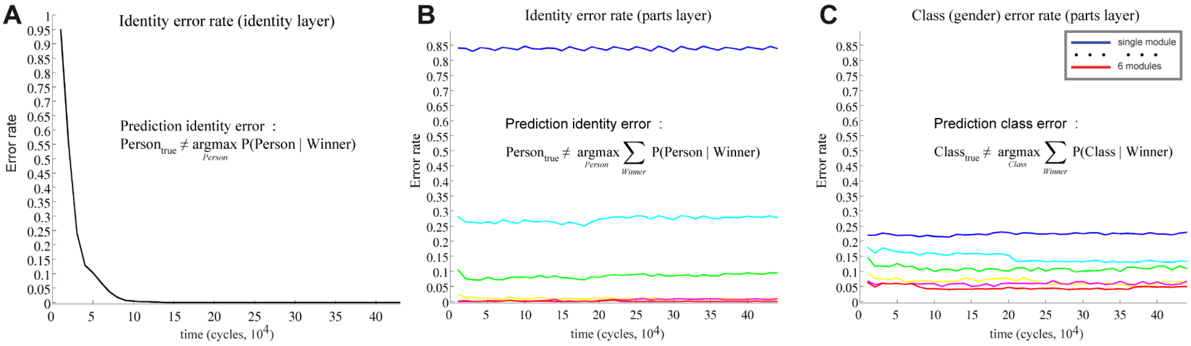


Figure 3.33: Learning error for the parts ($M = 6$ vocabulary modules, $N = 20$ units each) and identity (one identity module, $N = 120$ units) layer, 120 persons in the training set. **(A)** Person identity error on the identity layer. **(B)** Person identity error on the parts layer. **(C)** Gender error on the parts layer. On the identity layer, the error rate drops to zero. On the parts layer, the more modules are involved in identity or gender estimation, the less is the error rate. For 6 modules, both the identity error and gender error drop close to zero.

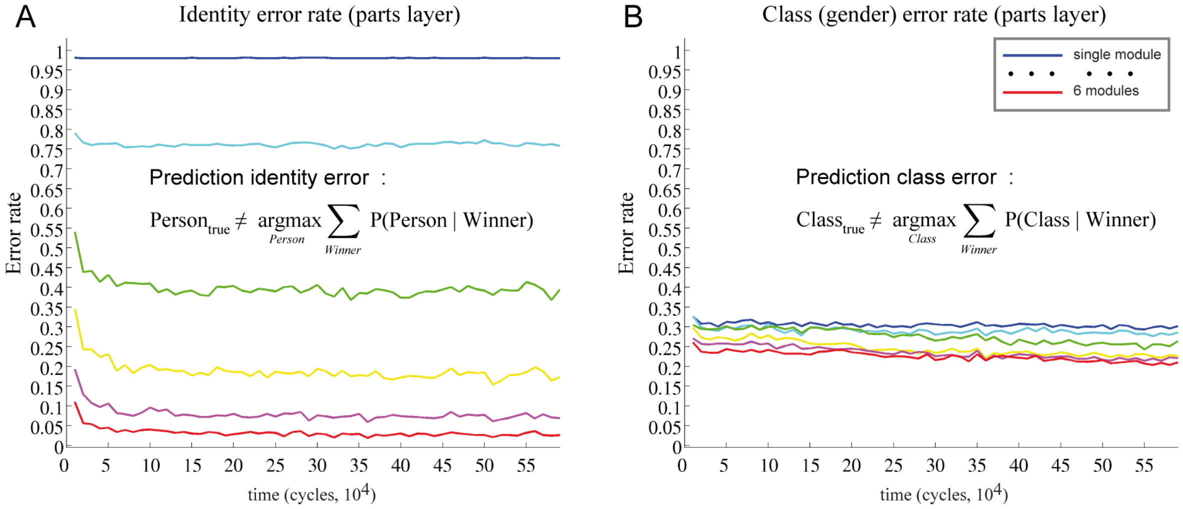


Figure 3.34: Learning error for the parts layer ($M = 6$ vocabulary modules, $N = 20$ units each), 1000 persons in the training set. **(A)** Identity error. **(B)** Gender error. Although being much lower than the chance level (50%), the gender error rate gets significantly larger (about 20%) than in case of learning 120 persons. The clusters of local appearance seems to be less consistently dominated by only male or only female features (see also Fig. 3.32 **(E)**), which consequently leads to more errors in gender classification.

Error type	Views, Error Rate (after learning 1000 persons)	
	Original	Emotion
Parts layer, Identity	$2.75\% \pm 0.22\%$	$70.63\% \pm 0.32\%$
Parts layer, Gender	$22.26\% \pm 0.41\%$	$24.69\% \pm 0.24\%$

Table 3.6: Generalization block error for the parts layer after learning 1000 persons. The identity error gets substantially larger for the alternative view.

affected by the increase of person load in case of learning 1000 persons, the error rate is still well beyond the chance level. This may point to a well-formed structure of the memory domain, which still manages to capture important characteristics of facial appearance despite the obvious content overload. In overall, the results suggest that expanding the memory network for more landmarks would provide a suitable substrate for storing a substantial number of compositional face identities in the range matching the real-world application scenarios.

3.5 Discussion

In this chapter, I addressed the question how a hierarchical memory domain for compositional representation of complex natural objects may arise in the visual cortex by unsupervised learning from sensory experience. To provide a functional answer to this question, a self-organizing cortical network architecture was proposed and implemented. The network was built of layers of distributed, recurrently interconnected cortical modules, making use of the neuronal model of unsupervised competitive learning elaborated in the previous chapter. Employing a decision cycle in the gamma range as an atomic fragment of ongoing competitive processing and learning, the network was able to form sparse memory traces for the incrementally presented natural face images of different persons. On the lower network layer, the vocabularies of reusable local face appearance elements were developed. These vocabularies were put into a global context by explicitly linking via associative lateral connectivity those appearance elements the memorized faces were composed of. On the higher layer, the higher-order symbols for face identities emerged. These identity units projected the compositional information about their constituent parts back to the vocabulary layer in generative fashion via top-down connectivity. The established memory traces held thus each a full compositional, generative description of a stored face. In its mature connectivity state, the system was able to recognize identity and gender of the persons from alternative face views not shown before. The network turned out to possess a variety of functionally relevant processing properties, such as the ability to sustain the stimulus-induced activity, to perform generative pattern completion, to induce bottom-up and top-down attention and to self-generate spontaneous activity, replaying the memory content in absence of the external input.

Generic memory architecture. The organization of the proposed memory system model supports the idea of universal cortical operations involving strong competitive and cooperative effects [von der Malsburg and Singer, 1988, Douglas et al., 1995], which are building up on essentially the same local circuitry and the same plasticity mechanisms utilized in different cortical areas [Mountcastle, 1997, Phillips and Singer, 1997, Douglas and Martin, 2004]. The implemented network architecture employed cortical modules developed in the previous chapter as elementary computational nodes. As already pointed out, each module may stand for a local cortical cluster of neuronal populations responding to similar features, shapes or objects [Fujita et al., 1992, Tanaka, 2003, Sato et al., 2009a].

The classical term "column" was avoided here intentionally, as the arrangement of the neuronal populations within the cluster can have a more complicated layout than a simple vertical orientation towards the cortical surface [Song et al., 2005, Yoshimura et al., 2005, Haider and McCormick, 2009]. Still, the functional nature of the module as a container for vocabulary of related similar elements holds here. Modules of this kind are found throughout the whole processing hierarchy in the cortex, independent of the modality [Mountcastle, 1997, Buxhoeveden and Casanova, 2002, Jones, 2000, Mountcastle, 2003, Lund et al., 2003, Rockland and Ichinohe, 2004]. When thinking about the processes underlying the memory formation and function, it becomes suggestive to hypothesize that these modules carry out the elementary generic operations necessary for maintaining those mnemonic processes. In the model network, these local operations amount to strong competition between the units within the module, bidirectional modification of the synapses the units receive from outside and homeostatic regulation of the unit activity.

The local competitive operation has been assumed to be a canonical operation performed in the cortical microcircuits in a vast number of previous works [Douglas et al., 1995, Hahnloser et al., 2000, Xie et al., 2002, Lücke, 2005, Spratling, 2008, Kouh and Poggio, 2008]. If the selection and amplification of a small fraction of neuronal resources at the expense of the rest indeed occurs within the modules along the processing hierarchy, an obvious analogy between the notion of the memory trace in the model network and the hypothetical memory trace in the cortex appears. Drawing upon this analogy, the cortical memory trace for a presented stimulus would consist of a sparse winner unit assembly formed during the signal propagation through the processing hierarchy and distributed all across its stages. This transient memory trace would serve as a basis for learning the compositional object identity, relying on the vocabularies of universal, reusable elements already established on the different hierarchical levels.

In the model network, learning happens simultaneously at all levels of the hierarchy. The same could be hypothesized for the cortical memory formation. The difference is of course that the universal vocabularies of low-level visual primitives are most probably acquired very early in the cortical development. The large part of connectivity modification would thus presumably occur in higher visual areas like IT, although accordingly specific task requirements (e.g., perceptual learning involving discrimination between low-level features) may also induce synaptic changes even in the early visual areas like V1 [Ahissar and Hochstein, 2004, Gilbert et al., 2009]. This is exactly the advantage of combinatorial flexibility given by such an universal vocabulary, which allows any novel complex object to be immediately instantiated and encoded into memory as a hierarchical composition of preexisting reusable elements of much lower complexity.

Learning the generative compositional object representation. The inherent compositional nature of natural visual scenes and objects makes their processing and learning tractable, as the full complexity can be reduced to the treatment of reoccurring elements and their relations. However, to benefit from this compositional nature, the memory system has first to acquire the ability to decompose the objects efficiently into re-usable elements of much lower complexity. This poses again a very hard learning task.

The network model implemented here solves simultaneously a number of core issues raised by this problem. First, it is able to establish vocabularies of local appearance elements that can be reused by multiple face objects stored in the memory. An arbitrary face can be thus represented in combinatorial fashion by picking best fit candidates from each local appearance vocabulary. Second, it is able at the same time to capture the compositional identity of individual faces explicitly in lateral and top-down connectivity, linking part and identity units into assemblies that correspond to memory traces of memorized faces. In such a network, the processing involves global cooperation within the assemblies and global competition between them, replacing the purely local competition between the units within the modules. The distributed decision making about faces becomes thus orchestrated by contextual support formed in the course of previous experience with visual objects.

Furthermore, it is possible to use the network in the recognition and the generative mode. In recognition mode, a very fast inference about a face presented on the input is possible within a single gamma cycle. A complex face object is immediately encoded as composition of few reusable part units and a corresponding identity unit. This sparse decomposition allows to reduce the learning of a full appearance to learning of the relations between few reusable elements of lower complexity taken from the established vocabularies. This makes processing and learning of novel faces extremely efficient, as a memory trace for a novel face can be created instantly, without need to re-build the vocabularies or insert new physical units that would have to be responsible for the new face.

In generative mode, the network is able to generate samples of previously memorized faces in form of activity corresponding to their full compositional representation, either by partial priming of corresponding memory traces or just by spontaneous memory replay occurring in absence of any external stimulation. From perspective of Bayesian generative learning [Lee and Mumford, 2003, Ulusoy and Bishop, 2005, Ommer and Buhmann, 2006], this functionality can be interpreted as capability to capture the face image generated by the global cause (face identity) as composition of many local causes (local appearance elements) including explicit representation of their interrelations. None of the recent neuronal modeling studies on unsupervised learning of natural visual objects were able to solve the problem of object representation in this explicit form [Serre et al., 2007c, Plebe and Domenella, 2007, Waydo and Koch, 2008, Wallis et al., 2008].

In order to make learning of global face appearance tractable in the network, the full input space of face objects was subdivided into lower dimensional subspaces, each corresponding to the local appearance space of a facial landmark. From the perspective of estimating the probability density for the face objects space, this partitioning can be interpreted as decomposition of the full density for faces into combination of densities for different local appearance subspaces. Importantly, this is not akin to factorization of the full distribution into independent components, as the local appearance subspace elements are linked via associative lateral and compositional top-down connectivity in the network and thus are explicitly dependent. This is a crucial difference to other approaches of density estimation by decomposition into subspaces, like for instance independent subspace analysis (ISA), where subspaces are assumed to be independent [Hyvärinen and Hoyer, 2000, Theis, 2007]. Here, the dependencies between the elements of different vocabularies are the essential basis for the memory traces. These dependencies define unit assemblies that represent compositional face identities stored in the network, rendering some combinations of units more probable than others. Furthermore, the number of elements within vocabularies does not have to be the same in the network, which means that the subspaces can have basis of different size. The subspaces may also have different dimensions, which is not the case for the standard ISA approach.

Competition and cooperation in structure and activity formation. For the learning to start at all, it is essential to have mechanisms for selecting a fraction of available neuronal resources that become clearly responsible for a current stimulus even in absence of differentiated, pre-established structure. So, learning prerequisites selection capability. This may sound like a circular paradox, as selection in turn appears to require some form of preceding learning. The strong local competition between the units within modules is exactly the right selection mechanism to resolve this circularity.

Of course, the selection via competition is purely random given the initial state of undifferentiated network structure, picking an arbitrary subset of winner units by enforced symmetry breaking in response to a presented face. However, these enforced response patterns offer sufficient playground for the learning to ignite. The high sparseness and amplification of activity produced by the network within a decision cycle makes it easy for synaptic plasticity to put previously random unit assembly in clear relation to the appearance of the presented face. Over repetitive stimulation, this correlation gets stronger and competition is not arbitrary anymore, but becomes more and more guided by the established synaptic connectivity.

This is a critical moment in memory formation, as many conflicting memory traces may strive for the right to be amplified. Misbehaving competition would be dangerous there, as it could create few strong or too many weak memory traces. Two adaptive mechanisms acting on the slow time scale in the network counteract these unfavorable tendencies. The emergence of strongly over- oder under-utilized traces is prevented due to the homeostatic activity regulation [Desai et al., 1999, Daoudal and Debanne, 2003, Maffei and Turrigiano, 2008], which reassures equal participation of units on memory formation. The segregation of the strongly conflicting traces is supported by the bidirectional nature of synaptic plasticity [Lisman, 1989, Artola et al., 1990, Bear, 1996], which favors synaptic potentiation between more coherently co-activated units while promoting depression between less coherently co-activated ones. The synaptic and unit competition thus enhance each other in the process of memory formation.

Crucially, local competition is not the only force shaping the memory formation. As lateral and top-down connectivity becomes more mature, local competition gets progressively accompanied by global cooperation and competition between the units across the modules mediated via lateral and top-down signaling. Opposed to the independent local decision making in the early learning phase, the decision making in the advanced learning phase is highly coordinated. In this phase, the interpretation of face stimuli takes into account both local and global context. The units that form the winner assembly and encode the presented face are those with a high degree of cooperation between each other. This cooperation is possible due to the existing coherent synaptic coupling between the assembly units. Established by the previous experience with the face stimuli, this contextual connectivity mediates signal exchange within the assembly and links locally ambiguous sensory cues to globally coherent interpretation of a presented face. This cooperation within the assembly is characteristic for the successful memory recall, being a signature for strong agreement between sensory and contextual cues and thus for successful recognition of a memorized face.

Two properties of the network were essential for proper signal exchange mediating cooperation between the units. First, the two-phase WTA coding scheme offers an opportunity to communicate clearly graded internal activity states of the modules across the network in a sufficiently large time window during the decision cycle. Second, separation of synapses in functionally different groups according to their hierarchical origin enables correct treatment of the incoming signals conveying different cues [Phillips and Singer, 1997, Friston, 2002, Larkum et al., 2009]. Remarkably, the signature of a failed recall is then the cooperation breakdown, which can be measured locally as disagreement between the bottom-up, lateral and top-down signals incoming from different levels of network hierarchy. This disagreement opens a possibility to compute a prediction error signal. This error signal could be used in perspective to modulate local learning in similar way as it is done in predictive coding scheme [Rao and Ballard, 1999]. For instance, the local plasticity could be deactivated in this manner if the local activity is already explained by the signals from higher stages, indicating that there is nothing more to learn.

The state of undifferentiated structure is however the worst-case scenario and not necessarily the initial condition for experience-driven learning in the cortex. Already in infants there may exist basis structures prepared for the acquisition of behaviorally highly relevant patterns, like for instance faces [Johnson et al., 1991]. It was hypothesized in numerous previous works that cooperation and competition phenomena driven by the environmental and intrinsically generated signals are involved in formation of cortical memory domain starting from this state [Willshaw and von der Malsburg, 1976, Changeux and Danchin, 1976, Pearson et al., 1987, Edelman, 1993]. Here, these phenomena could be tracked down in the network model to the level of unit-to-unit interactions and synaptic modifications, making more clear how selection and amplification may work to organize memory traces. The progress from a largely undifferentiated to a highly organized state via selection and amplification of only small subset of totally available alternatives is a general feature in evolutionary and ontogenetic development of biological organisms [Varela et al., 1974, Eigen and Schuster, 1977, Kauffman, 1993, von der Malsburg, 1995a]. The notion that the very same principles may guide activity and structure formation in the brain supports the view of learning as an optimization procedure adapting the nervous structure to the demands put on it by the environment [von der Malsburg and Singer, 1988, Edelman, 1993].

Gamma cycle as an atomic discrete fragment of ongoing processing. As discussed in the previous chapter (see Sec. 2.6), the distributed modules composing the network employ the gamma rhythm to make decisions about the incoming afferent inputs. Each cycle of the gamma rhythm constitutes a fragment of ongoing processing where competitive computation selects only few units to become active and represent the current input. In the network, this cycle is synchronized across the distributed modules. The synchronization is taken here for granted. In the cortex, there is mounting evidence that slow and fast ongoing oscillatory rhythms may synchronize their phase over long distance in response to demands of the current cognitive task [Gray et al., 1992, Rodriguez et al., 1999, Axmacher et al., 2006, Womelsdorf et al., 2007, Sauseng et al., 2008, Gregoriou et al., 2009]. The task-specific synchronization of distributed neuronal populations is thought to be the outcome of a self-organization process [Fries et al., 2007, Maldonado et al., 2008]. How this self-organization comes across is an important and challenging question, which has yet to be resolved and is a subject of further studies.

The functional role the decision cycle plays in the memory network model helps to elucidate the putative discrete nature of cortical processing hypothesized in a number of previous studies on perception of visual and auditive stimuli and on selective attention [Pöppel, 1997, VanRullen and Koch, 2003, VanRullen et al., 2005, Luo and Poeppel, 2007, VanRullen et al., 2007, Busch et al., 2009]. This view states that ongoing cortical processing may be composed of discrete successive cycles executed on the fast time scale of neuronal oscillations (10 to $200ms$), without elaborating much on the details of computation carried out in these discrete fragments. The decision cycle employed in the network defines a common reference time window for distributed cooperative-competitive decision making. It orchestrates not only activities, but also synaptic plasticity across the units. Sparse activity pattern generated within the cycle by WTA-like operation defines unambiguous temporal correlations between the winner assembly units, solving the binding problem by competition. The competition permits only one non-ambiguous interpretation of the current stimulus situation to appear within the perceptual fragment defined by the decision cycle. A decision made within one single cycle can be interpreted as an initial coarse hypothesis about the incoming face stimulus. Thus, memory recall in the network is possible within one short cycle of ongoing rhythm, which is in line with other modeling and experimental studies [de Almeida et al., 2007, Serre et al., 2007b, Thorpe et al., 1996, Fabre-Thorpe et al., 2001, VanRullen, 2007, Stanford et al., 2010]. However, this rapid processing does not have the purely feed-forward sweep style here, as it is biased in advance by the existing lateral and top-down connectivity.

Sparse activity pattern generated in a single cycle can be carried over to the successive cycles, providing an opportunity to refine the initial hypothesis by more elaborate recurrent processing. This refinement would then massively involve lateral and top-down connectivity, which has not the sufficient time to unfold its full influence within the short time of a single cycle. The hypothesis refinement could also take place upon a slower oscillation in the theta range. The slower theta rhythm could modulate the amplitude of the faster gamma rhythm on its top, which is indeed observed in cortical processing [Lisman and Idiart, 1995, Lisman, 2005, Jensen and Colgin, 2007, Lisman and Buzsáki, 2008, Axmacher et al., 2010] (see Fig. 3.35 for a simplified example without amplitude modulation). Such embedded oscillations could be a suitable mechanism to gradually increase the competitive pressure on progressive decision making, which in turn could correspond to the hypothetical transition from the pre-attentive first coarse hypothesis to the final revised hypothesis entering the awareness [Lamme and Roelfsema, 2000]. The same mechanism could be also used for memorizing item sequences, where each embedded gamma cycle would carry an item as a part of a longer sequence linked together within the slower theta rhythm [Jensen and Lisman, 1996, 2005, Luo and Poeppel, 2007].

Interestingly, whether activity generated within a cycle will influence decision making in the next one can be controlled by a simple tuning of the ongoing rhythms. In the one extreme, the stimulus-induced activity can be abolished, such that decision making in the next cycle is reset and has to start over again. In another extreme, the activity can be completely locked, so it persists in the successive cycles and cannot be abolished even if the external stimulus is removed or another stimulus is presented. This simple mechanism introduces necessary flexibility into discrete operating mode of ongoing processing, which has either to react to a novel incoming input or to keep the computed result in short-term working

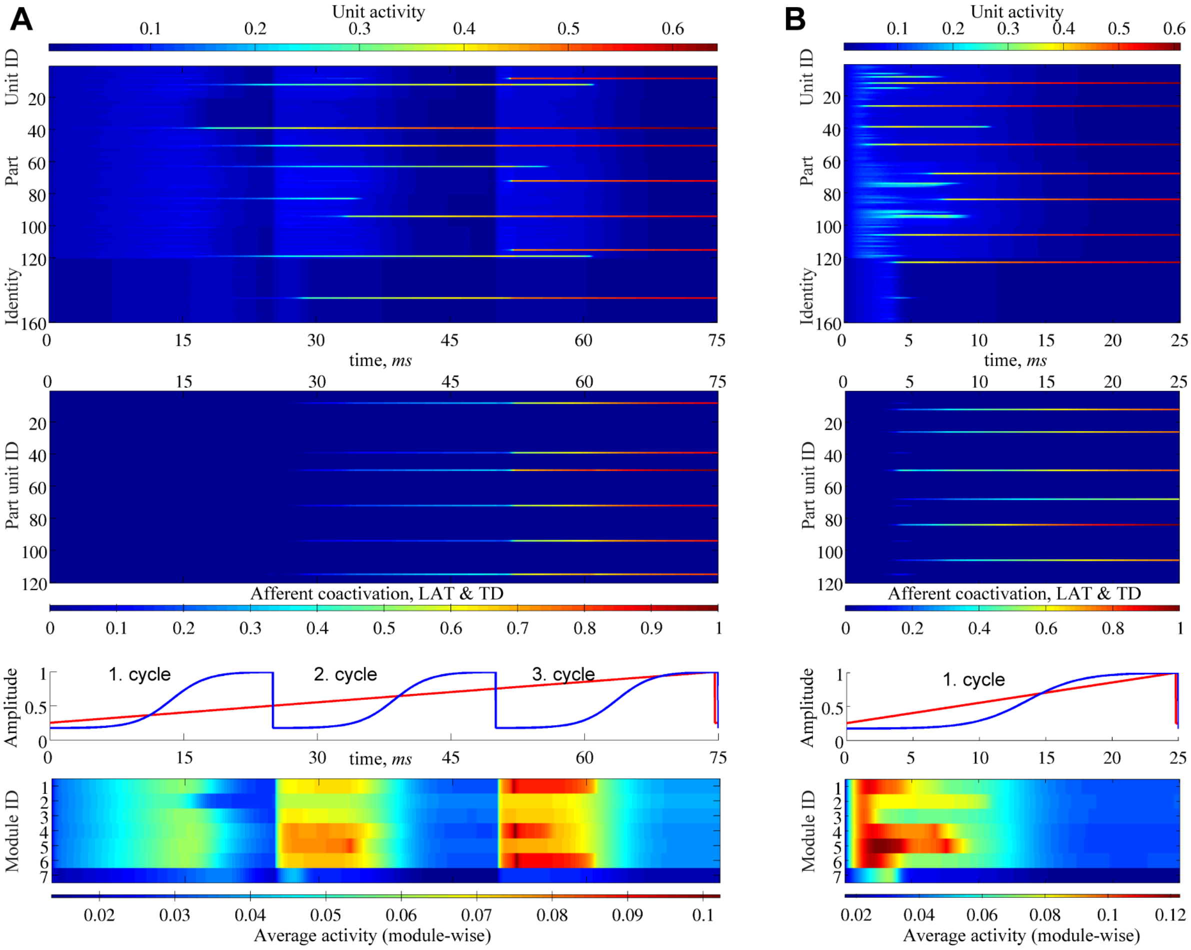


Figure 3.35: Embedding fast gamma rhythms (3 cycles, $25ms$ each) in slower theta rhythm (1 cycle, $75ms$). A simplified embedding example is shown in **(A)**, where the amplitude of the gamma rhythms is not modulated by the theta phase. The initial hypothesis created in the first cycle contains a number of wrong decisions about local facial appearance. This initial guess is refined in the successive cycles, improving local decisions and correcting them towards more coherent interpretation of the presented face, as indicated by a high final degree of coactivation of the afferents converging on the winner units. This refinement fails if only one decision cycle is given for the stimulus interpretation as shown in $(\mathbf{B})$. The final degree of afferent coactivation is lower there, indicating that local decisions do not match each other well and were thus less successful in achieving a coherent interpretation of the presented face.

memory, depending on the current task demand.

Highly sparse representation with option for localist coding. An essential property of the processing in the network is the sparseness of the activity generated in response to a face stimulus. This sparseness is the consequence of a hard WTA operation performed by the network modules, keeping only one unit per module active at the end of the decision cycle. As already discussed here before, sparse representations in the cortex are considered to support efficient coding [Barlow, 1961, Field, 1987, Simoncelli and Olshausen, 2001, Földiák, 2002, Olshausen and Field, 2004], rapid and robust learning [Hahnloser et al., 2002, Fiete et al., 2004, Asari et al., 2006], memory capacity [Palm, 1980, Tsodyks and Feigel'man, 1988, Buhmann et al., 1989, Amit and Treves, 1989, Rolls and Treves, 1990, Okada, 1996, Sommer and Palm, 1999, Rehn and Sommer, 2007] and metabolic efficiency [Levy and Baxter, 1996, Laughlin, 2001, Vincent et al., 2005]. In the network model, the sparse activity patterns result in sparse connectivity for memory traces formed during the learning, reducing interference effects between individual face representations stored in the memory. Furthermore, the ultra-sparse activity makes the read out of the network state extremely easy. A very simple procedure based on the history of unit win events can be thus employed to decode the identity or gender of a presented face.

By introducing the hierarchy, higher order symbols for person identity were obtained in form of the identity units on the higher layer. These units are used for a compact representation of individual faces shown during the learning phase, without discarding the information about their composition which is kept in the top-down connections projecting back to the lower vocabulary layer. The explicit preserving of compositional information is a crucial property of the generative representation developed by the network. Although the identity units have a highly localist appeal [Barlow, 1972, Bowers, 2009], being extremely specific in their response selectivity, their meaning is rooted in the part units distributed on the upstream layer. Thus, an identity unit is never activated in isolation, but is always accompanied by the activation of the part units that contribute to its specificity. This is of course a trivial statement for a stimulus-induced activation, but it is not trivial in case of evoking the same unit assembly in generative fashion by a top-down cue or by priming of only a subset of the part units that belong to the full assembly. In both cases, the network is able to generate the full compositional description of the memorized face identity by reactivating the whole assembly, comprising the part and the identity units. In terms of its highly specific selectivity, the identity unit can be indeed considered as analogous to a grandmother neuron (or a grandmother neuron population). However, this statement makes little sense in terms of the employed coding scheme, as the identity unit can never be observed being active all alone due to the existing discriminative and generative connectivity. Its meaning is conveyed not simply by its own state of activity, but also by the activity states of all other units the active identity unit directly causes.

The identity module demonstrates the potential to develop the localist kind of representation on the higher level of hierarchy. The dispute about the existence of such representation in the brain is highly controversial [Földiák, 2009, Bowers, 2009, 2010, Plaut and McClelland, 2010, Quiroga and Kreiman, 2010]. Single neurons that are highly specialized for a specific object or an individual face were observed experimentally in higher cortical areas like IT, PFC and MTL [Freedman et al., 2003, Quiroga et al., 2005, 2008, Bowers, 2009]. One can hypothesize that such units may emerge on the top of the processing hierarchy to represent very important things (VITs) or very important persons (VIPs). Those are objects or persons the processing is extensively dealing with on day-to-day basis, like a favorite teddy bear or a close friend. The high selectivity of such units does not necessarily mean though that they would not respond to any other stimulus, nor it means that a particular object activates only a single unit in the higher cortical regions. The only definite finding of the experimental studies so far is that the evoked activity pattern is sparse in both lifetime and in population sense, involving only a small fraction of available units to code for a given stimulus [Weliky et al., 2003, Quiroga et al., 2005, 2008]. This is also the property of the coding scheme employed by the model network. In principle, the system's functionality as associative memory doesn't need to rely on a localist representational scheme of the higher identity layer. It could use exclusively the sparse coding, utilized on the lower vocabulary layer, as well to represent identity and gender of presented faces. However, the compact generative representation offered by the identity units is also advantageous, as it allows to reactivate the full parts-based description of the memorized identity by simply reactivating one single dedicated unit. This is not possible for single part units, as they are always involved in coding for different memorized faces.

In general, the modules in the network are not stuck to coding via the strict hard winner-take-all operation. If a face is fixed on the input over multiple cycles, the network behavior resembles transient attractor dynamics [Friston, 1997, Rabinovich et al., 2008, Durstewitz and Deco, 2008], where different sparse assemblies get activated in a sequence. This is a completely different coding scheme with a much richer expression power, which is however not easy to read out anymore, it was not exploited here further. Potentially, it would be also possible to select multiple candidates from a single module to encode local appearance of a face in a single cycle. Here, I used very strong competition leading to a form of activity sparseness termed hard sparseness [Rehn and Sommer, 2007], limiting the number of active units to one per module. While this kind of sparse coding is advantageous for learning individual faces, it may be generally too sparse for representing coarser categories (like young, old, etc). However, the competition strength can in principle be adjusted arbitrarily in a task-dependent manner. This can be easily implemented by tuning the amplitude of the ongoing rhythms. In the brain, such tuning could be initiated by some kind of internal cortical signal or state, indicating the task-dependent need for the competition strength. The tuning of the competition strength would allow the formation of less sparse activity distributions, representing the stimulus on a coarser categorical level [Tanaka, 2003, Kim et al., 2008].

Attentional and generative mechanisms in the memory. The generative character of lateral and top-down connectivity and competitive nature of local computation endow the system with further general inherent capabilities. For instance, selective object-based top-down attention is naturally given in the model, because the priming of the identity units leads to facilitation of the corresponding part units via top-down and lateral connections. Such facilitation provides clear advantage in the competition against other alternatives during the memory recall. This priming can mediate covert attention directed to a specific object, promoting the pop out of its stored full compositional representation while suppressing the rest of the memory content. In the same fashion, priming a subset of part units has the effect of bottom-up parts-based attention mediated via lateral and top-down connectivity. Generally, the selection and amplification by competition can be interpreted as a universal attentional mechanism, which selectively focuses the neural resources on processing one object or category at a time by actively suppressing the rest [Bundesen, 1990, Tsotsos et al., 1995, Lee et al., 1999, Reynolds et al., 1999].

Remarkably, the reactivation of the full object representation is also possible in absence of any external stimuli or additional priming. The reactivation occurs in form of spontaneously self-generated activity emerging within the fragments of the ongoing decision cycles. Interestingly, spontaneously generated activity patterns that seem to reflect the internal memory structure can be also observed in the neocortex and hippocampus in the absence of intentional sensory input during restful waking or sleep [Kenet et al., 2003, Euston et al., 2007, Ji and Wilson, 2007, Pastalkova et al., 2008]. This memory replay can be interpreted from two different perspectives. Looking at the memory network as dynamical system with attractor states defined by the connectivity, the replay corresponds to visiting the attractor states transiently, offering the way to browse through the memory content in an ongoing manner. Relating the established memory structure to a generative model for face objects learned from the experience, the replay can be understood as ongoing sampling from this generative model that provides full description how each memorized face identity is composed from a number of local causes, i.e. its constituent local appearance elements. This puts the model in interesting relation to generative approaches explaining construction of explicit data representation in machine learning [Ulusoy and Bishop, 2005, Ommer and Buhmann, 2006]. Both perspectives suggest that this reprocessing procedure can be potentially useful for maintenance and reorganization of the memory domain in an off-line mode, resembling sleep or restful waking, as will be discussed in the next chapter.

Functional advantage over the purely feed-forward architecture. In neuronal modeling of visual object recognition, the feed-forward processing style still has strong dominant influence. This is due to many influential works that were able to demonstrate the ability to perform object recognition very rapidly in a single feed-forward sweep [Riesenhuber and Poggio, 1999, Thorpe et al., 2001, Serre et al., 2007b], achieving times comparable to what is known from experimental studies with higher primates and humans [Thorpe et al., 1996, VanRullen and Thorpe, 2001, Joubert et al., 2007].

Most of researchers advocating the idea of a rapid feed-forward sweep agree now that this can be only a model for a very early fragment of recognition process (below $150ms$ from the stimulus onset), which has to be refined subsequently by massive recurrent processing involving feedback connectivity [Lamme and Roelfsema, 2000, Delorme et al., 2004, Serre et al., 2007a]. In the implemented network model, the memory recall and recognition can be performed within one single short cycle corresponding roughly to $25ms$ of biological time. Already in this small fragment of recognition process, the fully recurrent network configuration were able to outperform the purely feed-forward opponent in two scenarios. The significant difference in identity and gender recognition performance in favor of fully recurrent architecture was observed on the alternative views that deviated stronger from the original views shown during the learning. This difference between the two configurations was even larger after excitability regularization procedure that improved the recognition performance of both configurations (elaborated discussion of this particular phenomenon will be postponed until the next chapter). These results suggest that the fully recurrent network is able to generalize over new data much better than the purely feed-forward configuration, which in turn performed well on original or only slightly deviating views.

At least two points are worth mentioning in face of these findings. First, the outcome indicates that these different processing strategies may prove more or less useful depending on a given situation. The feed-forward processing may already suffice to do a good and quick job when facing well-known, overlearned situations, where effortful disambiguation of locally available information is not required due to the strong familiarity of the sensory input (like it is here the case for local face appearance of the original views shown during the learning). However, the heavy dependence on feed-forward connectivity is punished as soon as novel data arrives and this strong familiarity is not given anymore. In this situation, locally available sensory information about the global stimulus is highly ambiguous and cannot be interpreted correctly anymore without additional support of contextual cues mediated by lateral and top-down connectivity. Thus, the recurrent processing is mostly beneficial in novel situations, which require effortful disambiguation of otherwise not clearly interpretable local information provided by less familiar or unknown stimuli. This is the case the organism gets confronted with rather often in the real world. In such setting, the recurrent processing becomes most probably an indispensable part of optimal perception and decision making, while the ultra-rapid feed-forward sweep can serve as the very first processing stage providing a fast, but coarse and "dirty" initial interpretation of the incoming stimuli [Lamme and Roelfsema, 2000, VanRullen, 2007].

Second, one may ask whether it is conceivable to expect the beneficial influence of recurrent intra- and interlayer connectivity on decision making in such a short time, if the time of signal propagation is taken into account. For two successive stages of processing, it is realistic to assume that signal exchange involving bottom-up and top-down connectivity can be accomplished within one cycle period of $25ms$. For the full processing hierarchy in the visual cortex consisting of about 6 stages, this would then amount to the total time needed to incorporate first recurrent feedback of about $150ms$. Another important fact is the existence of the ongoing activity in the cortex, which most probably bias and essentially modulate the perception of the incoming stimuli according to the object knowledge gained by the previous experience [Tsodyks et al., 1999, Kenet et al., 2003, Fiser et al., 2004]. This ongoing activity is surely heavily shaped by the recurrent connectivity, which is thus able to influence the interpretation of the incoming stimuli even before the stimulus onset.

Model predictions. A number of direct predictions can be derived from the system's behavior. Concerning the difference between the two network configurations, the hypothetical benefit of recurrent architecture for generalization capability could be tested in a behavioral experiment. Subjects should be instructed to learn to discriminate persons from a number of face images, for example by assigning each of them an arbitrary chosen name or index. After learning, the same discrimination task were to perform on the original and alternative face views of the same persons. Disrupting during the task the recurrent interareal connectivity in the IT, which can be done for instance by transcranial magnetic stimulation (TMS) [Cowey and Walsh, 2001, Lamme, 2006], should then have no substantial effect on recognition rates for original views if compared to the normal condition. On the contrary, for the alternative views the disruption of the recurrent connectivity should lead to a severe performance drop compared to the normal condition.

Another general prediction is that a failed memory recall should be accompanied by higher overall activation in the IT during the gamma or theta cycle, with the activity level of the remaining active spots at the end of the cycle being on contrary diminished. Reversely, a successful recall should be characterized by decreased overall activity in the IT and by increased activity in the remaining active spots at the cycle's end. This is also interpretable in terms of signaling the degree of decision certainty, the successful recall being accompanied by greater certainty about the recognition result. Further, a failed recall should induce much more depression (LTD) than potentiation (LTP), a successful recall much more LTP than LTD on the active synapses. In addition, if required to memorize and distinguish very similar stimuli, the recall of such an item should lead to a higher overall activation in the IT network than for items with less similar appearance. The active spots at cycle's end, on contrary, should exhibit a reduced activation due to the inhibition originating from the competing similar content. Again, interpretation of the activity level as degree of certainty is possible here: the more similar the stimuli to be discriminated, the lower is the winner activation signaling the decision made, indicating lower certainty about the recognition result. An interesting prediction concerning the bidirectional plasticity mechanism is the erasure of a memory trace after repetitive stimulus-induced recall if LTD/LTP transition threshold is shifted to the higher values, for example due to an artificial manipulation. This was indeed performed and observed in experiments of selective memory erasure in mice [Cao et al., 2008].

Innovations compared to existing approaches. The network model introduced here is able to build up a hierarchical memory domain for the persons shown during the learning in completely unsupervised fashion by just being exposed to the data. In contrast, many approaches to object and face recognition still do require heavy supervision, where at least the identity labels have to be provided together with the object images in the training set [Lawrence et al., 1997, Rowley et al., 1998, Dailey and Cottrell, 1999, Riesenhuber and Poggio, 2000, Viola and Jones, 2001, Garcia and Delakis, 2004, Murray and Kreutz-Delgado, 2007, Tong et al., 2008]. On the other hand, systems that can learn in largely unsupervised fashion from natural images are able to discriminate only between highly distinct objects, like for example planes vs. faces vs. horses [Ommer and Buhmann, 2006, Epshtein et al., 2008, Ranzato, 2009]. Unsupervised learning of individual differences on subordinate level, which is done here by memorizing compositional identity of different persons in the network, turns out to be extremely difficult especially for the object classes with small intraclass variance (like the object class of faces), where similarity between the class members is inherently high. The competitive and cooperative effects instantiated in the network endowed it with the capability to solve this task.

Another difficulty for the existing methods is often to employ a vocabulary of reusable elements that can be shared for representing multiple objects stored in the memory in combinatorial fashion. For instance, the previous approaches that used the columnar network architecture with hard-wired connectivity to perform face recognition required completely disjoint sets of units for storing each individual face [Jitsev, 2006, Wolfrum et al., 2008]. Here, such vocabularies are learned from experience with the face images.

Moreover, those methods that aim on learning the universal vocabularies rely usually on a sequential procedure, where learning is performed separately for each processing layer of the hierarchy. The modification there is always restricted to one processing layer while keeping the others fixed, proceeding sequentially from the bottom to the top of the hierarchy until all stages are done [Hinton et al., 2006, Serre et al., 2007b, Ranzato et al., 2008]. This is very hard to reconcile with ongoing plasticity processes observed in the cortex [Gilbert et al., 2009]. In the model network there is no need for such sequential procedure, as all layers can modify their representations simultaneously, developing in parallel full synaptic connectivity (bottom-up, lateral and top-down). Further, so far only very few approaches in machine learning show ability to learn the compositional nature of visual objects from natural images, capturing the parts and their relations explicitly in generative fashion to define the global object identity [Fergus et al., 2003, Ommer and Buhmann, 2006, Fei-Fei et al., 2006, Ommer and Buhmann, 2010]. These highly promising approaches rely on methods from Bayesian generative learning, where it is by far not clear how a neuronal plausible implementation would look like. The network model implemented here has demonstrated how such compositional representations could be created in a cortical memory architecture using neuronal plausible mechanisms of competitive learning. Finally, there is no need here for any time-dependent stop condition commonly defined by hand in many standard approaches to prevent destabilization of the learning procedure, as the learning in the network is self-stabilizing and thus life-long.

4

Autonomous off-line memory reprocessing in a sleep-like state and its functional consequences

While examining the ability of the model memory network to recognize persons from natural face images, a direct beneficial effect of homeostatic activity regulation on the memory performance was found and described in the previous chapter (Sec. 3.3). This effect does not depend on modification of individual synapses, as the observed improvement in recognition performance was achieved with disabled synaptic plasticity during the memory reprocessing. It was shown that the positive effect is mediated entirely by the mechanism of intrinsic plasticity of the units. The main origin of the positive effect turned out to be the regularization of excitability levels in the network.

One way to induce the improvement in recognition performance on alternative face views was to show repeatedly prepared blocks of novel data, causing homeostatic activity regulation to adjust the excitability levels of identity units towards a regularized state. While this presentation mode is not totally impossible in a natural learning scenario, this procedure cannot be considered as a technique to improve generalization, because the novel data has to be shown to the system in advance. The other way was to set the excitability levels to an equal value by hand. This in turn can be considered as a procedure improving generalization capability, as there is no need to present novel data in advance in this case. However, this technical shortcut is biologically not plausible. So, the question arises whether one could find a neuronal plausible way to make the regulatory process mediated by intrinsic plasticity run towards the state of regularized excitability levels, without falling back to manual procedures or showing the novel data to the network.

Another intriguing feature of the emerging memory architecture studied in the previous chapter was its ability to self-generate spontaneously activity patterns in absence of external stimuli, replaying the memory content in a sleep-like off-line regime (Sec. 3.2.4). In this regime, the units in the network get the opportunity to become active without bias imposed by external sensory input. What kind of outcome one would expect from this off-line memory reprocessing if the network spends a while in this "sleep" state decoupled from sensory inputs? If we disable synaptic plasticity, the connectivity structure would trivially stay the same. Keeping homeostatic activity regulation on would impose changes in excitability across the network units. How would the nature of these changes look like? In absence of external input, homeostatic activity regulation would be driven entirely by the intrinsically generated network activity. We would then expect that without intervening external stimulation it should become easier to downregulate those units that are initially more active and to upregulate those that are initially less active during the spontaneous replay. But this would provide us exactly the desired course of excitability tuning, namely the regularization of its levels across the units of the network. The expected outcome would be that after some time spent in the off-line state, the units would have almost uniform probability to become active during a cycle, their excitability levels converging close to some common value.

Motivated by these considerations, I'm going to probe in this chapter whether the off-line memory reprocessing via self-generated activity replay is indeed able to achieve the favorable network state of regularized excitability levels in absence of external sensory input. To test the actual effect of the off-line memory reprocessing on recognition performance and on generalization capability of the memory network, I will compare the error rates on alternative face views before and after the sleep-like mode. Furthermore, I will look for the potential differences between the purely feed-forward and fully recurrent network configuration in terms of the positive effect induced by the reprocessing in the off-line regime. Finally, I will discuss the relation of the observed phenomena to the hypothetical off-line memory reprocessing during the states of sleep or restful waking in the cortex and cognitive improvements observed afterwards.

4.1 Off-line memory reprocessing and generalization boost

4.1.1 Off-line regime setup and performance evaluation

The memory network which emerges after a prolonged on-line learning from natural face images (see Sec. 3.1.3, mature connectivity state taken after $4 \cdot 10^5$ cycles) is able to self-generate activity in the absence of the external stimuli (Sec. 3.2.4). It can be put in an off-line reprocessing mode by simply removing the images from the input. Consequently, the vocabulary modules on the lower memory layer do not receive any sensory stimulation over the bottom-up synapses in the off-line phase. The network continues then to run autonomously, reprocessing the memory content (Fig. 3.21). During this off-line phase, synaptic plasticity is disabled, the homeostatic activity regulation is active in the network.

To observe the changes in the excitability levels for a prolonged period, the off-line mode is run until no significant excitability changes can be detected anymore across the network units. The achieved distribution of excitability levels is then assessed. After the off-line reprocessing, the recognition performance of the network can be measured on alternative face views not shown before. The performance after the sleep-like state can be then compared to the performance of the original network before the off-line reprocessing by measuring the immediate generalization error (see Sec. 3.1.1) for both states. The evaluation is done for the parts and the identity layer separately. Again, the purely feed-forward network configuration and the fully recurrent network configuration can be compared in terms of the recognition performance after the off-line phase.

4.1.2 Results

Following the procedures described above, the memory network is put into the off-line mode after the prolonged on-line learning from natural face images. As already described in Sec. 3.2.4, the self-generated activity patterns tend to resemble stimulus-induced activity observed in the "wake" on-line

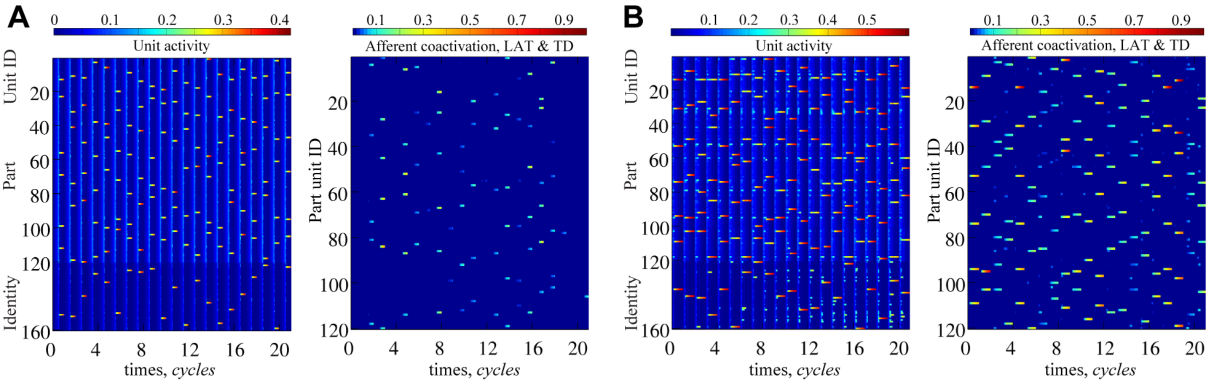


Figure 4.1: Activity formation in different off-line regimes. Depending on oscillation parameters, the self-generated activity can be more or less coherent, corresponding more or less to the previously memorized faces. In **(A)**, an outtake (20 cycles) from the standard off-line regime used here for memory reprocessing is shown (default oscillation parameters used), where the replay may contain both memorized and "phantasized" patterns. The afferent coactivation within the winner assemblies can be sometimes high and sometimes low there (see also Fig. 3.21). **(B)** shows an example of a "sleep" regime with stronger coherency of replayed activity patterns (oscillation tuning used, $\nu_{min} = 0.15$, $\omega_{min} = 0.5$). As indicated by the high level of afferent coactivation, the recalled assemblies correspond to the face representations already stored in the memory.

Oscillation Parameter	Sleep mode		
	Downregulation	Balanced	Upregulation
ω_{min}	0.25	0.1	0.1
ω_{max}	0.75	1.0	1.0
ν_{min}	0.005	0.15	0.25
ν_{max}	1.0	1.0	1.0

Table 4.1: Oscillation amplitudes for different "sleep" regimes, where the excitability levels get either downregulated, upregulated or stay roughly the same.

mode. The tendency to replay exactly the patterns that correspond to the memorized faces can be made weaker or stronger by adjusting the amplitude of the ongoing rhythms in the "sleep" state (Fig. 4.1).

First, let us concentrate on the changes in the excitability levels across the network caused by the homeostatic activity regulation, being the only adaptive mechanism active during the off-line mode. The time course of excitability change reveals clearly a global trend towards layerwise equalization of initially widespread unit excitability levels (Fig. 4.2). This trend is reflected in the decrease of standard deviation of intrinsic excitability on the both network layers, which drops close to zero within a short time spent in the off-line regime (around $6 \cdot 10^3$ cycles, see Fig. 4.2 **(B)**). The layerwise decrease in their standard deviation proves that excitability levels move closer together, approaching a certain common value. The flattening of the excitability landscape makes consequently the layerwise distribution of unit win events to become more uniform. This regularized state of excitability levels corresponds just precisely to the desired network state found to be beneficial for the memory function by previous experiments.

In addition to the flattening of excitability levels there is a particular direction the levels converge to in the network. In this particular case, the network excitability is downregulated. The direction of excitability regulation is determined by the amplitudes $\omega_{min}, \nu_{min}, \omega_{max}, \nu_{max}$ of the ongoing excitatory and inhibitory rhythms ω and ν. This is because these parameters influence how strong the units may become activated during the replay in the off-line phase. To get sure that the excitability regularization

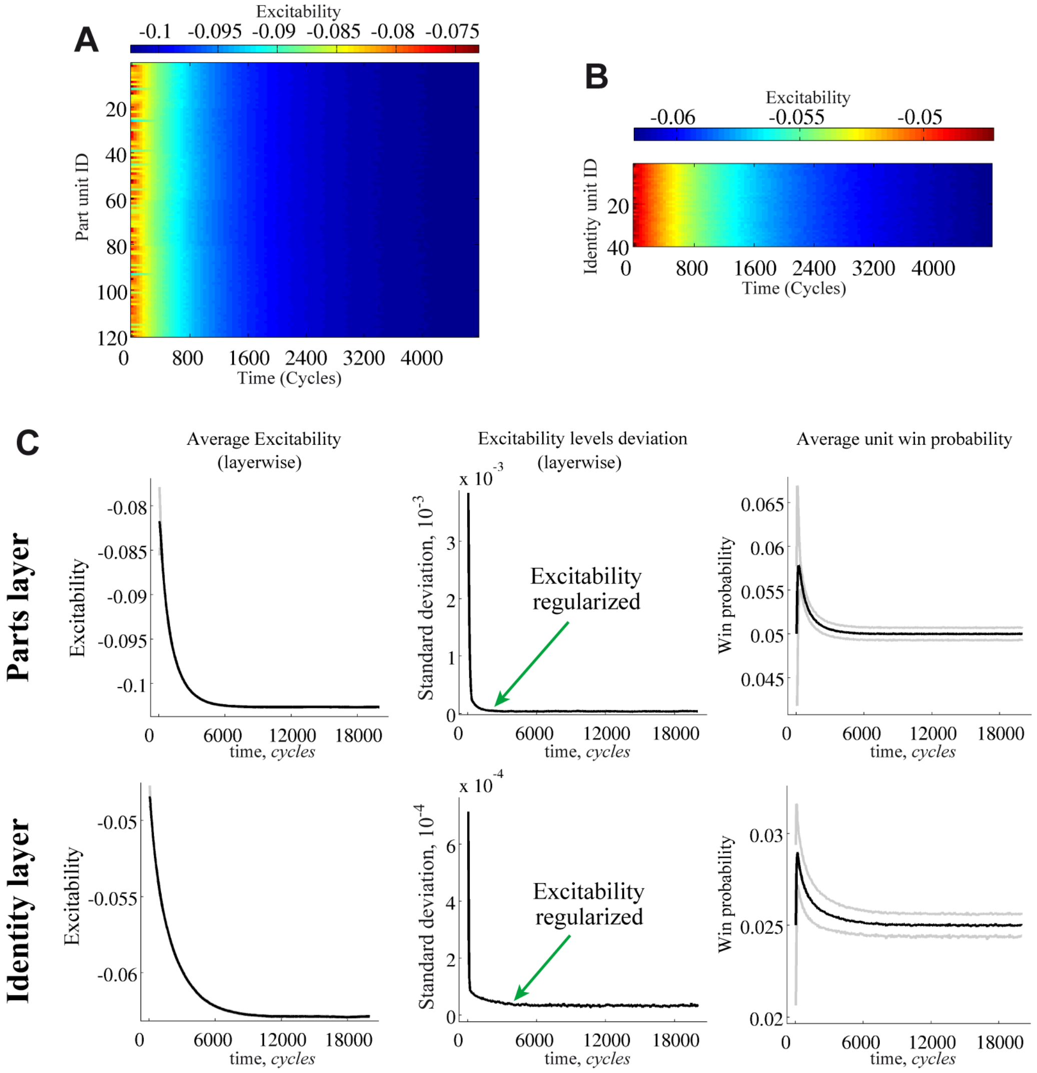


Figure 4.2: Excitability regulation during the sleep-like state (downregulating regime) on the lower vocabulary layer **(A)** and higher identity layer **(B)**. Excitability is downregulated on both layers. Excitability levels get regularized layerwise by moving closer together (standard deviation drops almost to zero on both layers). Regularization is accomplished already within a short time spent in the off-line regime (standard deviation settles down to the low value before $6 \cdot 10^3$ cycles have passed, which corresponds to biological time of about $150s$). The layerwise distribution of unit win events becomes consequently also more uniform.

is produced independent of the regulation direction, the oscillation amplitudes can be adjusted such that instead of excitability downregulation either upregulation or roughly balanced regulation of excitability occurs (Tab. 4.1). In both cases, the flattening of excitability levels is again observed. Noteworthy, if the regulation of excitability levels during the off-line regime has a balanced outcome (Fig. 4.3), where the levels converge to a common value close to the layerwise average excitability level of the original state, the units that were initially below the average excitability level get upregulated, and those above the average get downregulated during the off-line reprocessing. The excitability regularization phenomenon is also unchanged, if oscillations are tuned to increase the replay of coherent patterns already stored in the memory.

Back in the on-line mode (after $2 \cdot 10^4$ cycles), the system shows a tremendous boost of recognition

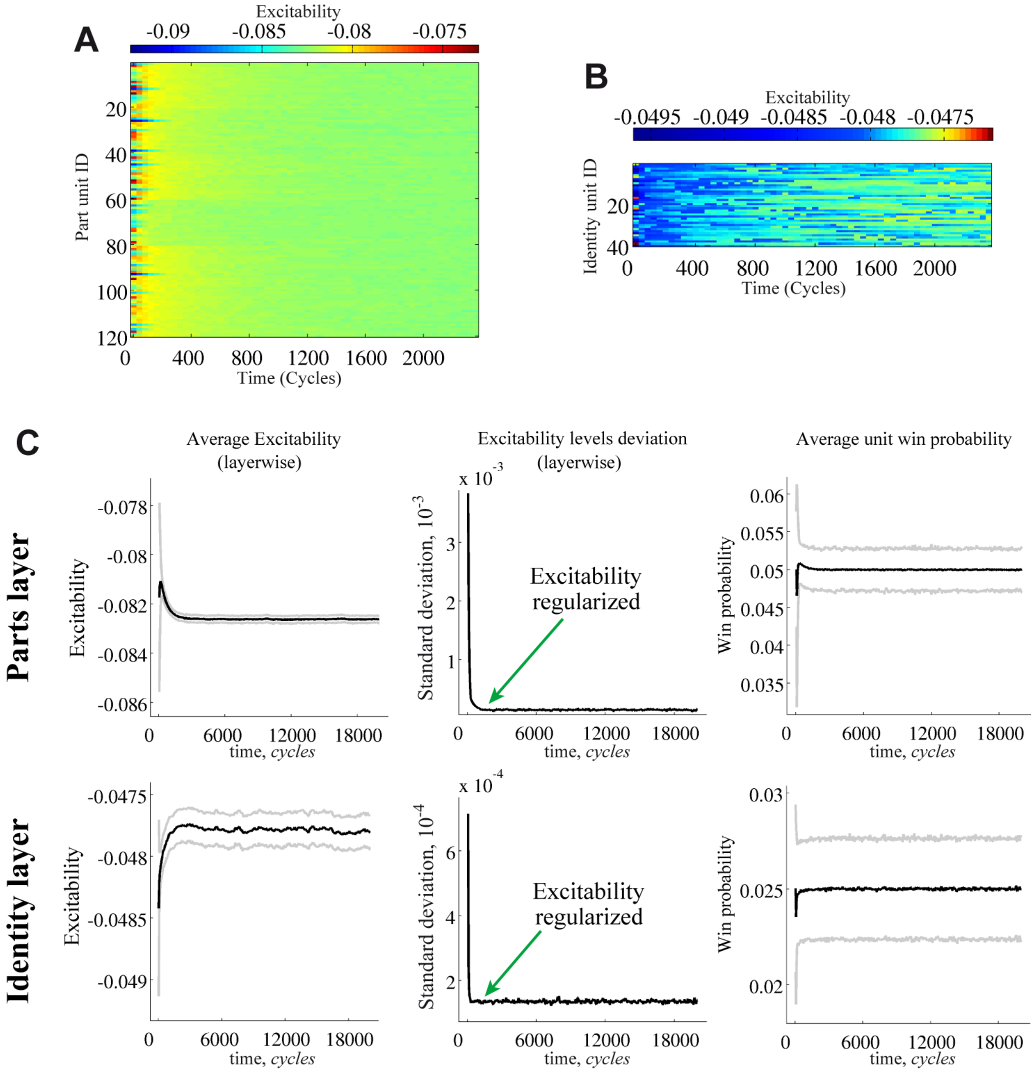


Figure 4.3: Excitability regulation during the sleep-like state (balanced regulation regime) on the lower vocabulary layer **(A)** and higher identity layer **(B)**. As in the previous case, excitability levels get regularized layerwise by moving closer together (standard deviation drops almost to zero on both layers) **(C)**. Regularization is achieved rapidly after about 2000 cycles (50s in biological time), the layerwise distribution of unit win events becomes more uniform. The excitability levels converge here to a common value in the vicinity of the layerwise average excitability level of the original state. Consequently, the units that were initially below the average excitability level get upregulated, and those above the average get downregulated during the off-line reprocessing (**(C)**, on the left).

performance compared to the state before the off-line reprocessing across all face views, the original and the alternative ones not shown before (Fig. 4.4). The strong drop in the identity error rate is observed for the identity layer as well as for the parts layer. Remarkably, the error drop is much more evident for the alternative views (up to 40%) than for the original views used during the learning.

The positive effect caused by the off-line reprocessing does not dependent on the particular direction the excitability regulation shifts the levels to, as shown in the Fig. 4.5. The recognition performance has no significant differences in the network states with downregulated, balanced or upregulated excitability levels after the sleep-like mode. This provides further evidence that the main cause of the observed

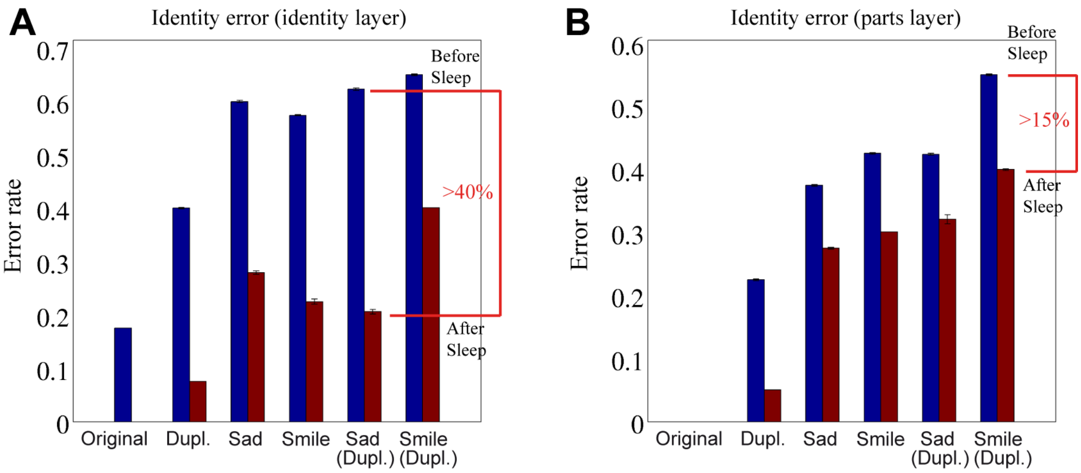


Figure 4.4: Comparison of recognition performance before and after the off-line memory reprocessing for the recurrent network configuration (downregulating regime used). Identity error rate for the identity **(A)** and the parts layer **(B)** is shown. The off-line reprocessing leads to a dramatic drop in error rate across all views, similar to the positive effect of the manual regularization procedure described in Sec. 3.3 (see also Fig. 3.25).

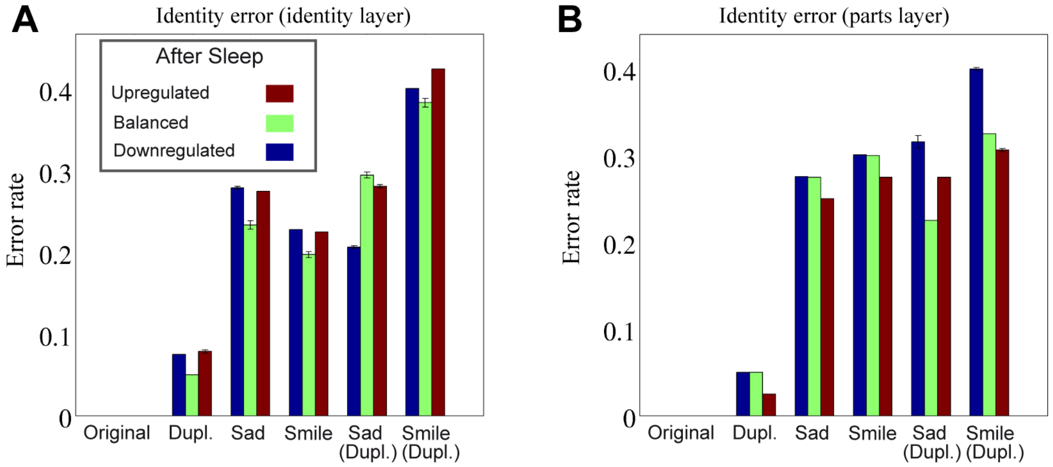


Figure 4.5: Effect of different off-line regimes on performance of recurrent network configuration. Identity error for the identity layer **(A)** and for the parts layer **(B)** is shown. The positive effect is qualitatively independent of the particular regularization direction, the performance is roughly the same in case of downregulation, balanced regulation or upregulation of the regularized excitability levels.

improvement in recognition performance is the regularization of excitability levels, and not the specific direction they are shifted to.

Interestingly, it can be again observed that the purely feed-forward network configuration cannot gain the same benefit from the excitability regulation as it does the fully recurrent network configuration (Fig. 4.6, 4.7). The gain in performance after the off-line reprocessing is significantly higher for the fully recurrent configuration (Fig. 4.7). The difference in performance between the two configurations after the sleep-like state becomes even stronger articulated if the oscillation tuning is employed, enhancing the influence of contextual signaling via lateral and top-down connectivity. The fully recurrent network architecture clearly outperforms the purely feed-forward configuration in both identity and gender recognition, showing much lower error rates for the parts and the identity layer. The off-line

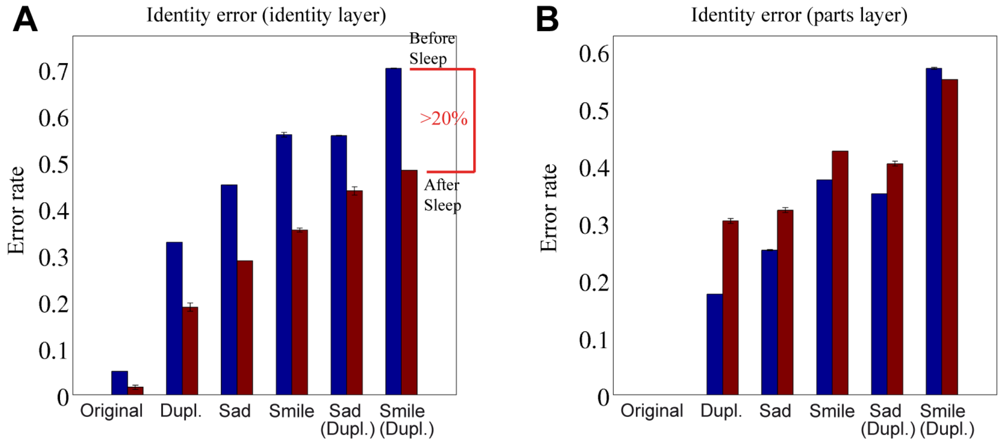


Figure 4.6: Analog to the positive effect observed for the recurrent network configuration, the purely feed-forward version is also able to benefit from the off-line reprocessing. Identity error rate for identity **(A)** and parts **(B)** layer is shown.

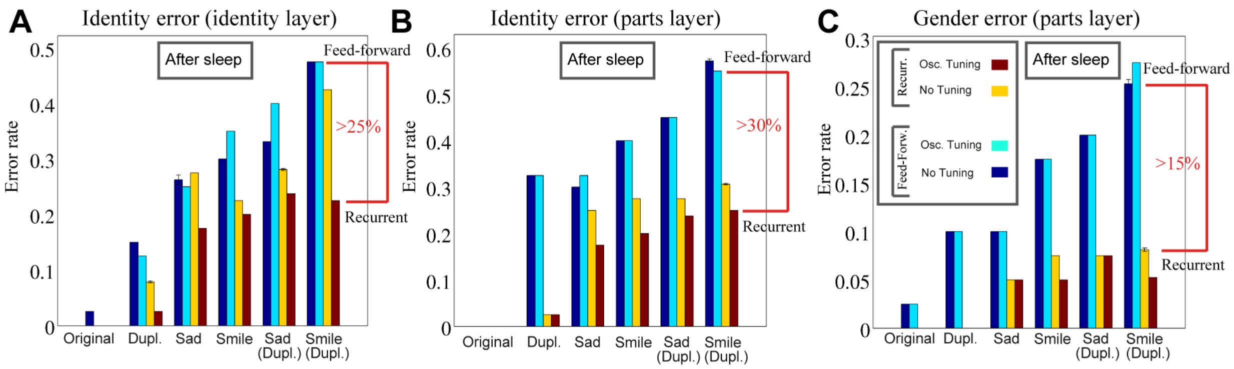


Figure 4.7: Comparison between recognition performance of feed-forward and recurrent configuration after the off-line reprocessing (upregulation regime used). Both configurations were tested with and without additional oscillation tuning (($\nu_{min} = 0.15$, $\omega_{min} = 0.5$)) after the off-line mode. Identity error for the identity layer **(A)** and the parts layer **(B)** and gender error for the parts layer **(B)** are shown. The recurrent configuration clearly outperforms the feed-forward configuration on most views, the advantage is further amplified by additional oscillation tuning.

reprocessing in the sleep-like state seems to be the right mechanism to improve the generalization capability of the fully recurrent memory network using the inherent ability of the network to self-generate activity in absence of external stimuli.

4.2 Discussion

In this chapter, a model of hierarchical visual memory formed by unsupervised learning was found to have strongly improved recognition performance after intrinsic excitability regulation during a sleep-like, off-line regime. This improvement over the initial condition before off-line memory reprocessing was stated for both the vocabulary and the identity network layer. The improvement in recognition performance was as strong or stronger on novel alternative face views than for the images presented during on-line learning, showing that off-line reprocessing is elevating the generalization capability of the memory network. Intriguingly, this functional improvement does not depend on synapse-specific plasticity, relying only on the tuning of synapse-unspecific, global excitability of units. It seems that there is no preferred direction the excitability levels should be regulated to. What really counts is that the regulatory process runs towards the equalization of intrinsic excitability across the memory units, which seems to be a favorable state of network organization after a prolonged on-line phase of unsupervised learning. Remarkably, although both the fully recurrent and purely feed-forward network configurations take benefit from the off-line reprocessing, the gain in recognition performance turns out to be significantly higher for the recurrent configuration. This indicates that the recurrent connectivity needs eventually additional maintenance in order to unfold its full capability to provide contextual support for the coherent decision making and improved memory recall. This maintenance is performed here by the system itself, in an autonomous mode that uses the inherent ability of the network to self-generate activity in the absence of external stimuli.

The causal link between regularization of intrinsic excitability in the memory network and subsequent improvement of memory function is very remarkable if interpreted from the perspective of hypothetically different memory processing strategies employed by the brain in waking and sleep states [Crick and Mitchison, 1983, Hopfield et al., 1983, Stickgold et al., 2001, Hobson and Pace-Schott, 2002]. The cortical processing and learning in the on-line regime of the wake state is biased by external stimuli which happen to occur with certain frequency in a given fragment of waking experience. In the limited episode of waking, one cannot expect to see a uniform sample of all the relevant objects, even if they could be drawn from an uniform distribution. The on-line learning has thus to operate on the restricted data available during the wakefulness. The non-uniform exposure to certain stimuli over a prolonged waking time (like it may happen while visiting an ocean beach on a hot summer day) could then lead to an overexpressed divergence of excitability levels in the cortical network, caused by the homeostatic activity regulation in response to imbalanced usage of neuronal units. This in turn could create a performance deficit for the recognition function, because the strong excitability bias induced by the specific stimuli experienced in the given waking period may influence and distort processing of other and related content. In its nature, the impairment would be similar to the overfitting phenomenon which may occur if the learner adapts too strongly to the limited data in the training set, or in other words, overlearns it. This bias, or prior, could be removed from the memory network in a sleep state. There, unbiased replay of memory content in form of spontaneously self-generated neuronal activity becomes possible, as indeed evidenced from numerous experimental findings [Euston et al., 2007, Ji and Wilson, 2007, Rasch and Born, 2007]. Without intervening external stimuli, this off-line memory reprocessing could flatten the artificial prior created during the wake period by simply regularizing network excitability levels, resetting the internal representation closer to the true environmental statistics and restoring the ability of the memory to cope with the external stimuli in an unbiased way.

Interestingly, there are some recent findings about beneficial effects of slow-wave sleep (SWS, NREM stage 4) and also restful waking on performance in cognitive tasks that were trained before, like memorizing word pair lists or sequences of finger tapping movements [Stickgold et al., 2001, Marshall and Born, 2007, Axmacher et al., 2008c]. This indicates that off-line memory reprocessing in the cortex may indeed enhance learning, improving declarative and non-declarative memory formation. Paradoxically, the conditions for conventional synapse-specific plasticity are rather unfavorable in the brain states where off-line memory reprocessing is thought to occur. In SWS, the low level of cholinergic neuromodulation and strongly diminished activity of plasticity-related early genes effectively disable individual, correlation-based synaptic potentiation [Hasselmo, 2006, Rasch and Born, 2007]. Short periods of restful waking and ultra-short naps [Axmacher et al., 2008c, Lahl et al., 2008] do not offer enough time for the rather slow classical synaptic plasticity to induce the beneficial functional changes in the memory structure. So it is unclear, what plasticity processes underlie the structural changes made during the off-line memory reprocessing, causing different memory function improvements observed experimentally after off-line states in a behavioral setting.

In the face of these findings, I hypothesize that a homeostatic activity regulation mechanism can be made responsible for the reorganization and optimization of the memory network's structure, causing synapse-unspecific changes of global neuronal excitability [Karmarkar and Buonomano, 2006, Ibata et al., 2008] that lead to the functional improvements observed in behavioral experiments. It can do so without relying on synapse-specific plasticity, and it can do it fast, so that already a short time spent in the off-line regime can be enough to get the improvement effect.

The main prediction of this hypothesis is that a regularization of excitability levels in the cortical memory network would be necessary after a prolonged time spent with on-line learning. If this excitability regularization wouldn't occur for some reason, the recognition performance, particularly with regard to the less familiar inputs requiring generalization, should be severely impaired because of the artificial bias imposed by the restricted stimuli set available during the on-line learning period. Then a neuronal process would be expected to initiate memory reprocessing that takes place in an off-line mode, either during a sleep state or during restful waking. This neuronal process would use the mechanism of homeostatic activity regulation to adjust the intrinsic excitabilities and equalize their levels across the units of the memory network, downregulating the units that are overactive during the replay and upregulating the units underactive during the replay. This off-line reprocessing should lead to a strong improvement in recognition and generalization capability back in the on-line wake state, removing the artificial bias induced in the previous period of on-line learning.

A related hypothesis, with focus on synaptic homeostasis, was provided by Tononi and Cirelli. The authors discuss homeostatic synaptic scaling performed during sleep as a mechanism to re-obtain the ability to potentiate synapses after a prolonged waking period [Tononi and Cirelli, 2003]. This perspective stresses rather the importance of local synaptic strength homeostasis [Sullivan and de Sa, 2008], as opposed to the activity homeostasis and global reorganization of the network state proposed here. It may well be the case that both synapse-unspecific homeostatic regulation mechanisms act in the off-line state in the memory network, the one equalizing the unit excitability to remove the artificial bias induced by the wake period, and the other rescaling the synapses to recover their plasticity when back in the on-line mode again.

The off-line regime examined here is only one of the possible alternatives how the off-line reprocessing can be performed in the studied memory network model. Different off-line modes can be induced by tuning the ongoing rhythms, for instance by simply changing the excitatory and inhibitory amplitude. In perspective, it would be of course interesting to investigate a second sleep-like state where off-line memory reprocessing occurs with activated synaptic plasticity and increased contextual influence via lateral and top-down pathways, mimicking the REM sleep phase. In this state, one may expect that the replay of the memory content, combined with active bidirectional plasticity [Lisman, 1989, Artola and Singer, 1993], could lead to a different quality of structural reorganization of the memory, including memory trace stabilization and amplification and reduction of interference between competing traces that overlap substantially.

This agenda pursues ultimately modeling of a complete wake-SWS-REM cycle and investigating the functional consequences of such a tri-phase memory processing for the organization and performance of memory. The phenomena studied in such a model can be then compared with experimental observations for each specific reprocessing stage. In addition, more elaborate modeling could include separate hippocampal and neocortical subsystems [Norman et al., 2005]. The model of tri-phase memory reprocessing would support the view of memory formation and maintenance as a constantly evolving dynamic multi-stage process having different functional properties during wake and sleep states. The ultimate aim of such a process is of course the optimization of memory function, delivering an evolutionary explanation for the different regimes of sleep being manifest in the nervous system of most living organisms [Cirelli and Tononi, 2008].

5

Résume and outlook

At the begin of this work, two ambitions stood in the foreground. Both of them aimed at elucidating the phenomena behind memory formation and learning in the brain on the level of cortical microcircuits. The first ambition was of a more generic, conceptual kind. It targeted better insight into neural mechanisms and generic cortical operations that may underlie the experience-driven self-organization of long-term memory for objects of natural complexity in the visual cortex. The second ambition aimed at the concrete implementation of those principles in a neuronal network architecture capable of unsupervised learning the compositional object identity from natural images.

Substantial success was achieved in both directions spanning the frame of this study. Neuronal mechanisms that presumably instantiate cooperative-competitive learning in local cortical microcircuits were postulated to be responsible for memory formation and maintenance in the visual cortex. The learning was taken to occur within discrete fragments of ongoing processing defined by the cycles of the fast gamma rhythm observed in the cortex. The postulated mechanisms were employed in a hierarchical memory network model. The network was able to memorize in an unsupervised manner a substantial number of persons from incrementally presented natural face images, employing a hierarchical compositional representation for the stored face objects. More concrete, following essential functionality was demonstrated by the system:

- The system employed hierarchical, compositional representation of face objects. It learned the local vocabularies of reusable parts, the associative relations between the parts and higher-order symbols for the global face identity simultaneously.

- The established compositional representation was of generative nature. It was possible to recreate the full compositional description of a memorized face in terms of all its parts by priming only its higher-order identity symbol or a subset of its parts.

- The system learned without supervision. Only face images were provided during the learning phase, without showing any person identity labels.

- The system formed simultaneously all kinds of connectivity - feed-forward bottom-up, recurrent lateral and top-down - within and between the network layers. Only neuronal plausible, generic mechanisms were used to drive activity and structure formation in the network.

- The unsupervised learning was self-stabilizing and life-long. There were no manually defined stop conditions that would freeze learning after some time.
- In the mature connectivity state, the network recognized reliably both person identity and gender from the alternative face views not presented during the learning.
- The system was able to recall a memorized face within a short time period of a single gamma cycle. This is comparable to the amount of time predicted by psychophysiological experiments on ultra-rapid visual object recognition.
- The network showed self-generated spontaneous memory replay in an off-line, sleep-like regime in absence of external stimuli. This off-line memory reprocessing had a direct functional benefit. The recognition performance was boosted after the off-line memory reprocessing particularly strongly on the alternative face views, indicating that off-line reprocessing improved in particular the generalization capability of the memory network. Surprisingly, the positive effect turned out to rely completely on the synapse-unspecific, homeostatic activity regulation across the memory network.

To my current knowledge, this functionality spectrum is not offered by any other existing neuronal model of object learning and recognition. Still, the implemented architecture is only a basic core which has to be developed further to provide at some point a universal visual memory domain for all kind of natural objects, stored in compositional fashion across the hierarchy of multiple, permanently self-organizing stages. On the way there, a number of hard problems has to be addressed for which no satisfactory solution is available so far.

5.1 Learning of transformation invariant object representation

One very important problem left here open is the problem of learning the transformation invariance. The question there is how to combine learning of appearance and learning of transformations that may occur to that appearance (like translation, scaling, rotation, etc), such that a memorized object can be recognized on the sensory input both in terms of its identity and the transformations applied to its image on the input [Tenenbaum and Freeman, 2000, Grimes and Rao, 2005, Miao and Rao, 2007, Culpepper and Olshausen, 2009, Memisevic and Hinton, 2010]. The current system relies on the fixed landmarks predefined on the face image to learn the vocabularies of local appearance. A system which in addition would be able to learn translation invariance, could also locate the right landmarks and signal explicitly the position of each part the face on the input is composed of.

Fortunately, there is a straightforward solution how translation invariance, and potentially its learning, could be incorporated seamlessly in the design of the system proposed here. In a series of previous works on translation-invariant object recognition, this solution was successfully tested using non-adaptive, hard-wired networks of distributed modules [Lücke, 2005, Keck, 2005, Jitsev, 2006, Lücke et al., 2008, Wolfrum et al., 2008]. Those networks employed the dynamic link architecture (DLA, [von der Malsburg, 2002a]) to establish point-to-point correspondences between the patterns on the input and disjunct object models in memory. The main idea was to use a separate domain for representing topological relations between the local appearance elements the objects are composed of. As an extension of the current system, this domain would be again simply a layer of distributed modules (Fig. 5.1). Each module from this topological layer would be responsible for determining and signaling a suitable position in the input to be mapped onto a specific part from the appearance memory domain. The mapping would be realized by the activities of the modules of the topological layer that would

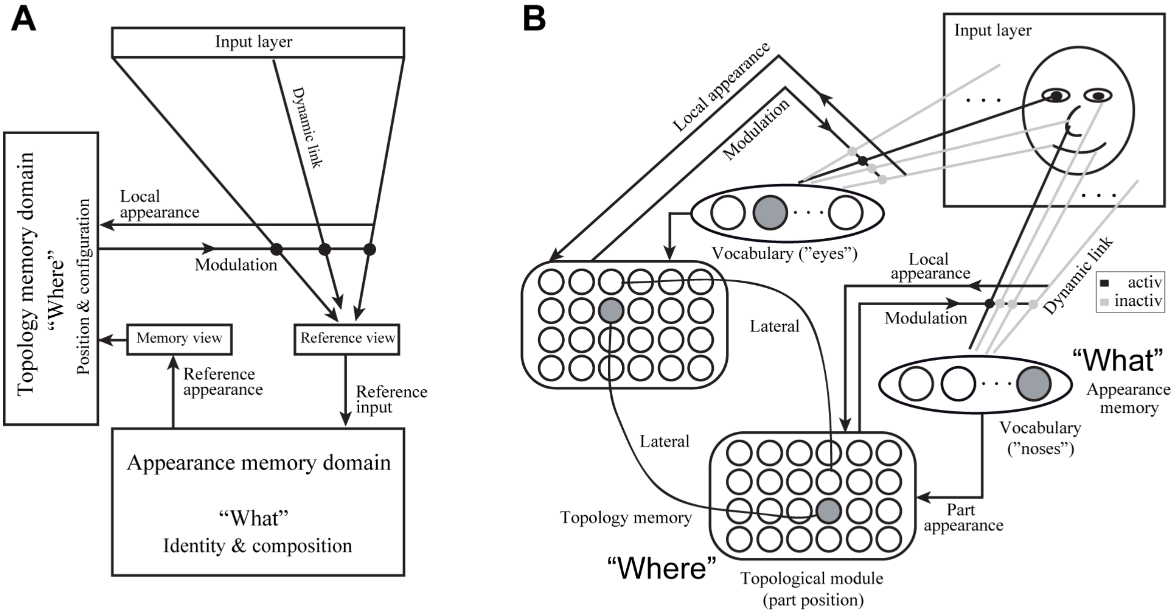


Figure 5.1: Extending the self-organizing memory domain for translation invariant learning. **(A)** General architecture, after [Jitsev, 2006] and [Wolfrum et al., 2008]. The self-organizing appearance memory domain (AM) is no longer connected directly to the fixed landmarks on the input image. The connectivity between the input layer and the AM is controlled by a self-organizing topology memory domain (TM), built, in analogy to the AM, of distributed modules. Each connection, or dynamic link, between a position on the input and a particular module in the AM can be either active or inactive depending on the activities of the TM modules. Each TM module signals a correspondence between a position on the input and a part from the AM, implementing a mapping from the input to the reference view. The mapping takes into account not only the similarities between the input and the reference memory appearance, but also the valid spatial arrangement of the parts into the whole. The information about the valid spatial arrangement of the parts is stored in the TM, while the information about the composition of parts into different object identities is stored in the AM. **(B)** The interplay between the appearance and topology memory. Each vocabulary module from the AM has a dedicated topology module on the TM. The vocabulary module provides the prototypical appearance of the memorized parts to its topology module (in a simplified case, this appearance can be a fixed average prototype, see text). The topology module decides about the position of the part on the input and routes the local appearance from the corresponding input position to the vocabulary module. The lateral connectivity in the TM is the memory for the topologically appropriate part arrangements, which have to be learned from the experience with the objects (the lateral connectivity in the AM, which is the memory for compositional object identity, is not shown). In a simpler learning task, the lateral connectivity in the TM can be pre-wired, only the AM has to be developed in this case. The system with a mature connectivity should be able to provide the compositional description of the face presented in any position on the input. The AM would signal the parts and the face identity, while the TM would explicitly assign a position on the input for each of the parts.

modulate the connectivity strength between the positions on the input and the parts in the appearance memory.

The extension of the system for translation invariant processing would add a subsystem dedicated to processing of spatial information about parts configuration. This subsystem would have its own memory domain that stores information about object topology in the lateral connectivity between the modules of the topological layer. The demand for learning in this system can be increased step by step, also increasing the complexity of the learning task. In the very first step, the topological layer can be pre-wired to contain the proper geometry of an average face [Wolfrum et al., 2008]. In addition, average local appearance of facial parts can be provided in advance, so that landmark detection is already functional before the learning starts to form the appearance memory domain. The learning task posed to this extended system would be identical to the task solved in this work, the only functional difference being the capability to perform autonomous landmark finding. In the most advanced step, one would abandon the hard-wiring of connectivity on the topological layer and also give up the assumption about the pre-existing knowledge of average face appearance. This is the hardest version of the learning task, where both appearance and topology memory domains have to be formed and maintained in an unsupervised fashion. The self-organization of the appearance memory domain was the main part of this work and first steps towards learning of lateral topological connectivity from natural images were already done in [Bouecke, 2006, Bouecke and Lücke, 2008].

In the same way one could think about adding further subsystems, each dedicated to learning of a particular transformation invariance. Some promising experiments were done for scale and rotation invariant object recognition using a compatible neuronal architecture that could be integrated into the network proposed here [Jitsev et al., 2008, Sato et al., 2009b]. Another interesting possibility would involve a separate dedicated memory domain for storing temporal sequences of objects that reside in the appearance memory domain. In such a temporal memory domain, the topological relations in space could be replaced by relations in time, and the object sequences could be stored in the lateral connectivity between the modules that signal for a certain position within the sequence. This line of research is at the very beginning, and concrete predictions about the outcome can be made only after a number of further intensive studies within the framework of the proposed network architecture.

5.2 Memory maintenance via off-line memory replay

Another important issue which has to be addressed in the further development of the system is the on-going life-long maintenance of the memory domain. Such maintenance has two general purposes. First, it has to increase the stability of the already memorized content, making the system more robust to temporal deviations from the long-term input statistics. Such deviations may include temporal removal of the relevant stimuli, or massive presentation of novel content on the expense of already familiar content. Second, it has to keep the memory domain flexible enough to be able to rapidly incorporate novel memories without getting too strong conflicts with existing memory traces. This problem setting is related to the well-known stability-plasticity dilemma which points out the difficulty to handle both tasks of keeping memorized and integrating novel content simultaneously [Grossberg, 1987b, Abraham and Robins, 2005].

In the current form, the proposed system already possesses the basic core functionality to cope with this difficulty. Its inherently sparse activity and synaptic structure provides good conditions to put new memories in the domain without disturbing the existing content too much. The stability of the created memory traces is secured by the strong synaptic coupling within the unit assemblies that form the individual traces. However, in the present architecture the phenomenon of catastrophic forgetting [Ratcliff, 1990, Robins, 1995, McClelland et al., 1995, French, 1999] may still occur if the system is confronted with the situation where the previously memorized objects are removed from the input for a longer period of time. The memory traces for those objects will most probably vanish over time and will be replaced by more recent content. On the other hand, putting a new object into an already stable memory network structure may become problematic, too. This is because the existing stable memory traces may dominate too strongly the memory attractor landscape, posing a dangerous destabilization force on the labile weak trace of the newly encoded content.

I hypothesize here that these issues can be solved by instantiating a full sleep-wake cycle, where the phases of experience-driven on-line learning alternate with off-line memory reprocessing performed in a sleep-like regime (Sec. 3.2.4, Chap. 4). In the off-line mode, where both synaptic and intrinsic

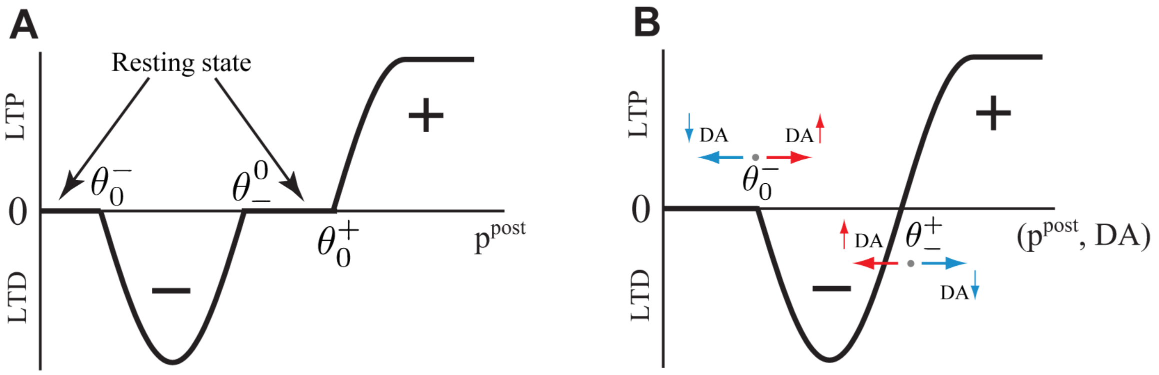


Figure 5.2: Extensions of bidirectional plasticity mechanism. **(A)** A form of bidirectional plasticity employing three sliding thresholds (after [Frégnac, 2002]). This hypothetical rule could potentially support the amplification of newly acquired, labile memory traces replayed in the off-line regime, while leaving already strong, stable traces unaffected. The essential property is hereby the additional neutral zone in the mid-range of the post-synaptic activity. **(B)** A modification of the bidirectional rule to contain a dopamine (DA) signal could render the system capable of reinforcement learning. DA could modulate the size and the position of the depression and potentiation zones by rapidly shifting the thresholds, making synaptic plasticity sensitive to the amount of the reward conveyed by the DA signaling.

plasticity would be enabled, catastrophic forgetting could be completely avoided by the self-generated replay of memory traces corresponding to stored objects. This memory trace reactivation would refresh the synaptic coupling within the corresponding assemblies, keeping the memories stable even if their sensory origin is temporarily not accessible from the environment for some reason [McClelland et al., 1995, Robins and McCallum, 1999, O'Reilly and Norman, 2002, Norman et al., 2005].

Stabilization and amplification of newly acquired weak memory traces can be supported by employing a more elaborate bidirectional plasticity mechanism, working in the on- and off-line regime. This mechanism would take use of three different sliding post-synaptic thresholds that apply for each synaptic site, as indeed suggested by some experimental evidence [Frégnac, 2002]. The potentiation of a synapse may occur there only if the synaptic site gets substantially more active than the average activation it received previously. Depression occurs in turn if the synaptic site gets substantially less active than in previous history of activation. The zone in between is the neutral zone, where no modification occurs (Fig. 5.2 (**A**)). For a stable, frequently accessed memory trace, the reactivation in the off-line mode would probably not affect the synaptic coupling, as the activity level would not be different from the previous average. In case of a weak, newly acquired memory trace, the replay would generate more activation on synaptic sites within the unit assembly than it was previously the case on average, so that potentiation of the synaptic coupling between the assembly units could be expected. The replay would thus amplify and stabilize specifically less stable traces, leaving the strong stable traces untouched. The same replay could also help to detect and remove spurious traces, that do not correspond to any proper memory content and are better forgotten.

5.3 Further forms of learning

In this work, the network model exclusively used unsupervised learning to build up the memory domain. The proposed architecture also offers interfaces that allow easy incorporation of other learning paradigms, like semisupervised and reinforcement learning. Semisupervised learning requires teacher signals that convey the labels for the presented input. These signals can be simply fed to the respective identity units. An identification compatible with the teacher signal would then further increase the probability for synaptic potentiation, whereas wrong identification would make depression more probable. In this manner, the teacher signals would be able to provide additional support for driving structure formation in the right direction.

Another type of signals that could potentially support learning in the system are reinforcement or reward signals, indicating whether the network response was good or not given task-related input. Reinforcement signals in the cortex are thought to be conveyed mainly via dopaminergic transmission [Schultz et al., 1997, Schultz, 2007]. Interestingly, dopaminergic modulation of synaptic plasticity seems to follow the same bidirectional nature as the plasticity mechanism employed in the network model [Reynolds et al., 2001, Reynolds and Wickens, 2002, Calabresi et al., 2007]. Dopamine signal may induce depression or potentiation on the synapse depending on the level of neurotransmitter release. The mechanism of dopaminergic plasticity modulation has been recently used in neuronal models of reinforcement learning [Florian, 2007, Legenstein et al., 2008, Vasilaki et al., 2009], so that it becomes suggestive to apply the same technique for extending learning capability of the proposed network architecture. The extension could be made in straightforward manner by adding into the plasticity rule a dependency on an external signal that would correspond to the diffuse release of dopamine. This signal could modulate the size and position of the depression and potentiation zones by shifting the transition thresholds on the postsynaptic side (Fig. 5.2 **(B)**). Making use of the ability to maintain persisting activity (Sec. 3.2.1), the associations between the states and successful actions (doing a saccade to the right place in the visual field) could be learned in this manner. This would open a venue towards implementation of active learning and active vision to drive memory formation in the network.

5.4 Epilog

The neural network architecture introduced in this work offers not only a unique functionality not matched by any current neuronal modeling approach to unsupervised learning of compositional object representation, but it also opens many intriguing perspectives toward better understanding the phenomenas of memory, learning and, in general, of basic principles behind the self-organization of a successful subsystem coordination across different time scales [von der Malsburg, 2002b].

From the current state of the model, two directions of research can be pursued simultaneously. The first direction leads to a more detailed neuronal modeling of the circuits, operations and plasticity mechanisms involved in the network self-organization. For instance, it would be a very challenging endeavour to dissect the microcircuit into explicit populations of excitatory and inhibitory cells that reside in different cortical layers [Douglas and Martin, 2004, Thomson and Lamy, 2007, Binzegger et al., 2009]. This would provide an opportunity to study different functional roles that the different populations may play within a cortical microcircuit, in particular with regard to the postulated competitive computation and learning carried out in gamma cycle frames. The question of how the gamma rhythms are generated and coordinated between distributed modules could also be addressed on a much more elaborated level. The explicit modeling of the microcircuit populations would also be interesting for examining the processing and learning over multiple successive gamma cycles that may be nested in a slower theta rhythm [Lisman, 2005, Jensen and Colgin, 2007, Lisman and Buzsáki, 2008]. Hypothetical functional roles of such processing, like sequential refinement of an initial coarse hypothesis about the stimulus, learning of items within a broader context or maintaining a multi-item working memory [Jensen and Lisman, 2005, Tort et al., 2009, Siegel et al., 2009, Axmacher et al., 2010], could be tested in this setting. Further analyses could concern the different forms of homeostatic and synaptic plasticity that may be active in a local microcircuit network on different time scales (like for instance short-term synaptic depression and facilitation [Markram et al., 1998, Barak and Tsodyks, 2007] or adaptive excitatory-inhibitory balance between the different populations [Karmarkar and Buonomano, 2006, Nelson and Turrigiano, 2008]). Finally, explicit spike-based modeling of the neuronal populations may be targeted. On this level of modeling, precise detailed predictions about the cortical function during phases of memory encoding, consolidation, re-consolidation and erasure should become possible.

The second direction points toward more technically motivated implementations of the system, ultimately aiming at real-world applications like automated object storage and recognition. To achieve such implementation, the employed neuronal mechanisms have to be translated into abstract language of probabilistic learning and inference. This translation can be done for instance by providing a description of the network operation in terms of learning and inference in a hierarchical graphical model [Lee and Mumford, 2003, Epshtein et al., 2008, Friston, 2008, George and Hawkins, 2009, Litvak and Ullman, 2009]. Inference of the compositional object identity given its image can be performed in such a model efficiently, using belief propagation algorithms that involve message passing across graph nodes [Pearl, 1988, Murphy et al., 1999, Yedidia et al., 2003]. Learning of a hierarchical graphical model for compositional object representation from natural image data is still a matter of ongoing research in machine learning, with a number of promising approaches proposed recently [Fergus et al., 2003, Fei-Fei et al., 2006, Fidler et al., 2009, Ommer and Buhmann, 2010, Zhu et al., 2010]. Potentially, the competitive learning implemented by neural mechanisms in the memory network model may give a hint how to rapidly construct good approximations for otherwise computationally extremely expensive Bayesian learning algorithms.

The functionality of structure self-organization and unsupervised learning demonstrated by the proposed network architecture provides thus a promising departure point for novel thrilling research. Following either direction will be surely a hard, but fruitful road to go, eventually giving rise to systems capable of discovering and storing complex regularities from natural sensory streams over multiple description levels. At the end, paraphrasing a well-known saying, if the brain were simple enough that we could easily understand it, we would definitely have only half the fun studying it.

Bibliography

L. F. Abbott and S. B. Nelson. Synaptic plasticity: taming the beast. *Nat Neurosci*, 3 Suppl:1178–1183, Nov 2000. doi: 10.1038/81453. URL `http://dx.doi.org/10.1038/81453`.

W. C. Abraham and M. F. Bear. Metaplasticity: the plasticity of synaptic plasticity. *Trends Neurosci.*, 19(4):126–130, Apr 1996.

W. C. Abraham and W. P. Tate. Metaplasticity: a new vista across the field of synaptic plasticity. *Prog. Neurobiol.*, 52(4):303–323, Jul 1997.

W. C. Abraham, S. E. Mason-Parker, M. F. Bear, S. Webb, and W. P. Tate. Heterosynaptic metaplasticity in the hippocampus in vivo: a BCM-like modifiable threshold for LTP. *Proc. Natl. Acad. Sci. U. S. A.*, 98(19):10924–10929, Sep 2001. doi: 10.1073/pnas.181342098. URL `http://dx.doi.org/10.1073/pnas.181342098`.

Wickliffe C Abraham. Metaplasticity: tuning synapses and networks for plasticity. *Nat. Rev. Neurosci.*, 9(5):387, May 2008. doi: 10.1038/nrn2356. URL `http://dx.doi.org/10.1038/nrn2356`.

Wickliffe C Abraham and Anthony Robins. Memory retention–the synaptic stability versus plasticity dilemma. *Trends Neurosci.*, 28(2):73–78, Feb 2005. doi: 10.1016/j.tins.2004.12.003. URL `http://dx.doi.org/10.1016/j.tins.2004.12.003`.

D.H. Ackley, G.E. Hinton, and T.J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9, 1985.

Stanley C. Ahalt, Ashok K. Krishnamurthy, Prakoon Chen, and Douglas E. Melton. Competitive learning algorithms for vector quantization. *Neural Netw.*, 3(3):277–290, 1990. ISSN 0893-6080. doi: http://dx.doi.org/10.1016/0893-6080(90)90071-R.

M. Ahissar and S. Hochstein. Task difficulty and the specificity of perceptual learning. *Nature*, 387 (6631):401–406, May 1997. doi: 10.1038/387401a0. URL `http://dx.doi.org/10.1038/387401a0`.

Merav Ahissar and Shaul Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.*, 8(10):457–464, Oct 2004. doi: 10.1016/j.tics.2004.08.011. URL `http://dx.doi.org/10.1016/j.tics.2004.08.011`.

Srinivas M. Aji and Robert J. McEliece. The generalized distributive law. *IEEE T. Inform. Theory.*, 46(2):325–343, 2000. URL `http://dblp.uni-trier.de/db/journals/tit/tit46.html#AjiM00`.

Mark V Albert, Adam Schnabel, and David J Field. Innate visual learning through spontaneous activity patterns. *PLoS Comput. Biol.*, 4(8):e1000137, 2008. doi: 10.1371/journal.pcbi.1000137. URL `http://dx.doi.org/10.1371/journal.pcbi.1000137`.

D. J. Amit and A. Treves. Associative memory neural network with low temporal spiking rates. *Proc. Natl. Acad. Sci. U. S. A.*, 86(20):7871–7875, Oct 1989.

Igor Antonov, Irina Antonova, Eric R Kandel, and Robert D Hawkins. Activity-dependent presynaptic facilitation and hebbian ltp are both required and interact during classical conditioning in aplysia. *Neuron*, 37(1):135–147, Jan 2003.

Igor Antonov, Eric R Kandel, and Robert D Hawkins. Presynaptic and postsynaptic mechanisms of synaptic plasticity and metaplasticity during intermediate-term memory formation in aplysia. *J. Neurosci.*, 30(16):5781–5791, Apr 2010. doi: 10.1523/JNEUROSCI.4947-09.2010. URL `http://dx.doi.org/10.1523/JNEUROSCI.4947-09.2010`.

Aristotle. (400 b.c.). de memoria et reminiscentia. In J. A. Anderson, A. Pellionisz, and E. Rosenfeld, editors, *Neurocomputing 2: Directions of Research*, pages 1–10. MIT Press, 1990.

A. Artola and W. Singer. Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends Neurosci.*, 16(11):480–487, Nov 1993.

A. Artola, S. Bröcher, and W. Singer. Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, 347(6288):69–72, Sep 1990. doi: 10.1038/347069a0. URL `http://dx.doi.org/10.1038/347069a0`.

Hiroki Asari, Barak A Pearlmutter, and Anthony M Zador. Sparse representations for the cocktail party problem. *J. Neurosci.*, 26(28):7477–7490, Jul 2006. doi: 10.1523/JNEUROSCI.1563-06.2006. URL `http://dx.doi.org/10.1523/JNEUROSCI.1563-06.2006`.

E. Aserinsky and N. Kleitman. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science*, 118(3062):273–274, Sep 1953.

D. Attwell and S. B. Laughlin. An energy budget for signaling in the grey matter of the brain. *J. Cereb. Blood Flow Metab.*, 21(10):1133–1145, Oct 2001. doi: 10.1097/00004647-200110000-00001. URL `http://dx.doi.org/10.1097/00004647-200110000-00001`.

Nikolai Axmacher, Florian Mormann, Guillen Fernández, Christian E Elger, and Juergen Fell. Memory formation by neuronal synchronization. *Brain Res. Rev.*, 52(1):170–182, Aug 2006. doi: 10.1016/j.brainresrev.2006.01.007. URL `http://dx.doi.org/10.1016/j.brainresrev.2006.01.007`.

Nikolai Axmacher, Christian E Elger, and Juergen Fell. Memory formation by refinement of neural representations: the inhibition hypothesis. *Behav. Brain Res.*, 189(1):1–8, May 2008a. doi: 10.1016/j.bbr.2007.12.018. URL `http://dx.doi.org/10.1016/j.bbr.2007.12.018`.

Nikolai Axmacher, Christian E Elger, and Juergen Fell. Ripples in the medial temporal lobe are relevant for human memory consolidation. *Brain*, 131(Pt 7):1806–1817, Jul 2008b. doi: 10.1093/brain/awn103. URL `http://dx.doi.org/10.1093/brain/awn103`.

Nikolai Axmacher, Sven Haupt, Guillén Fernández, Christian E Elger, and Juergen Fell. The role of sleep in declarative memory consolidation–direct evidence by intracranial EEG. *Cereb. Cortex*, 18(3):500–507, Mar 2008c. doi: 10.1093/cercor/bhm084. URL `http://dx.doi.org/10.1093/cercor/bhm084`.

Nikolai Axmacher, Melanie M Henseler, Ole Jensen, Ilona Weinreich, Christian E Elger, and Juergen Fell. Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc. Natl. Acad. Sci. U. S. A.*, 107(7):3228–3233, Feb 2010. doi: 10.1073/pnas.0911531107. URL `http://dx.doi.org/10.1073/pnas.0911531107`.

C. H. Bailey and M. Chen. Morphological basis of long-term habituation and sensitization in aplysia. *Science*, 220(4592):91–93, Apr 1983.

C. H. Bailey and E. R. Kandel. Structural changes accompanying memory storage. *Annu. Rev. Physiol.*, 55:397–426, 1993. doi: 10.1146/annurev.ph.55.030193.002145. URL `http://dx.doi.org/10.1146/annurev.ph.55.030193.002145`.

Omri Barak and Misha Tsodyks. Persistent activity in neural networks with dynamic synapses. *PLoS Comput. Biol.*, 3(2):e35, Feb 2007. doi: 10.1371/journal.pcbi.0030035. URL `http://dx.doi.org/10.1371/journal.pcbi.0030035`.

H. Barlow. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234, 1961. URL `http://redwood.berkeley.edu/w/images/f/fd/02-barlow-pr-1954.pdf`.

H. B. Barlow. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1 (4):371–394, 1972.

C. A. Barnes, B. L. McNaughton, S. J. Mizumori, B. W. Leonard, and L. H. Lin. Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog. Brain Res.*, 83:287–300, 1990.

M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *IEEE Trans. Neural Netw.*, 13(6):1450–1464, 2002. doi: 10.1109/TNN.2002.804287. URL `http://dx.doi.org/10.1109/TNN.2002.804287`.

A.G. Barto and P. Anandan. Pattern recognizing stochastic learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 15(3):360–375, 1985.

M. F. Bear. A synaptic basis for memory storage in the cerebral cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24):13453–13459, Nov 1996.

M. F. Bear, L. N. Cooper, and F. F. Ebner. A physiological basis for a theory of synapse modification. *Science*, 237(4810):42–48, Jul 1987.

Mark F Bear. Bidirectional synaptic plasticity: from theory to reality. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 358(1432):649–655, Apr 2003. doi: 10.1098/rstb.2002.1255. URL `http://dx.doi.org/10.1098/rstb.2002.1255`.

S. Becker and R. Zemel. Unsupervised learning with global objective functions. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1183–1187. Cambridge, MA: The MIT Press, 2nd edition, 2002.

I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.*, 94:115–147, 1987.

I. Biederman and E. E. Cooper. Priming contour-deleted images: evidence for intermediate representations in visual object recognition. *Cogn. Psychol.*, 23(3):393–419, Jul 1991.

E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2(1):32–48, Jan 1982.

T. Binzegger, R. J. Douglas, and K. A C Martin. Topology and dynamics of the canonical circuit of cat v1. *Neural Netw.*, 22(8):1071–1078, Oct 2009. doi: 10.1016/j.neunet.2009.07.011. URL `http://dx.doi.org/10.1016/j.neunet.2009.07.011`.

Niels Birbaumer and Robert F. Schmidt. *Biologische Psychologie*. Springer, Berlin, 2005. ISBN 3540254609.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

Blais, Intrator, Shouval, and Cooper. Receptive field formation in natural scene environments. comparison of single-cell learning rules. *Neural Comput*, 10(7):1797–1813, Sep 1998.

Clemens Boucsein, Martin Nawrot, Stefan Rotter, Ad Aertsen, and Detlef Heck. Controlling synaptic input patterns in vitro by dynamic photo stimulation. *J. Neurophysiol.*, 94(4):2948–2958, Oct 2005. doi: 10.1152/jn.00245.2005. URL `http://dx.doi.org/10.1152/jn.00245.2005`.

Jan D. Bouecke and Jörg Lücke. Learning of neural information routing for correspondence finding. In *ICANN '08: Proceedings of the 18th international conference on Artificial Neural Networks, Part II*, pages 557–566, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87558-1. doi: http://dx.doi.org/10.1007/978-3-540-87559-8_58.

J.D. Bouecke. Lernen von Merkmalsnachbarschaften in einem Netzwerk kortikaler Kolumnen zum Korrespondenzfinden in Bildern. Master's thesis, Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum, Germany, 2006.

Jeffrey S Bowers. On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychol. Rev.*, 116(1):220–251, Jan 2009. doi: 10.1037/a0014462. URL `http://dx.doi.org/10.1037/a0014462`.

Jeffrey S Bowers. More on grandmother cells and the biological implausibility of pdp models of cognition: a reply to plaut and mcclelland (2010) and quian quiroga and kreiman (2010). *Psychol. Rev.*, 117(1):300–6; discussion 289–90, 297–9, 306–8, Jan 2010. doi: 10.1037/a0018047. URL `http://dx.doi.org/10.1037/a0018047`.

Edward S Boyden, Feng Zhang, Ernst Bamberg, Georg Nagel, and Karl Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.*, 8(9):1263–1268, Sep 2005. doi: 10.1038/nn1525. URL `http://dx.doi.org/10.1038/nn1525`.

Björn Brembs. Operant conditioning in invertebrates. *Curr. Opin. Neurobiol.*, 13(6):710–717, Dec 2003.

Björn Brembs, Fred D Lorenzetti, Fredy D Reyes, Douglas A Baxter, and John H Byrne. Operant reward learning in aplysia: neuronal correlates and mechanisms. *Science*, 296(5573):1706–1709, May 2002. doi: 10.1126/science.1069434. URL `http://dx.doi.org/10.1126/science.1069434`.

Christoph Börgers, Steven Epstein, and Nancy J Kopell. Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. *Proc. Natl. Acad. Sci. U. S. A.*, 105 (46):18023–18028, Nov 2008. doi: 10.1073/pnas.0809511105. URL `http://dx.doi.org/10.1073/pnas.0809511105`.

Scott L Brincat and Charles E Connor. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.*, 7(8):880–886, Aug 2004. doi: 10.1038/nn1278. URL `http://dx.doi.org/10.1038/nn1278`.

Korbinian Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien, dargestellt auf Grund des Zellenbaues.* Johann Ambrosius Barth Verlag, Leipzig, 1909.

Buhmann, Divko, and Schulten. Associative memory with high information content. *Phys. Rev. A*, 39 (5):2689–2692, Mar 1989.

J. Bullier. *The visual neurosciences*, chapter Communications between cortical areas of the visual system, pages 522–540. Cambridge, MA: The MIT Press, 2003.

C. Bundesen. A theory of visual attention. *Psychol. Rev.*, 97(4):523–547, Oct 1990.

Andreas Burkhalter. Many specialists for suppressing cortical excitation. *Front. Neurosci.*, 2(2): 155–167, Dec 2008. doi: 10.3389/neuro.01.026.2008. URL `http://dx.doi.org/10.3389/neuro.01.026.2008`.

Thomas Burwick. On the relevance of local synchronization for establishing a winner-take-all functionality of the gamma cycle. *Neurocomput.*, 72(7-9):1525–1533, 2009. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2008.11.029.

Niko A Busch, Julien Dubois, and Rufin VanRullen. The phase of ongoing eeg oscillations predicts visual perception. *J. Neurosci.*, 29(24):7869–7876, Jun 2009. doi: 10.1523/JNEUROSCI.0113-09.2009. URL `http://dx.doi.org/10.1523/JNEUROSCI.0113-09.2009`.

Daniel P Buxhoeveden and Manuel F Casanova. The minicolumn hypothesis in neuroscience. *Brain*, 125(Pt 5):935–951, May 2002.

György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *Science*, 304(5679): 1926–1929, Jun 2004. doi: 10.1126/science.1099745. URL `http://dx.doi.org/10.1126/science.1099745`.

G. Buzsáki and J. J. Chrobak. Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. *Curr. Opin. Neurobiol.*, 5(4):504–510, Aug 1995.

Paolo Calabresi, Barbara Picconi, Alessandro Tozzi, and Massimiliano Di Filippo. Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci.*, 30(5):211–219, May 2007. doi: 10.1016/j.tins.2007.03.001. URL `http://dx.doi.org/10.1016/j.tins.2007.03.001`.

Xiaohua Cao, Huimin Wang, Bing Mei, Shuming An, Liang Yin, L. Phillip Wang, and Joe Z Tsien. Inducible and selective erasure of memories in the mouse brain via chemical-genetic manipulation. *Neuron*, 60(2):353–366, Oct 2008. doi: 10.1016/j.neuron.2008.08.027. URL `http://dx.doi.org/10.1016/j.neuron.2008.08.027`.

M. Carandini and D. L. Ringach. Predictions of a recurrent model of orientation selectivity. *Vision Res.*, 37(21):3061–3071, Nov 1997.

T. J. Carew, R. D. Hawkins, and E. R. Kandel. Differential classical conditioning of a defensive withdrawal reflex in aplysia californica. *Science*, 219(4583):397–400, Jan 1983.

G. C. Castellani, E. M. Quinlan, L. N. Cooper, and H. Z. Shouval. A biophysical model of bidirectional synaptic plasticity: dependence on ampa and nmda receptors. *Proc. Natl. Acad. Sci. U. S. A.*, 98 (22):12772–12777, Oct 2001. doi: 10.1073/pnas.201404598. URL `http://dx.doi.org/10.1073/pnas.201404598`.

Michele Cavazzini, Tim Bliss, and Nigel Emptage. Ca2+ and synaptic plasticity. *Cell Calcium*, 38(3-4):355–367, 2005. doi: 10.1016/j.ceca.2005.06.013. URL `http://dx.doi.org/10.1016/j.ceca.2005.06.013`.

J. P. Changeux and A. Danchin. Selective stabilisation of developing synapses as a mechanism for the specification of neuronal networks. *Nature*, 264(5588):705–712, 1976.

K. Cho, J. P. Aggleton, M. W. Brown, and Z. I. Bashir. An experimental test of the role of postsynaptic calcium levels in determining synaptic strength using perirhinal cortex of rat. *J. Physiol. (Lond.)*, 532(Pt 2):459–466, Apr 2001.

Chiara Cirelli and Giulio Tononi. Is sleep essential? *PLoS Biol.*, 6(8):e216, Aug 2008. doi: 10.1371/journal.pbio.0060216. URL `http://dx.doi.org/10.1371/journal.pbio.0060216`.

J. D. Cohen, K. Dunbar, and J. L. McClelland. On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychol. Rev.*, 97(3):332–361, Jul 1990.

N. J. Cohen and L. R. Squire. Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science*, 210(4466):207–210, Oct 1980.

Charles E Connor, Scott L Brincat, and Anitha Pasupathy. Transformation of shape information in the ventral pathway. *Curr. Opin. Neurobiol.*, 17(2):140–147, Apr 2007. doi: 10.1016/j.conb.2007.03.002. URL `http://dx.doi.org/10.1016/j.conb.2007.03.002`.

B. W. Connors and M. J. Gutnick. Intrinsic firing patterns of diverse neocortical neurons. *Trends Neurosci.*, 13(3):99–104, Mar 1990.

Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2-3):393 – 405, 1990. ISSN 0004-3702. doi: DOI:10.1016/0004-3702(90)90060-D. URL `http://www.sciencedirect.com/science/article/B6TYF-47X29XG-3H/2/f8a895d2ec1685901f93b201c47348e2`.

Stanley Coren, Lawrence M. Ward, and James T. Enns. *Sensation & Perception*. Harcourt College Pub, 1999. ISBN 015506889X.

A. Cowey and V. Walsh. Tickling the brain: studying visual sensation, perception and cognition by transcranial magnetic stimulation. *Prog. Brain Res.*, 134:411–425, 2001.

F. Crick and G. Mitchison. The function of dream sleep. *Nature*, 304(5922):111–114, 1983.

Benjamin Culpepper and Bruno Olshausen. Learning transport operators for image manifolds. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 423–431. 2009.

Mark O Cunningham, Miles A Whittington, Andrea Bibbig, Anita Roopun, Fiona E N LeBeau, Angelika Vogt, Hannah Monyer, Eberhard H Buhl, and Roger D Traub. A role for fast rhythmic bursting neurons in cortical gamma oscillations in vitro. *Proc. Natl. Acad. Sci. U. S. A.*, 101(18):7152–7157, May 2004. doi: 10.1073/pnas.0402060101. URL `http://dx.doi.org/10.1073/pnas.0402060101`.

M. Dailey and G. Cottrell. Organization of face and object recognition in modular neural network models. *Neural Netw.*, 12(7-8):1053–1074, Oct 1999.

Gaël Daoudal and Dominique Debanne. Long-term plasticity of intrinsic excitability: learning rules and mechanisms. *Learn. Mem.*, 10(6):456–465, 2003. doi: 10.1101/lm.64103. URL `http://dx.doi.org/10.1101/lm.64103`.

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, Jul 1985.

A. S. Dave and D. Margoliash. Song replay during sleep and computational rules for sensorimotor vocal learning. *Science*, 290(5492):812–816, Oct 2000.

Graeme W Davis. Homeostatic control of neural activity: from phenomenology to molecular design. *Annu. Rev. Neurosci.*, 29:307–323, 2006. doi: 10.1146/annurev.neuro.28.061604.135751. URL `http://dx.doi.org/10.1146/annurev.neuro.28.061604.135751`.

Licurgo de Almeida, Marco Idiart, and John E Lisman. Memory retrieval time and memory capacity of the ca3 network: role of gamma frequency oscillations. *Learn. Mem.*, 14(11):795–806, Nov 2007. doi: 10.1101/lm.730207. URL `http://dx.doi.org/10.1101/lm.730207`.

Licurgo de Almeida, Marco Idiart, and John E Lisman. A second function of gamma frequency oscillations: an ewinner-take-all mechanism selects which cells fire. *J. Neurosci.*, 29(23):7497–7503, Jun 2009. doi: 10.1523/JNEUROSCI.6044-08.2009. URL `http://dx.doi.org/10.1523/JNEUROSCI.6044-08.2009`.

Dominique Debanne, Gaël Daoudal, Valérie Sourdet, and Michaël Russier. Brain plasticity and ion channels. *J. Physiol. Paris*, 97(4-6):403–414, 2003. doi: 10.1016/j.jphysparis.2004.01.004. URL `http://dx.doi.org/10.1016/j.jphysparis.2004.01.004`.

Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics, 3rd Edition*. Addison Wesley, Boston, 3rd edition, 10 2001. ISBN 9780201524888.

Arnaud Delorme, Guillaume A Rousselet, Marc J-M Macé, and Michèle Fabre-Thorpe. Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Brain Res. Cogn. Brain Res.*, 19(2):103–113, Apr 2004. doi: 10.1016/j.cogbrainres.2003.11.010. URL `http://dx.doi.org/10.1016/j.cogbrainres.2003.11.010`.

Sophie Deneve. Bayesian spiking neurons ii: learning. *Neural Comput.*, 20(1):118–145, Jan 2008. doi: 10.1162/neco.2008.20.1.118. URL `http://dx.doi.org/10.1162/neco.2008.20.1.118`.

N. S. Desai, L. C. Rutherford, and G. G. Turrigiano. Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.*, 2(6):515–520, Jun 1999. doi: 10.1038/9165. URL `http://dx.doi.org/10.1038/9165`.

Duane Desieno. Adding a conscience to competitive learning. In *Proc. ICNN'88, Int. Conf. on Neural Networks*, pages 117–124, Piscataway, NJ, 1988. IEEE Service Center.

R. Desimone. Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24):13494–13499, Nov 1996.

Russell L. DeValois and Karen K. DeValois. *Spatial Vision (Oxford Psychology Series)*. Oxford University Press, USA, 1990. ISBN 019506657X.

Susanne Diekelmann and Jan Born. The memory function of sleep. *Nat. Rev. Neurosci.*, 11(2):114–126, Feb 2010. doi: 10.1038/nrn2762. URL `http://dx.doi.org/10.1038/nrn2762`.

R. J. Douglas and K. A. Martin. A functional microcircuit for cat visual cortex. *J. Physiol. (Lond.)*, 440:735–769, 1991.

R. J. Douglas, C. Koch, M. Mahowald, K. A. Martin, and H. H. Suarez. Recurrent excitation in neocortical circuits. *Science*, 269(5226):981–985, Aug 1995.

Rodney J Douglas and Kevan A C Martin. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27: 419–451, 2004. doi: 10.1146/annurev.neuro.27.070203.144152. URL `http://dx.doi.org/10.1146/annurev.neuro.27.070203.144152`.

Rodney J Douglas and Kevan A C Martin. Recurrent neuronal circuits in the neocortex. *Curr. Biol.*, 17 (13):R496–R500, Jul 2007. doi: 10.1016/j.cub.2007.04.024. URL `http://dx.doi.org/10.1016/j.cub.2007.04.024`.

Kenji Doya. Metalearning and neuromodulation. *Neural Netw.*, 15(4-6):495–506, 2002.

Daniel Durstewitz and Gustavo Deco. Computational significance of transient dynamics in cortical networks. *Eur. J. Neurosci.*, 27(1):217–227, Jan 2008. doi: 10.1111/j.1460-9568.2007.05976.x. URL `http://dx.doi.org/10.1111/j.1460-9568.2007.05976.x`.

Emrah Düzel, Will D Penny, and Neil Burgess. Brain oscillations and memory. *Curr. Opin. Neurobiol.*, Feb 2010. doi: 10.1016/j.conb.2010.01.004. URL `http://dx.doi.org/10.1016/j.conb.2010.01.004`.

G. M. Edelman. Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115–125, Feb 1993.

M. Eigen and P. Schuster. A principle of natural self-organization. *Naturwissenschaften*, 64(11):541–565, 1977.

A. K. Engel, P. König, A. K. Kreiter, T. B. Schillen, and W. Singer. Temporal coding in the visual cortex: new vistas on integration in the nervous system. *Trends Neurosci.*, 15(6):218–226, Jun 1992.

Boris Epshtein and Shimon Ullman. Semantic hierarchies for recognizing objects and parts. In *CVPR*. IEEE Computer Society, 2007. URL `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2007.html#EpshteinU07`.

Boris Epshtein, Ita Lifshitz, and Shimon Ullman. Image interpretation by a single bottom-up top-down cycle. *Proc. Natl. Acad. Sci. U. S. A.*, 105(38):14298–14303, Sep 2008. doi: 10.1073/pnas.0800968105. URL `http://dx.doi.org/10.1073/pnas.0800968105`.

David R Euston, Masami Tatsuno, and Bruce L McNaughton. Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science*, 318(5853):1147–1150, Nov 2007. doi: 10.1126/science.1148979. URL `http://dx.doi.org/10.1126/science.1148979`.

M. Fabre-Thorpe, A. Delorme, C. Marlot, and S. Thorpe. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J. Cogn. Neurosci.*, 13(2):171–180, Feb 2001.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, Apr 2006. doi: 10.1109/TPAMI.2006.79. URL `http://dx.doi.org/10.1109/TPAMI.2006.79`.

Daniel E Feldman. Synaptic mechanisms for plasticity in neocortex. *Annu. Rev. Neurosci.*, 32:33–55, 2009. doi: 10.1146/annurev.neuro.051508.135516. URL `http://dx.doi.org/10.1146/annurev.neuro.051508.135516`.

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

R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, 2003. URL `http://www-prima.inrialpes.fr/DEA_IVR/fergus03.pdf`.

S Fidler, M Boben, and A Leonardis. Learning hierarchical compositional representations of object structure. In B. Schiele S. Dickinson, A. Leonardis and M. J. Tarr, editors, *Object Categorization: Computer and Human Vision Perspectives*. Cambridge university press, 2009.

D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, Dec 1987.

Ila R Fiete, Richard H R Hahnloser, Michale S Fee, and H. Sebastian Seung. Temporal sparseness of the premotor drive is important for rapid learning in a neural network model of birdsong. *J. Neurophysiol.*, 92(4):2274–2282, Oct 2004. doi: 10.1152/jn.01133.2003. URL `http://dx.doi.org/10.1152/jn.01133.2003`.

Stefan Fischer, Manfred Hallschmid, Anna Lisa Elsner, and Jan Born. Sleep forms memory for finger skills. *Proc. Natl. Acad. Sci. U. S. A.*, 99(18):11987–11991, Sep 2002. doi: 10.1073/pnas.182178199. URL `http://dx.doi.org/10.1073/pnas.182178199`.

József Fiser and Richard N Aslin. Statistical learning of new visual feature combinations by infants. *Proc. Natl. Acad. Sci. U. S. A.*, 99(24):15822–15826, Nov 2002. doi: 10.1073/pnas.232472899. URL `http://dx.doi.org/10.1073/pnas.232472899`.

József Fiser and Richard N Aslin. Encoding multielement scenes: statistical learning of visual feature hierarchies. *J. Exp. Psychol. Gen.*, 134(4):521–537, Nov 2005. doi: 10.1037/0096-3445.134.4.521. URL `http://dx.doi.org/10.1037/0096-3445.134.4.521`.

József Fiser, Chiayu Chiu, and Michael Weliky. Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature*, 431(7008):573–578, Sep 2004. doi: 10.1038/nature02907. URL `http://dx.doi.org/10.1038/nature02907`.

R. V. Florian. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Comput*, 19(6):1468–1502, Jun 2007. doi: 10.1162/neco.2007.19.6.1468. URL `http://dx.doi.org/10.1162/neco.2007.19.6.1468`.

P. Földiák. Forming sparse representations by local anti-hebbian learning. *Biol. Cybern.*, 64(2):165–170, 1990.

P Földiák. Sparse coding in the primate cortex. In Michael A Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1064–1068. Cambridge, MA: The MIT Press, second edition, 2002.

Peter Földiák. Neural coding: non-local but explicit and conceptual. *Curr. Biol.*, 19(19):R904–R906, Oct 2009. doi: 10.1016/j.cub.2009.08.020. URL `http://dx.doi.org/10.1016/j.cub.2009.08.020`.

David J Foster and Matthew A Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683, Mar 2006. doi: 10.1038/nature04587. URL `http://dx.doi.org/10.1038/nature04587`.

D. J. Freedman, M. Riesenhuber, T. Poggio, and E. K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, Jan 2001. doi: 10.1126/science.291.5502.312. URL `http://dx.doi.org/10.1126/science.291.5502.312`.

David J Freedman, Maximilian Riesenhuber, Tomaso Poggio, and Earl K Miller. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.*, 23(12):5235–5246, Jun 2003.

French. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.*, 3(4):128–135, Apr 1999.

Brendan J. Frey and David J. C. MacKay. A revolution: Belief propagation in graphs with cycles. In *In Neural Information Processing Systems*, pages 479–485. MIT Press, 1997.

Y. Frégnac. Hebbian synaptic plasticity. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 515–522. Cambridge, MA: The MIT Press, 2002.

Pascal Fries, Danko Nikolić, and Wolf Singer. The gamma cycle. *Trends Neurosci.*, 30(7):309–316, Jul 2007. doi: 10.1016/j.tins.2007.05.005. URL `http://dx.doi.org/10.1016/j.tins.2007.05.005`.

K. J. Friston. Transients, metastability, and neuronal dynamics. *Neuroimage*, 5(2):164–171, Feb 1997. doi: 10.1006/nimg.1997.0259. URL `http://dx.doi.org/10.1006/nimg.1997.0259`.

Karl Friston. Functional integration and inference in the brain. *Prog. Neurobiol.*, 68(2):113–143, Oct 2002.

Karl Friston. Hierarchical models in the brain. *PLoS Comput. Biol.*, 4(11):e1000211, Nov 2008. doi: 10.1371/journal.pcbi.1000211. URL `http://dx.doi.org/10.1371/journal.pcbi.1000211`.

Lluís Fuentemilla, Will D Penny, Nathan Cashdollar, Nico Bunzeck, and Emrah Düzel. Theta-coupled periodic replay in working memory. *Curr. Biol.*, 20(7):606–612, Apr 2010. doi: 10.1016/j.cub.2010.01.057. URL `http://dx.doi.org/10.1016/j.cub.2010.01.057`.

I. Fujita, K. Tanaka, M. Ito, and K. Cheng. Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360(6402):343–346, Nov 1992. doi: 10.1038/360343a0. URL `http://dx.doi.org/10.1038/360343a0`.

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

J. M. Fuster. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *J. Neurophysiol.*, 36(1):61–78, Jan 1973.

J. M. Fuster and G. E. Alexander. Neuron activity related to short-term memory. *Science*, 173(997): 652–654, Aug 1971.

S. Gais, W. Plihal, U. Wagner, and J. Born. Early sleep triggers memory for early visual discrimination skills. *Nat. Neurosci.*, 3(12):1335–1339, Dec 2000. doi: 10.1038/81881. URL `http://dx.doi.org/10.1038/81881`.

Steffen Gais and Jan Born. Declarative memory consolidation: mechanisms acting during human sleep. *Learn. Mem.*, 11(6):679–685, 2004a. doi: 10.1101/lm.80504. URL `http://dx.doi.org/10.1101/lm.80504`.

Steffen Gais and Jan Born. Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proc. Natl. Acad. Sci. U. S. A.*, 101(7):2140–2144, Feb 2004b. doi: 10.1073/pnas.0305404101. URL `http://dx.doi.org/10.1073/pnas.0305404101`.

M. Galarreta and S. Hestrin. A network of fast-spiking cells in the neocortex connected by electrical synapses. *Nature*, 402(6757):72–75, Nov 1999. doi: 10.1038/47029. URL `http://dx.doi.org/10.1038/47029`.

Christophe Garcia and Manolis Delakis. Convolutional face finder: a neural architecture for fast and robust face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1408–1423, Nov 2004. doi: 10.1109/TPAMI.2004.104. URL `http://dx.doi.org/10.1109/TPAMI.2004.104`.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Comput.*, 4(1):1–58, 1992. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1992.4.1.1.

Dileep George and Jeff Hawkins. Towards a mathematical theory of cortical micro-circuits. *PLoS Comput. Biol.*, 5(10):e1000532, Oct 2009. doi: 10.1371/journal.pcbi.1000532. URL `http://dx.doi.org/10.1371/journal.pcbi.1000532`.

Wulfram Gerstner. Population dynamics of spiking neurons: Fast transients, asynchronous states, and locking. *Neural Comput.*, 12(1):43–89, 2000. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/089976600300015899.

Charles D Gilbert, Wu Li, and Valentin Piëch. Perceptual learning and adult cortical plasticity. *J. Physiol.*, 587(Pt 12):2743–2751, Jun 2009. doi: 10.1113/jphysiol.2009.171488. URL `http://dx.doi.org/10.1113/jphysiol.2009.171488`.

M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends Neurosci.*, 15(1):20–25, Jan 1992.

Anatoli Gorchetchnikov. Memory model with unsupervised sequential learning: the effect of threshold self-adjustment. In *ACM-SE 37: Proceedings of the 37th annual Southeast regional conference (CD-ROM)*, page 16, New York, NY, USA, 1999. ACM. ISBN 1-58113-128-3. doi: http://doi.acm.org/10.1145/306363.306385.

C. M. Gray and D. A. McCormick. Chattering cells: superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex. *Science*, 274(5284):109–113, Oct 1996.

C. M. Gray, P. König, A. K. Engel, and W. Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338(6213):334–337, Mar 1989. doi: 10.1038/338334a0. URL `http://dx.doi.org/10.1038/338334a0`.

C. M. Gray, A. K. Engel, P. König, and W. Singer. Synchronization of oscillatory neuronal responses in cat striate cortex: temporal properties. *Vis. Neurosci.*, 8(4):337–347, Apr 1992.

Georgia G Gregoriou, Stephen J Gotts, Huihui Zhou, and Robert Desimone. High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science*, 324(5931):1207–1210, May 2009. doi: 10.1126/science.1171402. URL `http://dx.doi.org/10.1126/science.1171402`.

David B Grimes and Rajesh P N Rao. Bilinear sparse coding for invariant vision. *Neural Comput.*, 17(1):47–73, Jan 2005. doi: 10.1162/0899766052530893. URL `http://dx.doi.org/10.1162/0899766052530893`.

S. Grossberg. Neural pattern discrimination. *J. Theor. Biol.*, 27(2):291–337, May 1970.

S. Grossberg. Adaptive pattern classification and universal recoding: Ii. feedback, expectation, olfaction, illusions. *Biol. Cybern.*, 23(4):187–202, Aug 1976a.

S. Grossberg. Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biol. Cybern.*, 23(3):121–134, Jul 1976b.

S. Grossberg. How does a brain build a cognitive code? *Psychol. Rev.*, 87(1):1–51, Jan 1980.

S. Grossberg. Cortical dynamics of three-dimensional form, color, and brightness perception: Ii. binocular theory. *Percept. Psychophys.*, 41(2):117–158, Feb 1987a.

Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23 – 63, 1987b. ISSN 0364-0213. URL `http://www.sciencedirect.com/science/article/B6W48-4FW6JX3-3/2/e1ba23e5c25e895c71a0da8362f54d3b`.

R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, Jun 2000. doi: 10.1038/35016072. URL `http://dx.doi.org/10.1038/35016072`.

Richard H R Hahnloser, Alexay A Kozhevnikov, and Michale S Fee. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature*, 419(6902):65–70, Sep 2002. doi: 10.1038/nature00974. URL `http://dx.doi.org/10.1038/nature00974`.

Bilal Haider and David A McCormick. Rapid neocortical dynamics: cellular and network mechanisms. *Neuron*, 62(2):171–189, Apr 2009. doi: 10.1016/j.neuron.2009.04.008. URL `http://dx.doi.org/10.1016/j.neuron.2009.04.008`.

Kenneth D Harris, Jozsef Csicsvari, Hajime Hirase, George Dragoi, and György Buzsáki. Organization of cell assemblies in the hippocampus. *Nature*, 424(6948):552–556, Jul 2003. doi: 10.1038/nature01834. URL `http://dx.doi.org/10.1038/nature01834`.

M. E. Hasselmo and J. L. McClelland. Neural models of memory. *Curr. Opin. Neurobiol.*, 9(2):184–188, Apr 1999.

Michael E Hasselmo. The role of acetylcholine in learning and memory. *Curr. Opin. Neurobiol.*, 16(6): 710–715, Dec 2006. doi: 10.1016/j.conb.2006.09.002. URL `http://dx.doi.org/10.1016/j.conb.2006.09.002`.

R. D. Hawkins, T. W. Abrams, T. J. Carew, and E. R. Kandel. A cellular mechanism of classical conditioning in aplysia: activity-dependent amplification of presynaptic facilitation. *Science*, 219 (4583):400–405, Jan 1983.

Kenneth J Hayworth and Irving Biederman. Neural evidence for intermediate representations in object recognition. *Vision Res.*, 46(23):4024–4031, Nov 2006. doi: 10.1016/j.visres.2006.07.015. URL `http://dx.doi.org/10.1016/j.visres.2006.07.015`.

M. Herrmann, E. Ruppin, and M. Usher. A neural model of the dynamic activation of memory. *Biol. Cybern.*, 68(5):455–463, 1993.

John Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the theory of neural computation*. Addison-Wesley Pub. Co., Redwood City, Calif., 1991. ISBN 0201515601 0201503956.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, Jul 2006. doi: 10.1162/neco.2006.18.7.1527. URL `http://dx.doi.org/10.1162/neco.2006.18.7.1527`.

J. A. Hobson, R. W. McCarley, and P. W. Wyzinski. Sleep cycle oscillation: reciprocal discharge by two brainstem neuronal groups. *Science*, 189(4196):55–58, Jun 1975.

J. Allan Hobson and Edward F Pace-Schott. The cognitive neuroscience of sleep: neuronal systems, consciousness and learning. *Nat. Rev. Neurosci.*, 3(9):679–693, Sep 2002. doi: 10.1038/nrn915. URL `http://dx.doi.org/10.1038/nrn915`.

J. J. Hopfield, D. I. Feinstein, and R. G. Palmer. 'unlearning' has a stabilizing effect in collective memories. *Nature*, 304(5922):158–159, 1983.

Horn and Usher. Neural networks with dynamical thresholds. *Phys. Rev. A*, 40(2):1036–1044, Jul 1989.

D. H. Hubel and T. N. Wiesel. Functional architecture of macaque visual cortex. *Proc. R. Soc. Lond. B*, 198:1 – 59, 1977.

P. T. Huerta and J. E. Lisman. Heightened synaptic plasticity of hippocampal ca1 neurons during a cholinergically induced rhythmic state. *Nature*, 364(6439):723–725, Aug 1993. doi: 10.1038/364723a0. URL `http://dx.doi.org/10.1038/364723a0`.

P. T. Huerta and J. E. Lisman. Bidirectional synaptic plasticity induced by a single burst during cholinergic theta oscillation in ca1 in vitro. *Neuron*, 15(5):1053–1063, Nov 1995.

A. Hyvärinen and P. Hoyer. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.*, 12(7):1705–1720, Jul 2000.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13(4-5):411–430, 2000.

Keiji Ibata, Qian Sun, and Gina G Turrigiano. Rapid synaptic scaling induced by changes in postsynaptic firing. *Neuron*, 57(6):819–826, Mar 2008. doi: 10.1016/j.neuron.2008.02.031. URL `http://dx.doi.org/10.1016/j.neuron.2008.02.031`.

Anil K. Jain and Farshid Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern Recogn.*, 24(12):1167–1186, 1991. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/0031-3203(91)90143-S.

O. Jensen and J. E. Lisman. Theta/gamma networks with slow nmda channels learn sequences and encode episodic memory: role of nmda channels in recall. *Learn. Mem.*, 3(2-3):264–278, 1996.

Ole Jensen and Laura L Colgin. Cross-frequency coupling between neuronal oscillations. *Trends Cogn. Sci.*, 11(7):267–269, Jul 2007. doi: 10.1016/j.tics.2007.05.003. URL `http://dx.doi.org/10.1016/j.tics.2007.05.003`.

Ole Jensen and John E Lisman. Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends Neurosci.*, 28(2):67–72, Feb 2005. doi: 10.1016/j.tins.2004.12.001. URL `http://dx.doi.org/10.1016/j.tins.2004.12.001`.

Daoyun Ji and Matthew A Wilson. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.*, 10(1):100–107, Jan 2007. doi: 10.1038/nn1825. URL `http://dx.doi.org/10.1038/nn1825`.

J. Jitsev and C. von der Malsburg. Off-line memory reprocessing following on-line unsupervised learning strongly improves recognition performance in a hierarchical visual memory. In *International Joint Conference on Neural Networks (IJCNN), Special session on Organic Computing*, pages 1–8. IEEE World Congress on Computational Intelligence, WCCI, Barcelona, Spain, July 2010. doi: 10.1109/IJCNN.2010.5596765. URL `http://dx.doi.org/10.1109/IJCNN.2010.5596765`.

J. Jitsev, Y.D. Sato, and C. von der Malsburg. A neural system for scale and orientation invariant correspondence finding. In *Proc. Computational and Systems Neuroscience (COSYNE)*, 2008.

Jenia Jitsev. Mustererkennung in einem neuronalen Modell kortikaler Makrokolumnennetzwerke (Pattern recognition in a neuronal model based on cortical macrocolumn networks). Master's thesis, Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum, Germany, April 2006. URL `ftp://ftp.neuroinformatik.ruhr-uni-bochum.de/pub/manuscripts/IRINI/irini2006-02/irini2006-02.pdf`.

Jenia Jitsev and Christoph von der Malsburg. Experience-driven formation of parts-based representations in a model of layered visual memory. *Front. Comput. Neurosci., Special Issue on "Complex Systems Science and Brain Dynamics"*, 3:15, Sep 2009. ISSN 1662-5188. doi: 10.3389/neuro.10.015.2009. URL `http://dx.doi.org/10.3389/neuro.10.015.2009`.

M. H. Johnson, S. Dziurawiec, H. Ellis, and J. Morton. Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1-2):1–19, Aug 1991.

E. G. Jones. Microcolumns in the cerebral cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 97:5019 – 5021, 2000.

J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1233–1258, Dec 1987.

Olivier R Joubert, Guillaume A Rousselet, Denis Fize, and Michèle Fabre-Thorpe. Processing scene context: fast categorization and object interference. *Vision Res.*, 47(26):3286–3297, Dec 2007. doi: 10.1016/j.visres.2007.09.013. URL `http://dx.doi.org/10.1016/j.visres.2007.09.013`.

C. Jutten and J. Hérault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

E. R. Kandel, J. H. Schwartz, and R. M. Jessell. *Principles of neural science*. Prentice-Hall International Inc., 2000.

N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, 17(11):4302–4311, Jun 1997.

Uma R Karmarkar and Dean V Buonomano. Different forms of homeostatic plasticity are engaged with distinct temporal profiles. *Eur. J. Neurosci.*, 23(6):1575–1584, Mar 2006. doi: 10.1111/j.1460-9568.2006.04692.x. URL `http://dx.doi.org/10.1111/j.1460-9568.2006.04692.x`.

Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, USA, 1 edition, June 1993. ISBN 0195079515. URL `http://www.worldcat.org/isbn/0195079515`.

Christian Keck. Korrespondenzfindung mit Netzwerken neuronaler Kolumnen - Entwicklung einer Simulation und Anwendung auf Bilder. Master's thesis, Institut für Neuroinformatik, Ruhr-Universität Bochum, 2005.

Tal Kenet, Dmitri Bibitchkov, Misha Tsodyks, Amiram Grinvald, and Amos Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–956, Oct 2003. doi: 10.1038/nature02078. URL `http://dx.doi.org/10.1038/nature02078`.

C. Keysers, D.-K. Xiao, P. Földiák, and D. I. Perrett. The speed of sight. *J. Cognitive Neuroscience*, 13(1):90–101, 2001. ISSN 0898-929X. doi: http://dx.doi.org/10.1162/089892901564199.

Yihwa Kim, Boris B Vladimirskiy, and Walter Senn. Modulating the granularity of category formation by global cortical states. *Front. Comput. Neurosci.*, 2:1, 2008. doi: 10.3389/neuro.10.001.2008. URL `http://dx.doi.org/10.3389/neuro.10.001.2008`.

Holle Kirchner and Simon J Thorpe. Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Res.*, 46(11):1762–1776, May 2006. doi: 10.1016/j.visres.2005.10.002. URL `http://dx.doi.org/10.1016/j.visres.2005.10.002`.

Szabolcs Káli and Peter Dayan. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat. Neurosci.*, 7(3):286–294, Mar 2004. doi: 10.1038/nn1202. URL `http://dx.doi.org/10.1038/nn1202`.

P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992.

T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin Heidelberg, 3rd edition, 2001.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, 43: 59–69, 1982.

C. Kotropoulos, A. Tefas, and I. Pitas. Frontal face authentication using morphological elastic graph matching. *IEEE Trans. Image Process.*, 9(4):555–560, 2000. doi: 10.1109/83.841933. URL `http://dx.doi.org/10.1109/83.841933`.

Minjoon Kouh and Tomaso Poggio. A canonical neural circuit for cortical nonlinear operations. *Neural Comput.*, 20(6):1427–1451, 2008. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.2008.02-07-466.

Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE T. Inform. Theory.*, 47:498–519, 2001. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.1570`.

M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE T. Comput.*, 42(3): 300–311, 1993.

Olaf Lahl, Christiane Wispel, Bernadette Willigens, and Reinhard Pietrowsky. An ultra short episode of sleep is sufficient to promote declarative memory performance. *J. Sleep Res.*, 17(1):3–10, Mar 2008. doi: 10.1111/j.1365-2869.2008.00622.x. URL `http://dx.doi.org/10.1111/j.1365-2869.2008.00622.x`.

V. A. Lamme and P. R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.*, 23(11):571–579, Nov 2000.

Victor A F Lamme. Zap! magnetic tricks on conscious and unconscious vision. *Trends Cogn. Sci.*, 10(5):193–195, May 2006. doi: 10.1016/j.tics.2006.03.002. URL `http://dx.doi.org/10.1016/j.tics.2006.03.002`.

Carien S Lansink, Pieter M Goltstein, Jan V Lankelma, Ruud N J M A Joosten, Bruce L McNaughton, and Cyriel M A Pennartz. Preferential reactivation of motivationally relevant information in the ventral striatum. *J. Neurosci.*, 28(25):6372–6382, Jun 2008. doi: 10.1523/JNEUROSCI.1054-08.2008. URL `http://dx.doi.org/10.1523/JNEUROSCI.1054-08.2008`.

Matthew E Larkum, Thomas Nevian, Maya Sandler, Alon Polsky, and Jackie Schiller. Synaptic integration in tuft dendrites of layer 5 pyramidal neurons: a new unifying principle. *Science*, 325(5941): 756–760, Aug 2009. doi: 10.1126/science.1171958. URL `http://dx.doi.org/10.1126/science.1171958`.

S. B. Laughlin. Energy as a constraint on the coding and processing of sensory information. *Curr. Opin. Neurobiol.*, 11(4):475–480, Aug 2001.

S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: a convolutional neural-network approach. *IEEE Trans. Neural Netw.*, 8(1):98–113, 1997. doi: 10.1109/72.554195. URL `http://dx.doi.org/10.1109/72.554195`.

Joachim Lübke and Dirk Feldmeyer. Excitatory signal flow and connectivity in a cortical column: focus on barrel cortex. *Brain Struct. Funct.*, 212(1):3–17, Jul 2007. doi: 10.1007/s00429-007-0144-2. URL `http://dx.doi.org/10.1007/s00429-007-0144-2`.

J. Lücke. *Information Processing and Learning in Networks of Cortical Columns*. PhD thesis, Ruhr Universität Bochum, 2005.

D. K. Lee, L. Itti, C. Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nat. Neurosci.*, 2(4):375–381, Apr 1999. doi: 10.1038/7286. URL `http://dx.doi.org/10.1038/7286`.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.*, 20(7):1434–1448, Jul 2003.

Robert Legenstein, Dejan Pecevski, and Wolfgang Maass. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS Comput. Biol.*, 4(10): e1000180, Oct 2008. doi: 10.1371/journal.pcbi.1000180. URL `http://dx.doi.org/10.1371/journal.pcbi.1000180`.

Peter Lennie. The cost of cortical computation. *Curr. Biol.*, 13(6):493–497, Mar 2003.

William B. Levy and Robert A. Baxter. Energy efficient neural codes. *Neural Comput.*, 8(3):531–543, 1996. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1996.8.3.531.

J. Lisman. A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proc. Natl. Acad. Sci. U. S. A.*, 86(23):9574–9578, Dec 1989.

J. Lisman. The CaM kinase II hypothesis for the storage of synaptic memory. *Trends Neurosci.*, 17 (10):406–412, Oct 1994.

J. E. Lisman and M. A. Idiart. Storage of 7 +/- 2 short-term memories in oscillatory subcycles. *Science*, 267(5203):1512–1515, Mar 1995.

John Lisman. The theta/gamma discrete phase code occuring during the hippocampal phase precession may be a more general brain coding scheme. *Hippocampus*, 15(7):913–922, 2005. doi: 10.1002/hipo.20121. URL `http://dx.doi.org/10.1002/hipo.20121`.

John Lisman and György Buzsáki. A neural coding scheme formed by the combined function of gamma and theta oscillations. *Schizophr. Bull.*, 34(5):974–980, Sep 2008. doi: 10.1093/schbul/sbn060. URL `http://dx.doi.org/10.1093/schbul/sbn060`.

Shai Litvak and Shimon Ullman. Cortical circuitry implementing graphical models. *Neural Comput.*, 21(11):3010–3056, Nov 2009. doi: 10.1162/neco.2009.05-08-783. URL `http://dx.doi.org/10.1162/neco.2009.05-08-783`.

Jia Liu, Alison Harris, and Nancy Kanwisher. Perception of face parts and face configurations: An fmri study. *J. Cogn. Neurosci.*, Mar 2009. doi: 10.1162/jocn.2009.21203. URL `http://dx.doi.org/10.1162/jocn.2009.21203`.

Z. Liu, J. Golowasch, E. Marder, and L. F. Abbott. A model neuron with activity-dependent conductances regulated by multiple calcium sensors. *J. Neurosci.*, 18(7):2309–2320, Apr 1998.

Hartmut S Loos, Dagmar Wieczorek, Rolf P Würtz, Christoph von der Malsburg, and Bernhard Horsthemke. Computer-based recognition of dysmorphic faces. *Eur. J. Hum. Genet.*, 11(8):555–560, Aug 2003. doi: 10.1038/sj.ejhg.5200997. URL `http://dx.doi.org/10.1038/sj.ejhg.5200997`.

R. Lorente de No. The cerebral cortex: architecture, intracortical connections and motor projections. In J.Fulton, editor, *Physiology of the nervous system*, pages 291–301. Oxford University Press, 1938.

K. Louie and M. A. Wilson. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29(1):145–156, Jan 2001.

David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society. ISBN 0-7695-0164-8.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60 (2):91–110, 2004. ISSN 0920-5691. doi: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

Jörg Lücke. Receptive field self-organization in a model of the fine structure in V1 cortical columns. *Neural Comput.*, 21(10):2805–2845, Oct 2009. doi: 10.1162/neco.2009.07-07-584. URL `http://dx.doi.org/10.1162/neco.2009.07-07-584`.

Jörg Lücke, Christian Keck, and Christoph von der Malsburg. Rapid convergence to feature layer correspondences. *Neural Comput.*, 20(10):2441–2463, 2008. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.2008.06-07-539.

Jennifer S Lund, Alessandra Angelucci, and Paul C Bressloff. Anatomical substrates for functional columns in macaque monkey primary visual cortex. *Cereb. Cortex*, 13(1):15–24, Jan 2003.

Huan Luo and David Poeppel. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–1010, Jun 2007. doi: 10.1016/j.neuron.2007.06.004. URL `http://dx.doi.org/10.1016/j.neuron.2007.06.004`.

Arianna Maffei and Alfredo Fontanini. Network homeostasis: a matter of coordination. *Curr. Opin. Neurobiol.*, 19(2):168–173, Apr 2009. doi: 10.1016/j.conb.2009.05.012. URL `http://dx.doi.org/10.1016/j.conb.2009.05.012`.

Arianna Maffei and Gina G Turrigiano. Multiple modes of network homeostasis in visual cortical layer 2/3. *J. Neurosci.*, 28(17):4377–4384, Apr 2008. doi: 10.1523/JNEUROSCI.5298-07.2008. URL `http://dx.doi.org/10.1523/JNEUROSCI.5298-07.2008`.

S. Makeig, M. Westerfield, T. P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. J. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295(5555):690–694, Jan 2002. doi: 10.1126/science.1066168. URL `http://dx.doi.org/10.1126/science.1066168`.

Pedro Maldonado, Cecilia Babul, Wolf Singer, Eugenio Rodriguez, Denise Berger, and Sonja Grün. Synchronization of neuronal responses in primary visual cortex of monkeys viewing natural images. *J. Neurophysiol.*, 100(3):1523–1532, Sep 2008. doi: 10.1152/jn.00076.2008. URL `http://dx.doi.org/10.1152/jn.00076.2008`.

S. Marcelja. Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am.*, 70 (11):1297–1300, Nov 1980.

E. Marder, L. F. Abbott, G. G. Turrigiano, Z. Liu, and J. Golowasch. Memory from the dynamics of intrinsic membrane currents. *Proc. Natl. Acad. Sci. U. S. A.*, 93(24):13481–13486, Nov 1996.

Eve Marder and Jean-Marc Goaillard. Variability, compensation and homeostasis in neuron and network function. *Nat. Rev. Neurosci.*, 7(7):563–574, Jul 2006. doi: 10.1038/nrn1949. URL `http://dx.doi.org/10.1038/nrn1949`.

Eve Marder and Astrid A Prinz. Modeling stability in neuron and network function: the role of activity in homeostasis. *Bioessays*, 24(12):1145–1154, Dec 2002. doi: 10.1002/bies.10185. URL `http://dx.doi.org/10.1002/bies.10185`.

H. Markram, Y. Wang, and M. Tsodyks. Differential signaling via the same axon of neocortical pyramidal neurons. *Proc. Natl. Acad. Sci. U. S. A.*, 95(9):5323–5328, Apr 1998.

Henry Markram, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg, and Caizhi Wu. Interneurons of the neocortical inhibitory system. *Nat. Rev. Neurosci.*, 5(10):793–807, Oct 2004. doi: 10.1038/nrn1519. URL `http://dx.doi.org/10.1038/nrn1519`.

Lisa Marshall and Jan Born. The contribution of sleep to hippocampus-dependent memory consolidation. *Trends Cogn. Sci.*, 11(10):442–450, Oct 2007. doi: 10.1016/j.tics.2007.09.001. URL `http://dx.doi.org/10.1016/j.tics.2007.09.001`.

Lisa Marshall, Halla Helgadóttir, Matthias Mölle, and Jan Born. Boosting slow oscillations during sleep potentiates memory. *Nature*, 444(7119):610–613, Nov 2006. doi: 10.1038/nature05278. URL `http://dx.doi.org/10.1038/nature05278`.

S. J. Martin, P. D. Grimwood, and R. G. Morris. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu. Rev. Neurosci.*, 23:649–711, 2000. doi: 10.1146/annurev.neuro.23.1.649. URL `http://dx.doi.org/10.1146/annurev.neuro.23.1.649`.

A.M. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC Technical Report 24, June 1998.

Luis M Martinez, Qingbo Wang, R. Clay Reid, Cinthi Pillai, José-Mañuel Alonso, Friedrich T Sommer, and Judith A Hirsch. Receptive field structure varies with layer in the primary visual cortex. *Nat. Neurosci.*, 8(3):372–379, Mar 2005. doi: 10.1038/nn1404. URL `http://dx.doi.org/10.1038/nn1404`.

J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.*, 102(3):419–457, Jul 1995.

Rumelhart D. McClelland, J. An interactive activation model of context effect in letter perception. *Cognit. Psychol.*, 23:1–44, 1981.

Roland Memisevic and Geoffrey E Hinton. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Comput.*, Feb 2010. doi: 10.1162/neco.2010.01-09-953. URL `http://dx.doi.org/10.1162/neco.2010.01-09-953`.

Xu Miao and Rajesh P N Rao. Learning the lie groups of visual invariance. *Neural Comput.*, 19(10): 2665–2693, Oct 2007. doi: 10.1162/neco.2007.19.10.2665. URL `http://dx.doi.org/10.1162/neco.2007.19.10.2665`.

E. K. Miller, L. Li, and R. Desimone. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci.*, 13(4):1460–1478, Apr 1993.

E. K. Miller, C. A. Erickson, and R. Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.*, 16(16):5154–5167, Aug 1996.

Greg Miller. Optogenetics. shining new light on neural circuits. *Science*, 314(5806):1674–1676, Dec 2006. doi: 10.1126/science.314.5806.1674. URL `http://dx.doi.org/10.1126/science.314.5806.1674`.

Kenneth D. Miller and David J. C. MacKay. The role of constraints in hebbian learning. *Neural Comput.*, 6(1):100–126, 1994. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1994.6.1.100.

B. Milner, L. R. Squire, and E. R. Kandel. Cognitive neuroscience and the study of memory. *Neuron*, 20(3):445–468, Mar 1998.

Brenda Milner, Suzanne Corkin, and H.-L. Teuber. Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of h.m. *Neuropsychologia*, 6(3):215 – 234, 1968. ISSN 0028-3932. doi: DOI:10.1016/0028-3932(68)90021-3. URL `http://www.sciencedirect.com/science/article/B6T0D-45RD8SB-22/2/c0d88ebeb701aa59bf89d5f6d0755b54`.

Marvin Minsky. Steps toward artificial intelligence. In *Computers and Thought*, pages 406–450. McGraw-Hill, 1961.

Marvin L. Minsky. *Computation: finite and infinite machines*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1967. ISBN 0-13-165563-9.

Mortimer Mishkin, Leslie G. Ungerleider, and Kathleen A. Macko. Object vision and spatial vision: two cortical pathways. *Trends Neurosci.*, 6:414 – 417, 1983. ISSN 0166-2236. doi: DOI:10.1016/0166-2236(83)90190-X. URL `http://www.sciencedirect.com/science/article/B6T0V-482YTDD-6S/2/03a01de83b0c8fc0eaeab7723aa238d1`.

V. B. Mountcastle. The columnar organization of the neocortex. *Brain*, 120 (Pt 4):701–722, Apr 1997.

Vernon B. Mountcastle. Introduction to special issue. *Cereb. Cortex*, 13(1):2–4, January 2003.

K. P. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Uncertainty in Artificial Intelligence*, pages 467–475, 1999. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.5538`.

Joseph F Murray and Kenneth Kreutz-Delgado. Visual recognition and inference using dynamic overcomplete sparse learning. *Neural Comput.*, 19(9):2301–2352, Sep 2007. doi: 10.1162/neco.2007.19.9.2301. URL `http://dx.doi.org/10.1162/neco.2007.19.9.2301`.

Sacha B Nelson and Gina G Turrigiano. Strength through diversity. *Neuron*, 60(3):477–482, Nov 2008. doi: 10.1016/j.neuron.2008.10.020. URL `http://dx.doi.org/10.1016/j.neuron.2008.10.020`.

A. Ngezahayo, M. Schachner, and A. Artola. Synaptic activity modulates the induction of bidirectional synaptic changes in adult mouse hippocampus. *J. Neurosci.*, 20(7):2451–2458, Apr 2000.

Kenneth A Norman, Ehren L Newman, and Adler J Perotte. Methods for reducing interference in the complementary learning systems model: oscillating inhibition and autonomous memory rehearsal. *Neural Netw.*, 18(9):1212–1228, Nov 2005. doi: 10.1016/j.neunet.2005.08.010. URL `http://dx.doi.org/10.1016/j.neunet.2005.08.010`.

Masato Okada. Notions of associative memory and sparse coding. *Neural Netw.*, 9(8):1429–1458, Nov 1996.

Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends Cogn. Sci.*, 11(12): 520–527, Dec 2007. doi: 10.1016/j.tics.2007.09.009. URL `http://dx.doi.org/10.1016/j.tics.2007.09.009`.

B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333–339, May 1996. doi: 10.1088/0954-898X/7/2/014. URL `http://dx.doi.org/10.1088/0954-898X/7/2/014`.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Res.*, 37(23):3311–3325, Dec 1997.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.*, 14(4): 481–487, Aug 2004. doi: 10.1016/j.conb.2004.07.007. URL `http://dx.doi.org/10.1016/j.conb.2004.07.007`.

B. Ommer and J. Buhmann. Learning compositional categorization models. *Computer Vision–ECCV 2006*, pages 316–329, 2006.

Björn Ommer and Joachim M Buhmann. Learning the compositional nature of visual object categories for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(3):501–516, Mar 2010. doi: 10.1109/TPAMI.2009.22. URL `http://dx.doi.org/10.1109/TPAMI.2009.22`.

R. C. O'Reilly. Generalization in interactive networks: the benefits of inhibitory competition and hebbian learning. *Neural Comput*, 13(6):1199–1241, Jun 2001.

Randall C. O'Reilly. Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2(11):455–462, 1998. ISSN 1364-6613. doi: DOI:10.1016/S1364-6613(98)01241-8. URL `http://www.sciencedirect.com/science/article/B6VH9-3V79YK9-D/2/dc07e90a878257b5c61097c5cbab2f61`.

Randall C. O'Reilly and Yuko Munakata. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press, Cambridge, MA, USA, 2000. ISBN 0262650541.

Randall C. O'Reilly and Kenneth A. Norman. Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends Cogn. Sci.*, 6(12):505–510, Dec 2002.

G. Palm. On associative memory. *Biol. Cybern.*, 36(1):19–31, 1980.

Eva Pastalkova, Vladimir Itskov, Asohan Amarasingham, and György Buzsáki. Internally generated cell assembly sequences in the rat hippocampus. *Science*, 321(5894):1322–1327, Sep 2008. doi: 10.1126/science.1159775. URL `http://dx.doi.org/10.1126/science.1159775`.

Anitha Pasupathy and Charles E Connor. Population coding of shape in area v4. *Nat. Neurosci.*, 5(12): 1332–1338, Dec 2002. doi: 10.1038/nn972. URL `http://dx.doi.org/10.1038/nn972`.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

J. C. Pearson, L. H. Finkel, and G. M. Edelman. Plasticity in the organization of adult cerebral cortical maps: a computer simulation based on neuronal group selection. *J. Neurosci.*, 7(12):4209–4223, Dec 1987.

Philippe Peigneux, Steven Laureys, Sonia Fuchs, Arnaud Destrebecqz, Fabienne Collette, Xavier Delbeuck, Christophe Phillips, Joel Aerts, Guy Del Fiore, Christian Degueldre, André Luxen, Axel Cleeremans, and Pierre Maquet. Learned material content and acquisition level modulate cerebral reactivation during posttraining rapid-eye-movements sleep. *Neuroimage*, 20(1):125–134, Sep 2003.

C. M A Pennartz, E. Lee, J. Verheul, P. Lipa, C. A. Barnes, and B. L. McNaughton. The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. *J. Neurosci.*, 24(29):6446–6456, Jul 2004. doi: 10.1523/JNEUROSCI.0575-04.2004. URL `http://dx.doi.org/10.1523/JNEUROSCI.0575-04.2004`.

D. I. Perrett, J. K. Hietanen, M. W. Oram, and P. J. Benson. Organization and functions of cells responsive to faces in the temporal cortex. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 335(1273): 23–30, Jan 1992. doi: 10.1098/rstb.1992.0003. URL `http://dx.doi.org/10.1098/rstb.1992.0003`.

A. Peters and C. Sethares. Myelinated axons and the pyramidal cell modules in monkey primary visual cortex. *J. Comp. Neurol.*, 365:232 – 255, 1996.

A. Peters, J. M. Cifuentes, and C. Sethares. The organization of pyramidal cells in area 18 of the rhesus monkey. *Cereb. Cortex*, 7:405 – 421, 1997.

P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1090–1104, 2000. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/34.879790.

W. A. Phillips and W. Singer. In search of common foundations for cortical computation. *Behav. Brain Sci.*, 20(4):657–83; discussion 683–722, Dec 1997.

David C Plaut and James L McClelland. Locating object knowledge in the brain: comment on bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychol. Rev.*, 117(1):284–288, Jan 2010. doi: 10.1037/a0017101. URL `http://dx.doi.org/10.1037/a0017101`.

Alessio Plebe and Rosaria Grazia Domenella. Object recognition by artificial cortical maps. *Neural Netw.*, 20(7):763–780, Sep 2007. doi: 10.1016/j.neunet.2007.04.027. URL `http://dx.doi.org/10.1016/j.neunet.2007.04.027`.

Frédéric Pouille, Antonia Marin-Burgin, Hillel Adesnik, Bassam V Atallah, and Massimo Scanziani. Input normalization by global feedforward inhibition expands cortical dynamic range. *Nat. Neurosci.*, 12(12):1577–1585, Dec 2009. doi: 10.1038/nn.2441. URL `http://dx.doi.org/10.1038/nn.2441`.

Ernst Pöppel. A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, 1 (2):56 – 61, 1997. ISSN 1364-6613. doi: DOI:10.1016/S1364-6613(97)01008-5. URL `http://www.sciencedirect.com/science/article/B6VH9-3XBTXD2-Y/2/077e9f68e80e366183101f96828b9eeb`.

Dale Purves, George J. Augustine, David Fitzpatrick, William C. Hall, Anthony-Samuel Lamantia, James O. McNamara, and S. Mark Williams. *Neuroscience*. Sinauer Associates, 2004. ISBN 0878937250.

R. Quian Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, Jun 2005. doi: 10.1038/nature03687. URL `http://dx.doi.org/10.1038/nature03687`.

R. Quian Quiroga, G. Kreiman, C. Koch, and I. Fried. Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends Cogn. Sci.*, 12(3):87–91, Mar 2008. doi: 10.1016/j.tics.2007.12.003. URL `http://dx.doi.org/10.1016/j.tics.2007.12.003`.

Rodrigo Quian Quiroga and Gabriel Kreiman. Measuring sparseness in the brain: comment on bowers (2009). *Psychol. Rev.*, 117(1):291–297, Jan 2010. doi: 10.1037/a0016917. URL `http://dx.doi.org/10.1037/a0016917`.

M. I. Rabinovich, R. Huerta, A. Volkovskii, H. D. Abarbanel, M. Stopfer, and G. Laurent. Dynamical coding of sensory information with competitive networks. *J. Physiol. Paris*, 94(5-6):465–471, 2000.

Mikhail I Rabinovich, Ramón Huerta, Pablo Varona, and Valentin S Afraimovich. Transient cognitive dynamics, metastability, and decision making. *PLoS Comput. Biol.*, 4(5):e1000072, May 2008. doi: 10.1371/journal.pcbi.1000072. URL `http://dx.doi.org/10.1371/journal.pcbi.1000072`.

Marc'Aurelio Ranzato. *Unsupervised Learning of Feature Hierarchies*. PhD thesis, New York University, May 2009. URL `http://www.cs.nyu.edu/~{}ranzato/publications/ranzato-phd-thesis.pdf`.

Marc'Aurelio Ranzato, Fu-Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. Computer Vision and Pattern Recognition Conference (CVPR'07)*. IEEE Press, 2007.

Marc'Aurelio Ranzato, Y-Lan Boureau, and Yann LeCun. Sparse feature learning for deep belief networks. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1185–1192. MIT Press, Cambridge, MA, 2008.

R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, 2(1):79–87, Jan 1999. doi: 10.1038/4580. URL `http://dx.doi.org/10.1038/4580`.

Björn Rasch and Jan Born. Maintaining memories by reactivation. *Curr. Opin. Neurobiol.*, 17(6):698–703, Dec 2007. doi: 10.1016/j.conb.2007.11.007. URL `http://dx.doi.org/10.1016/j.conb.2007.11.007`.

Björn Rasch, Christian Büchel, Steffen Gais, and Jan Born. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science*, 315(5817):1426–1429, Mar 2007. doi: 10.1126/science.1138581. URL `http://dx.doi.org/10.1126/science.1138581`.

R. Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychol. Rev.*, 97(2):285–308, Apr 1990.

Leila Reddy and Nancy Kanwisher. Coding of visual objects in the ventral stream. *Curr. Opin. Neurobiol.*, 16(4):408–414, Aug 2006. doi: 10.1016/j.conb.2006.06.004. URL `http://dx.doi.org/10.1016/j.conb.2006.06.004`.

Martin Rehn and Friedrich T Sommer. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J. Comput. Neurosci.*, 22(2):135–146, Apr 2007. doi: 10.1007/s10827-006-0003-9. URL `http://dx.doi.org/10.1007/s10827-006-0003-9`.

J. H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas v2 and v4. *J. Neurosci.*, 19(5):1736–1753, Mar 1999.

J. N. Reynolds, B. I. Hyland, and J. R. Wickens. A cellular mechanism of reward-related learning. *Nature*, 413(6851):67–70, Sep 2001. doi: 10.1038/35092560. URL `http://dx.doi.org/10.1038/35092560`.

John N J Reynolds and Jeffery R Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw.*, 15(4-6):507–521, 2002.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–1025, Nov 1999. doi: 10.1038/14819. URL `http://dx.doi.org/10.1038/14819`.

M. Riesenhuber and T. Poggio. Models of object recognition. *Nat. Neurosci.*, 3 Suppl:1199–1204, Nov 2000. doi: 10.1038/81479. URL `http://dx.doi.org/10.1038/81479`.

Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.*, 88(1):455–463, Jul 2002.

A. Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

A. Robins and S. McCallum. The consolidation of learning during sleep: comparing the pseudorehearsal and unlearning accounts. *Neural Netw.*, 12(7-8):1191–1206, Oct 1999.

Kathleen S Rockland and Noritaka Ichinohe. Some thoughts on cortical minicolumns. *Exp. Brain Res.*, 158(3):265–277, Oct 2004. doi: 10.1007/s00221-004-2024-9. URL `http://dx.doi.org/10.1007/s00221-004-2024-9`.

E. Rodriguez, N. George, J. P. Lachaux, J. Martinerie, B. Renault, and F. J. Varela. Perception's shadow: long-distance synchronization of human brain activity. *Nature*, 397(6718):430–433, Feb 1999. doi: 10.1038/17120. URL `http://dx.doi.org/10.1038/17120`.

E. T. Rolls and M. J. Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.*, 73(2):713–726, Feb 1995.

Edmund Rolls. *Computational Neuroscience: A Comprehensive Approach*, chapter The Operation of Memory Systems in the Brain. Chapman & Hall/CRC Mathematical & Computational Biology. Chapman and Hall/CRC, 2004. ISBN 978-1-58488-362-3. URL `http://dx.doi.org/10.1201/9780203494462.ch16`.

Edmund T. Rolls and Gustavo Deco. *Computational Neuroscience of Vision*. Oxford University Press, USA, 2002. ISBN 0198524889.

E.T. Rolls and A. Treves. The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network: Computation in Neural Systems*, 1(4):407–421, 1990.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6):386–408, Nov 1958.

Guillaume A Rousselet, Simon J Thorpe, and Michèle Fabre-Thorpe. How parallel is visual processing in the ventral pathway? *Trends Cogn. Sci.*, 8(8):363–370, Aug 2004. doi: 10.1016/j.tics.2004.06.003. URL `http://dx.doi.org/10.1016/j.tics.2004.06.003`.

Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:23–38, 1998. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/34.655647.

Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.*, 20(10):2526–2563, Oct 2008. doi: 10.1162/neco.2008.03-07-486. URL `http://dx.doi.org/10.1162/neco.2008.03-07-486`.

David E. Rumelhart and David Zipser. Feature discovery by competitive learning. *Cognitive Science*, 9 (1):75–112, 1985. URL `http://cognitrn.psych.indiana.edu/rgoldsto/cogsci/Rumelhart.pdf`.

Magdalena Sanhueza, Charmian C McIntyre, and John E Lisman. Reversal of synaptic memory by ca2+/calmodulin-dependent protein kinase ii inhibitor. *J. Neurosci.*, 27(19):5190–5199, May 2007. doi: 10.1523/JNEUROSCI.5049-06.2007. URL `http://dx.doi.org/10.1523/JNEUROSCI.5049-06.2007`.

Takayuki Sato, Go Uchida, and Manabu Tanifuji. Cortical columnar organization is reconsidered in inferior temporal cortex. *Cereb. Cortex*, 19(8):1870–1888, Aug 2009a. doi: 10.1093/cercor/bhn218. URL `http://dx.doi.org/10.1093/cercor/bhn218`.

Y.D. Sato, J. Jitsev, and C von der Malsburg. A visual object recognition system invariant to scale and rotation. *Neural Network World (ICANN Special Issue)*, 19(5):529–544, 2009b.

Paul Sauseng, Wolfgang Klimesch, Walter R Gruber, and Niels Birbaumer. Cross-frequency phase synchronization: a brain mechanism of memory matching and attention. *Neuroimage*, 40(1):308–317, Mar 2008. doi: 10.1016/j.neuroimage.2007.11.032. URL `http://dx.doi.org/10.1016/j.neuroimage.2007.11.032`.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275 (5306):1593–1599, Mar 1997.

Wolfram Schultz. Behavioral dopamine signals. *Trends Neurosci.*, 30(5):203–210, May 2007. doi: 10.1016/j.tins.2007.03.007. URL `http://dx.doi.org/10.1016/j.tins.2007.03.007`.

W. B. Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry*, 20(1):11–21, Feb 1957.

Oliver G. Selfridge. Pandemonium: A paradigm for learning. *National Physical Laboratory*, 10: 513–529, 1958 1958.

Thomas Serre, Gabriel Kreiman, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, and Tomaso Poggio. A quantitative theory of immediate visual recognition. *Prog. Brain Res.*, 165:33–56, 2007a. doi: 10.1016/S0079-6123(06)65004-8. URL `http://dx.doi.org/10.1016/S0079-6123(06)65004-8`.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U. S. A.*, 104(15):6424–6429, Apr 2007b. doi: 10.1073/pnas.0700622104. URL `http://dx.doi.org/10.1073/pnas.0700622104`.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3): 411–426, Mar 2007c. doi: 10.1109/TPAMI.2007.56. URL `http://dx.doi.org/10.1109/TPAMI.2007.56`.

Gordon M G Shepherd and Karel Svoboda. Laminar and columnar organization of ascending excitatory projections to layer 2/3 pyramidal neurons in rat barrel cortex. *J. Neurosci.*, 25(24):5670–5679, Jun 2005. doi: 10.1523/JNEUROSCI.1173-05.2005. URL `http://dx.doi.org/10.1523/JNEUROSCI.1173-05.2005`.

Shy Shoham, Daniel H O'Connor, Dmitry V Sarkisov, and Samuel S-H Wang. Rapid neurotransmitter uncaging in spatially defined patterns. *Nat. Methods*, 2(11):837–843, Nov 2005. doi: 10.1038/nmeth793. URL `http://dx.doi.org/10.1038/nmeth793`.

A. G. Siapas and M. A. Wilson. Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron*, 21(5):1123–1128, Nov 1998.

M. Sidman, L.T. Stoddard, and J.P. Mohr. Some additional quantitative observations of immediate memory in a patient with bilateral hippocampal lesions. *Neuropsychologia*, 6 (3):245 – 254, 1968. ISSN 0028-3932. doi: DOI:10.1016/0028-3932(68)90023-7. URL `http://www.sciencedirect.com/science/article/B6T0D-45RD8SB-24/2/51c4359d92140dc6adf19761ac3a2197`.

Markus Siegel, Melissa R Warden, and Earl K Miller. Phase-dependent neuronal coding of objects in short-term memory. *Proc. Natl. Acad. Sci. U. S. A.*, 106(50):21341–21346, Dec 2009. doi: 10.1073/pnas.0908193106. URL `http://dx.doi.org/10.1073/pnas.0908193106`.

Herbert Alexander Simon. *The shape of automation for men and management*. Harper & Row New York,, [1st ed.] edition, 1965. School of Commerce, Accounts, and Finance, New York, 1960.

E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, 24:1193–1216, 2001. doi: 10.1146/annurev.neuro.24.1.1193. URL `http://dx.doi.org/10.1146/annurev.neuro.24.1.1193`.

W. Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24(1):49–65, 111–25, Sep 1999.

W. E. Skaggs and B. L. McNaughton. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, 271(5257):1870–1873, Mar 1996.

Vikaas S Sohal, Feng Zhang, Ofer Yizhar, and Karl Deisseroth. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature*, 459(7247):698–702, Jun 2009. doi: 10.1038/nature07991. URL `http://dx.doi.org/10.1038/nature07991`.

D. C. Somers, S. B. Nelson, and M. Sur. An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.*, 15(8):5448–5465, Aug 1995.

Friedrich Sommer and Gunther Palm. Improved bidirectional retrieval of sparse patterns stored by hebbian learning. *Neural Netw.*, 12(2):281–297, Mar 1999.

Sen Song, Per Jesper Sjöström, Markus Reigl, Sacha Nelson, and Dmitri B Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.*, 3(3):e68, Mar 2005. doi: 10.1371/journal.pbio.0030068. URL `http://dx.doi.org/10.1371/journal.pbio.0030068`.

M. W. Spratling and M. H. Johnson. Preintegration lateral inhibition enhances unsupervised learning. *Neural Comput.*, 14(9):2157–2179, 2002. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/089976602320264033.

Michael W Spratling. Reconciling predictive coding and biased competition models of cortical function. *Front. Comput. Neurosci.*, 2:4, 2008. doi: 10.3389/neuro.10.004.2008. URL `http://dx.doi.org/10.3389/neuro.10.004.2008`.

Nelson Spruston. Pyramidal neurons: dendritic structure and synaptic integration. *Nat. Rev. Neurosci.*, 9(3):206–221, Mar 2008. doi: 10.1038/nrn2286. URL `http://dx.doi.org/10.1038/nrn2286`.

L. R. Squire and S. Zola-Morgan. The medial temporal lobe memory system. *Science*, 253(5026): 1380–1386, Sep 1991.

Larry R Squire. Memory and brain systems: 1969-2009. *J. Neurosci.*, 29(41):12711–12716, Oct 2009. doi: 10.1523/JNEUROSCI.3575-09.2009. URL `http://dx.doi.org/10.1523/JNEUROSCI.3575-09.2009`.

L.R Squire. *Memory and Brain*. New York: Oxford University Press, 1987.

L.R. Squire and C.E.L. Stark. *Topics in Integrative Neuroscience: From Cells to Cognition.*, chapter Memory Systems, pages 243–264. Oxford University Press, 2008.

Terrence R Stanford, Swetha Shankar, Dino P Massoglia, M. Gabriela Costello, and Emilio Salinas. Perceptual decision making in less than 30 milliseconds. *Nat. Neurosci.*, 13(3):379–385, Mar 2010. doi: 10.1038/nn.2485. URL `http://dx.doi.org/10.1038/nn.2485`.

M. Steriade, D. A. McCormick, and T. J. Sejnowski. Thalamocortical oscillations in the sleeping and aroused brain. *Science*, 262(5134):679–685, Oct 1993.

R. Stickgold, L. James, and J. A. Hobson. Visual discrimination learning requires sleep after training. *Nat. Neurosci.*, 3(12):1237–1238, Dec 2000. doi: 10.1038/81756. URL `http://dx.doi.org/10.1038/81756`.

R. Stickgold, J. A. Hobson, R. Fosse, and M. Fosse. Sleep, learning, and dreams: off-line memory reprocessing. *Science*, 294(5544):1052–1057, Nov 2001. doi: 10.1126/science.1063530. URL `http://dx.doi.org/10.1126/science.1063530`.

Robert Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278, Oct 2005. doi: 10.1038/nature04286. URL `http://dx.doi.org/10.1038/nature04286`.

Thomas J Sullivan and Virginia R de Sa. Sleeping our way to weight normalization and stable learning. *Neural Comput.*, 20(12):3111–3130, Dec 2008. doi: 10.1162/neco.2008.04-07-502. URL `http://dx.doi.org/10.1162/neco.2008.04-07-502`.

Harvey A Swadlow. Fast-spike interneurons and feedforward inhibition in awake sensory neocortex. *Cereb. Cortex*, 13(1):25–32, Jan 2003.

J. Szentágothai. The ferrier lecture, 1977. the neuron network of the cerebral cortex: a functional interpretation. *Proc. R. Soc. Lond. B Biol. Sci.*, 201(1144):219–248, May 1978.

K. Tanaka. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, 19:109–139, 1996. doi: 10.1146/annurev.ne.19.030196.000545. URL `http://dx.doi.org/10.1146/annurev.ne.19.030196.000545`.

K. Tanaka. Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex*, 13(1):90–99, 2003.

J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Comput.*, 12(6):1247–1283, Jun 2000.

Fabian J. Theis. Towards a general independent subspace analysis. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1361–1368. MIT Press, Cambridge, MA, 2007.

Alex M Thomson and A. Peter Bannister. Interlaminar connections in the neocortex. *Cereb. Cortex*, 13(1):5–14, Jan 2003.

Alex M Thomson and Christophe Lamy. Functional maps of neocortical local circuitry. *Front. Neurosci.*, 1(1):19–42, Nov 2007. doi: 10.3389/neuro.01.1.1.002.2007. URL `http://dx.doi.org/10.3389/neuro.01.1.1.002.2007`.

S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381 (6582):520–522, Jun 1996. doi: 10.1038/381520a0. URL `http://dx.doi.org/10.1038/381520a0`.

S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Netw.*, 14(6-7):715–725, 2001.

Paul Tiesinga and Terrence J Sejnowski. Cortical enlightenment: are attentional gamma oscillations driven by ing or ping? *Neuron*, 63(6):727–732, Sep 2009. doi: 10.1016/j.neuron.2009.09.009. URL `http://dx.doi.org/10.1016/j.neuron.2009.09.009`.

Matthew H Tong, Carrie A Joyce, and Garrison W Cottrell. Why is the fusiform face area recruited for novel categories of expertise? a neurocomputational investigation. *Brain Res.*, 1202:14–24, Apr 2008. doi: 10.1016/j.brainres.2007.06.079. URL `http://dx.doi.org/10.1016/j.brainres.2007.06.079`.

Giulio Tononi and Chiara Cirelli. Sleep and synaptic homeostasis: a hypothesis. *Brain Res. Bull.*, 62 (2):143–150, Dec 2003.

Adriano B L Tort, Robert W Komorowski, Joseph R Manns, Nancy J Kopell, and Howard Eichenbaum. Theta-gamma coupling increases during the learning of item-context associations. *Proc. Natl. Acad. Sci. U. S. A.*, Nov 2009. doi: 10.1073/pnas.0911331106. URL `http://dx.doi.org/10.1073/pnas.0911331106`.

R. D. Traub, M. A. Whittington, I. M. Stanford, and J. G. Jefferys. A mechanism for generation of long-range synchronous fast oscillations in the cortex. *Nature*, 383(6601):621–624, Oct 1996. doi: 10.1038/383621a0. URL `http://dx.doi.org/10.1038/383621a0`.

Jochen Triesch. Synergies between intrinsic and synaptic plasticity mechanisms. *Neural Comput.*, 19(4):885–909, Apr 2007. doi: 10.1162/neco.2007.19.4.885. URL `http://dx.doi.org/10.1162/neco.2007.19.4.885`.

Doris Y Tsao, Winrich A Freiwald, Roger B H Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, Feb 2006. doi: 10.1126/science.1119983. URL `http://dx.doi.org/10.1126/science.1119983`.

M. Tsodyks, T. Kenet, A. Grinvald, and A. Arieli. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286(5446):1943–1946, Dec 1999.

M. V. Tsodyks and M. V. Feigel'man. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.*, 6(2):101–105, 1988. URL `http://stacks.iop.org/0295-5075/6/101`.

John K. Tsotsos, Sean M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artif. Intell.*, 78(1-2):507–545, 1995. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/0004-3702(95)00025-9.

K. Tsunoda, Y. Yamane, M. Nishizaki, and M. Tanifuji. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nat. Neurosci.*, 4(8):832–838, Aug 2001. doi: 10.1038/90547. URL `http://dx.doi.org/10.1038/90547`.

M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, pages 586–591, 1991a. doi: 10.1109/CVPR.1991.139758. URL `http://dx.doi.org/10.1109/CVPR.1991.139758`.

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cogn. Neurosci.*, 3(1):71–86, 1991b. ISSN 0898-929X. doi: http://dx.doi.org/10.1162/jocn.1991.3.1.71.

M. R. Turner. Texture discrimination by gabor functions. *Biol. Cybern.*, 55(2-3):71–82, 1986.

Gina G Turrigiano. The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, 135(3):422–435, Oct 2008. doi: 10.1016/j.cell.2008.10.008. URL `http://dx.doi.org/10.1016/j.cell.2008.10.008`.

Shimon Ullman, Michel Vidal-Naquet, and Erez Sali. Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.*, 5(7):682–687, Jul 2002. doi: 10.1038/nn870. URL `http://dx.doi.org/10.1038/nn870`.

Ilkay Ulusoy and Christopher M. Bishop. Generative versus discriminative methods for object recognition. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 258–265, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2372-2. doi: http://dx.doi.org/10.1109/CVPR.2005.167.

J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biol. Sci.*, 265(1394):359–366, Mar 1998. doi: 10.1098/rspb.1998.0303. URL `http://dx.doi.org/10.1098/rspb.1998.0303`.

R. VanRullen. The power of the feed-forward sweep. *Advances in Cognitive Psychology*, 3(1):167–176, 2007.

R. VanRullen and S. J. Thorpe. Is it a bird? is it a plane? ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, 30(6):655–668, 2001.

Rufin VanRullen and Christof Koch. Is perception discrete or continuous? *Trends Cogn. Sci.*, 7(5): 207–213, May 2003.

Rufin VanRullen, Leila Reddy, and Christof Koch. Attention-driven discrete sampling of motion perception. *Proc. Natl. Acad. Sci. U. S. A.*, 102(14):5291–5296, Apr 2005. doi: 10.1073/pnas.0409172102. URL `http://dx.doi.org/10.1073/pnas.0409172102`.

Rufin VanRullen, Thomas Carlson, and Patrick Cavanagh. The blinking spotlight of attention. *Proc. Natl. Acad. Sci. U. S. A.*, 104(49):19204–19209, Dec 2007. doi: 10.1073/pnas.0707316104. URL `http://dx.doi.org/10.1073/pnas.0707316104`.

F. G. Varela, H. R. Maturana, and R. Uribe. Autopoiesis: the organization of living systems, its characterization and a model. *Curr. Mod. Biol.*, 5(4):187–196, May 1974.

Eleni Vasilaki, Nicolas Frémaux, Robert Urbanczik, Walter Senn, and Wulfram Gerstner. Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. *PLoS Comput. Biol.*, 5(12):e1000586, Dec 2009. doi: 10.1371/journal.pcbi.1000586. URL `http://dx.doi.org/10.1371/journal.pcbi.1000586`.

Benjamin T Vincent, Roland J Baddeley, Tom Troscianko, and Iain D Gilchrist. Is the early visual system optimised to be energy efficient? *Network*, 16(2-3):175–190, 2005.

W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, Feb 2000.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:511–518, April 2001. ISSN 1063-6919. doi: 10.1109/CVPR.2001.990517. URL `http://dx.doi.org/10.1109/CVPR.2001.990517`.

C. von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100, Dec 1973.

C. von der Malsburg. The correlation theory of brain function. Internal Report 81-2, MPI Biophysical Chemistry, 1981. Reprinted in E. Domany, J. L. van Hemmen, and K. Schulten, Editors, *Models of Neural Networks II*, chapter 2, pages 95–119. Springer, Berlin, 1994.

C. von der Malsburg. Network self-organization in the ontogenesis of the mammalian visual system. In S. F. Zornetzer, J. Davis, C. Lau, and T. McKenna, editors, *An Introduction to Neural and Electronic Networks*, pages 447–462. Academic Press, 2nd edition, 1995a.

C. von der Malsburg. Binding in models of perception and brain function. *Curr. Opin. Neurobiol.*, 5 (4):520–526, Aug 1995b.

C. von der Malsburg. The what and why of binding: The modeler´s perspective. *Neuron*, 24(1):95–104, 1999.

C. von der Malsburg. Dynamic link architecture. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 365–368. Cambridge, MA: The MIT Press, 2nd edition, 2002a.

C. von der Malsburg. Self-organization and the brain. In M.A. Arbib, editor, *The handbook of brain theory and neural networks*, pages 1002–1005. Cambridge, MA: The MIT Press, Cambridge, MA, USA, 2 edition, 2002b.

C. von der Malsburg and W. Singer. Principles of cortical network organization. In P. Rakic and W. Singer, editors, *Neurobiology of Neocortex*, pages 69–99. Wiley, New York, 1988.

von Economo K. and GN. Koskinas. *Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen*. Springer Verlag, 1925.

T. Wagner, N. Axmacher, K. Lehnertz, C. E. Elger, and J. Fell. Sleep-dependent directional coupling between human neocortex and hippocampus. *Cortex*, Jun 2009. doi: 10.1016/j.cortex.2009.05.012. URL `http://dx.doi.org/10.1016/j.cortex.2009.05.012`.

Ullrich Wagner, Steffen Gais, Hilde Haider, Rolf Verleger, and Jan Born. Sleep inspires insight. *Nature*, 427(6972):352–355, Jan 2004. doi: 10.1038/nature02223. URL `http://dx.doi.org/10.1038/nature02223`.

Ullrich Wagner, Naveen Kashyap, Susanne Diekelmann, and Jan Born. The impact of post-learning sleep vs. wakefulness on recognition memory for faces with different facial expressions. *Neurobiol. Learn. Mem.*, 87(4):679–687, May 2007. doi: 10.1016/j.nlm.2007.01.004. URL `http://dx.doi.org/10.1016/j.nlm.2007.01.004`.

Matthew P Walker, Tiffany Brakefield, Alexandra Morgan, J. Allan Hobson, and Robert Stickgold. Practice with sleep makes perfect: sleep-dependent motor skill learning. *Neuron*, 35(1):205–211, Jul 2002.

Guy Wallis, Ulrike E Siebeck, Kellie Swann, Volker Blanz, and Heinrich H Bülthoff. The prototype effect revisited: Evidence for an abstract feature model of face recognition. *J. Vis.*, 8(3):20.1–2015, 2008. doi: 10.1167/8.3.20. URL `http://dx.doi.org/10.1167/8.3.20`.

E. T. Walters and J. H. Byrne. Associative conditioning of single sensory neurons suggests a cellular mechanism for learning. *Science*, 219(4583):405–408, Jan 1983.

Stephen Waydo and Christof Koch. Unsupervised learning of individuals and categories from images. *Neural Comput.*, 20(5):1165–1178, 2008. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.2007.03-07-493.

Y. Weiss. Correctness of local probability in graphical models with loops. *Neural Comput.*, 12(1): 1–41, Jan 2000.

Michael Weliky, József Fiser, Ruskin H Hunt, and David N Wagner. Coding of natural scenes in primary visual cortex. *Neuron*, 37(4):703–718, Feb 2003.

Valérie Wespatat, Frank Tennigkeit, and Wolf Singer. Phase sensitivity of synaptic modifications in oscillating cells of rat visual cortex. *J. Neurosci.*, 24(41):9067–9075, Oct 2004. doi: 10.1523/JNEUROSCI.2221-04.2004. URL `http://dx.doi.org/10.1523/JNEUROSCI.2221-04.2004`.

Günter Westphal and Rolf P Würtz. Combining feature- and correspondence-based methods for visual object recognition. *Neural Comput.*, Mar 2009. doi: 10.1162/neco.2009.12-07-675. URL `http://dx.doi.org/10.1162/neco.2009.12-07-675`.

M. A. Whittington, R. D. Traub, and J. G. Jefferys. Synchronized oscillations in interneuron networks driven by metabotropic glutamate receptor activation. *Nature*, 373(6515):612–615, Feb 1995. doi: 10.1038/373612a0. URL `http://dx.doi.org/10.1038/373612a0`.

D. J. Willshaw and C. von der Malsburg. How patterned neural connections can be set up by self-organization. *Proc. R. Soc. Lond. B Biol. Sci.*, 194(1117):431–445, Nov 1976.

M. A. Wilson and B. L. McNaughton. Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679, Jul 1994.

L. Wiskott. *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*, volume 53. Verlag Harri Deutsch, Thun, Frankfurt am Main, 1995.

L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE T. Pattern. Anal.*, 19(7):775–779, 1997.

Philipp Wolfrum, Christian Wolff, Jörg Lücke, and Christoph von der Malsburg. A recurrent dynamic model for correspondence-based face recognition. *J. Vis.*, 8(7):34.1–3418, 2008. doi: 10.1167/8.7.34. URL `http://dx.doi.org/10.1167/8.7.34`.

Thilo Womelsdorf, Jan-Mathijs Schoffelen, Robert Oostenveld, Wolf Singer, Robert Desimone, Andreas K Engel, and Pascal Fries. Modulation of neuronal interactions through neuronal synchronization. *Science*, 316(5831):1609–1612, Jun 2007. doi: 10.1126/science.1139597. URL `http://dx.doi.org/10.1126/science.1139597`.

Xiaohui Xie, Richard H R Hahnloser, and H. Sebastian Seung. Selectively grouping neurons in recurrent networks of lateral inhibition. *Neural Comput.*, 14(11):2627–2646, Nov 2002. doi: 10.1162/089976602760408008. URL `http://dx.doi.org/10.1162/089976602760408008`.

L. Xu, A. Krzyzak, and E. Oja. Rival penalized competitive learning for clustering analysis, rbf net, and curve detection. *IEEE Trans. Neural Netw.*, 4(4):636–649, 1993. doi: 10.1109/72.238318. URL `http://dx.doi.org/10.1109/72.238318`.

Lei Xu. Byy harmony learning, structural rpcl, and topological self-organizing on mixture models. *Neural Netw.*, 15(8-9):1125–1151, 2002.

Xiangmin Xu and Edward M Callaway. Laminar specificity of functional input to distinct types of inhibitory cortical neurons. *J. Neurosci.*, 29(1):70–85, Jan 2009. doi: 10.1523/JNEUROSCI.4104-08.2009. URL `http://dx.doi.org/10.1523/JNEUROSCI.4104-08.2009`.

Haishan Yao, Lei Shi, Feng Han, Hongfeng Gao, and Yang Dan. Rapid learning in cortical coding of visual scenes. *Nat. Neurosci.*, 10(6):772–778, Jun 2007. doi: 10.1038/nn1895. URL `http://dx.doi.org/10.1038/nn1895`.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Generalized belief propagation. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *NIPS*, pages 689–695. MIT Press, 2000. URL `http://dblp.uni-trier.de/db/conf/nips/nips2000.html#YedidiaFW00`.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. pages 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003. ISBN 1-55860-811-7.

Yumiko Yoshimura and Edward M Callaway. Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nat. Neurosci.*, 8(11):1552–1559, Nov 2005. doi: 10.1038/nn1565. URL `http://dx.doi.org/10.1038/nn1565`.

Yumiko Yoshimura, Jami L M Dantzker, and Edward M Callaway. Excitatory cortical neurons form fine-scale functional networks. *Nature*, 433(7028):868–873, Feb 2005. doi: 10.1038/nature03252. URL `http://dx.doi.org/10.1038/nature03252`.

M. P. Young and S. Yamane. Sparse population coding of faces in the inferotemporal cortex. *Science*, 256(5061):1327–1331, May 1992.

Yong-Chun Yu, Ronald S Bultje, Xiaoqun Wang, and Song-Hai Shi. Specific synapses develop preferentially among sister excitatory neurons in the neocortex. *Nature*, 458(7237):501–504, Mar 2009. doi: 10.1038/nature07722. URL `http://dx.doi.org/10.1038/nature07722`.

Alan L. Yuille and Norberto M. Grzywacz. A winner-take-all mechanism based on presynaptic inhibition feedback. *Neural Comput.*, 1(3):334–347, 1989. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1989.1.3.334.

Wei Zhang and David J Linden. The other side of the engram: experience-driven changes in neuronal intrinsic excitability. *Nat. Rev. Neurosci.*, 4(11):885–900, Nov 2003. doi: 10.1038/nrn1248. URL `http://dx.doi.org/10.1038/nrn1248`.

Zuohua Zhang and Dana H. Ballard. Distributed synchrony. *Neurocomputing*, 44-46: 715 – 720, 2002. ISSN 0925-2312. doi: 10.1016/S0925-2312(02)00463-0. URL `http://www.sciencedirect.com/science/article/B6V10-45F90MH-C/2/7110ce966c2a11a68b05ca428ece5cd1`.

Long Zhu, Yuanhao Chen, and Alan Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1029–1043, 2010. ISSN 0162-8828. doi: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.65.

Index

active learning, 128
adaptation (fatigue), 58

bag of features, 17, 18
bars test, 54, 55
Bayesian generative learning, 104, 112
belief propagation, 23, 129
binding, 62, 78, 92, 106
bistability, 31

combinatorial explosion, 9, 16
competitive learning, 25, 45, 51, 59, 61, 129
  rival penalized learning, 60
competitive processing, 11, 24, 25, 27, 32, 36, 39, 43, 58, 103, 104, 128
  and cooperation, 66, 77, 78, 103, 105
  as attentional mechanism, 88, 109
  graded amplification, 35, 36, 60, 63, 105
  selection and amplification, 37, 38, 60, 66, 77, 105
  two-phase winner-take-all (WTA) computation, 34, 39, 55, 56, 60, 62, 63, 105
compositional object representation, 8, 9, 22, 23, 53, 65, 72, 77, 103, 123
  combinatorial power, 65, 111
  generative nature of, 65, 75, 104, 108, 123
  reusable vocabularies, 21, 23, 26, 53, 65, 72, 74, 111, 123
context-dependent processing, *see* contextual processing
contextual connectivity, 10, 36, 66, 72, 73, 105, 118, 120
  generative nature, 88, 109
  improving generalization capability, 84, 110
contextual processing, 10, 20, 21, 38, 66, 78, 103, 105, 110
  coordination, 79, 80
  error signal, 80, 105
cortical cluster, column, *see* cortical module
cortical microcircuits, 6, 59, 128
  cortical layers, 6, 36
  fine-scale excitatory subnetworks, 6, 25, 27, 28, 59
  minicolumn, 6
cortical module, 7, 24, 59, 68, 102
  model, 25, 27
curse of dimensionality, 16

dead unit problem, 60
decision cycle, *see* gamma cycle
dopamine, 128
  role in reward signaling, 128
dynamic link architecture, DLA, 124

elastic graph matching, 18, 20, 22
explicit memory, 3

feed-forward inhibition, 39
feed-forward sweep, 20, 109, 110

Gabor filtering, 17, 47
gamma cycle, 11, 24, 27, 36, 38, 55, 66, 106, 110, 128
  as atom of discrete processing, 61, 106, 123
  cycle skipping, 61
  embedding in the theta rhythm, 106, 128
  probabilistic coding, 56, 57, 61, 90
  role in competitive learning, 60, 62
  role in competitive processing, 56, 61, 77, 106
generalization error, 68, 82
generative model, 109
grandmother neuron, 108
graphical model, 22, 129

HMAX architecture, 18, 22

implicit memory, 3
independent component analysis, ICA, 17
independent subspace analysis, ISA, 104
intrinsic plasticity, 10, 26, 41, 93

  - homeostatic excitability regulation, 12, 26, 41, 51, 58, 60, 74, 93, 104, 114, 119, 124
  - improving generalization capability, 93, 110, 113, 120

invariance problem, 17
  - transformation invariance, 124, 126

lateral geniculate nucleus, LGN, 54
localist representation, 108

medial temporal lobe, MTL, 3, 4, 108
memory consolidation, 4, 12, 72
memory recall, 3–5, 9, 14, 110, 120, 124
  - failed vs. successful, 111

memory replay, 14, 24, 92, 104, 109, 113–115, 124, 127
  - against catastrophic forgetting, 127
  - during sleep, 109, 120

memory trace, 10, 11, 53, 66, 72–74, 103, 121, 126, 127
  - consistency, 75
  - generative nature, 86, 104
  - interference, 108

mixture of Gaussians model, 61

Neocognitron, 19

off-line memory reprocessing, 12, 14, 15, 24, 92, 109, 113, 114, 120, 121, 126
  - improving generalization capability, 117, 119, 124

ongoing cortical activity, 110

Pandemonium model, 16
perceptual learning, 5, 10, 103
prefrontal cortex, PFC, 68, 108
priming, 86, 89, 109
principal component analysis, PCA, 17, 22

recurrent connectivity, 6, 7, 9, 20, 24, 35, 66, 84, 106, 110, 120, 123
reinforcement learning, 128
REM sleep, 13, 121
reward signal, 128

scale-invariant feature transform, SIFT, 17
semisupervised learning, 127
sleep stages, 12
slow wave sleep, SWS, 13, 120
  - and memory consolidation, 13, 120

sparse coding, 62, 107, 108
sparseness
  - hard sparseness, 109
  - of activity, 78, 104, 107, 108
  - synaptic, 70, 73, 108

stability-plasticity dilemma, 126
  - catastrophic forgetting, 126

synaptic plasticity, 10, 41, 103
  - Artola-Bröcher-Singer rule, ABS, 42
  - bidirectional nature, 25, 41, 42, 60, 62, 104, 127, 128
  - Bienenstock-Cooper-Munro rule, BCM, 42
  - caricaturizing effect, 26
  - competition, 43, 51, 60, 105
  - dopaminergic modulation, 128
  - long-term depression, LTD, 43
  - long-term potentiation, LTP, 43
  - metaplasticity, 42
  - role of the gamma rhythm, 11, 62, 106
  - short-term depression, facilitation, 129
  - sliding threshold, 25, 42

synaptic scaling, 44
  - during sleep, 121

synaptic separation, 35, 36, 63, 105

transient attractor dynamics, 59, 63, 89, 108

unsupervised learning, 1, 15, 23, 26, 47, 55, 66, 68, 111, 123
  - life-long learning, 61, 112, 124, 126
  - of sequences, 126
  - of transformation invariance, 125

visual cortex, 5, 65
  - fusiform face area, FFA, 7
  - inferotemporal cortex, IT, 7, 68, 108, 111
  - occipital face area, OFA, 7
  - primary visual cortex, V1, 7, 17, 55
  - speed of object recognition, 110
  - ventral visual pathway, 7

winner-take-all, WTA, *see* competitive processing
working memory, 86, 107
  - multi-item working memory, 128
  - persisting activity, 85

List of Figures

1.1 Sea slug and BigDog 2
1.2 Brodmann areas and mapping of cortical pathways 4
1.3 Medial temporal lobe and hippocampal formation 5
1.4 Various forms of memory 6
1.5 Cortical microcircuits, cell and synaptic specificity 7
1.6 Functional excitatory fine-scale subnetworks 8
1.7 Modules and neuron selectivity along the ventral visual pathway 9
1.8 Parallel processing pathways in visual cortex 10
1.9 Processing hierarchy of the ventral pathway of the visual cortex 11
1.10 Vocabulary of universal elements and combinatorial representational power 12
1.11 Role of context in visual recognition 12
1.12 Multiple forms of plasticity in a cortical circuit 13
1.13 Generation of local rhythms in cortical microcircuit 14
1.14 Sleep and memory function 15
1.15 Pandemonium model of hierarchical processing 16
1.16 Invariance problem in vision 17
1.17 Gabor filtering and low-level visual features 18
1.18 HMAX architecture 19
1.19 Elastic graph matching 21
1.20 Hierarchical graphical models 22

2.1 A cortical module, schematic illustration 26
2.2 Vector field of the winner unit. 28
2.3 Vector field of the challenger unit. 30
2.4 Comparing the behavior of winner and challenger unit. 31
2.5 Vector field in the phase space of a two-units module. 33
2.6 Bifurcation diagram and eigenvalues. 34
2.7 2-Phase WTA coding in a single module. 37
2.8 Flexibility of 2-phase WTA coding. 38
2.9 Module circuit diagram. 40
2.10 Bidirectional plasticity and competitive learning. 43
2.11 Face databases used for learning. 46
2.12 Gabor filter bank. 47
2.13 Learning procedure for single module. 48
2.14 Learning procedure for distributed isolated modules. 49
2.15 Learning error course and synaptic structure formation for single module. 50
2.16 Unit selectivity in distributed isolated modules, 120 persons in the training set 52
2.17 Learning error for distributed isolated modules, 120 persons in the training set 52

2.18 Unit selectivity in distributed isolated modules, 1000 persons in the training set 53
2.19 Learning error for distributed isolated modules, 1000 persons in the training set 54
2.20 Unsupervised component extraction : bars test. 55
2.21 Unsupervised formation of an overcomplete basis for natural image patches. 56
2.22 Gamma coding scheme, crossed bars example. 57
2.23 Gamma coding scheme, natural image patch example. 58

3.1 Hierarchical visual memory model 67
3.2 Learning procedure for the memory network. 69
3.3 Network structure formation, lower layer bottom-up and lateral connectivity 71
3.4 Network structure formation, higher layer bottom-up and top-down connectivity 72
3.5 Time course of structure organization. 73
3.6 Excitability regulation time course 74
3.7 Average excitability time course 75
3.8 Unit selectivity on the vocabulary layer of the memory network 76
3.9 Unit selectivity on the identity layer of the memory network 76
3.10 Connectivity consistency 77
3.11 Activity formation during memory recall 78
3.12 Activity formation and signal cooperation 79
3.13 Improvement of afferent signal coordination during the learning 80
3.14 Signatures of successful and failed recall 81
3.15 Learning error for the identity layer, feed-forward vs. recurrent architecture. 82
3.16 Learning error for the parts layer, 40 persons in the training set 83
3.17 Locking activity after stimulus removal 85
3.18 Memory recall induced by partial bottom-up cues 87
3.19 Memory recall induced by top-down identity cue 88
3.20 Memory recall and encoding over multiple gamma cycles 89
3.21 Memory replay in absence of external stimuli 90
3.22 Coherent storage-based replay in off-line regime. 91
3.23 Improving recognition performance on blocks of novel data, time course for different views 93
3.24 Excitability regulation during exposure to novel data blocks 94
3.25 Recognition performance of the recurrent network configuration before and after excitability regularization 95
3.26 Effect of different regularization procedures on performance of recurrent network configuration 95
3.27 Recognition performance of the feed-forward network configuration before and after excitability regularization 96
3.28 Comparison between performance of feed-forward and recurrent configuration after the excitability regularization 96
3.29 Network structure formed by learning 120 persons 97
3.30 Network structure formed by learning 1000 persons 98
3.31 Unit selectivity on the vocabulary layer, learning 120 persons 99
3.32 Unit selectivity on the vocabulary layer, learning 1000 persons 100
3.33 Learning error for the parts and identity layer, 120 persons in the training set 100
3.34 Learning error for the parts layer, 1000 persons in the training set 101
3.35 Embedding fast gamma rhythms in slower theta rhythm 107

4.1 Activity formation in different off-line regimes 115
4.2 Excitability regulation during the sleep-like state, downregulation regime 116
4.3 Excitability regulation during sleep-like state, balanced regulation regime 117
4.4 Recognition performance of the recurrent network configuration before and after the off-line regime 118
4.5 Effect of different off-line regimes on performance of recurrent network configuration . 118
4.6 Recognition performance of the feed-forward network configuration before and after the off-line regime 119
4.7 Comparison between performance of feed-forward and recurrent configuration after the off-line regime 119

5.1 Learning translation invariant object representation 125
5.2 Extensions of bidirectional plasticity mechanism 127

List of Tables

2.1 Model parameters overview 44

3.1 Generalization block error on the identity layer, fully recurrent versus purely feed-forward memory configuration. 83
3.2 Generalization block error on the parts layer, fully recurrent versus purely feed-forward memory configuration. 84
3.3 Generalization block error for gender on the parts layer, fully recurrent versus purely feed-forward memory configuration. 84
3.4 Performance improvement on blocks of novel data 93
3.5 Generalization block error for the identity and parts layer after learning 120 persons . . 99
3.6 Generalization block error for the parts layer after learning 1000 persons 101

4.1 Oscillation amplitudes for different sleep regimes 115

Kurzfassung

Gegenwärtig besteht immer noch ein enormer Abstand zwischen der Lernfähigkeit von künstlichen und biologischen Informationsverarbeitungssystemen. Dieser Abstand ließe sich durch eine bessere Einsicht in die höheren Funktionen des Gehirns wie Lernen und Gedächtnis mit Sicherheit verringern. Der visuelle Kortex etwa wird als der Ort vermutet, wo die komplexen visuellen Inhalte gelernt und im Langzeitgedächtnis aufbewahrt werden. Zudem wird angenommen, dass der visuelle Kortex im Wahrnehmungsvorgang die Objekte innerhalb kürzester Zeit in ihre Bestandteile zerlegt, um sie entlang der hierarchischen Verarbeitungspfade durch die Komposition von Elementen niedrigerer Komplexität darzustellen und bei Bedarf so zu speichern. Wie eine derartige, kompositionell-hierarchische Gedächtnisstruktur durch die visuelle Erfahrung zustande kommen kann, ist noch weitgehend ungeklärt.

Um einen Fortschritt in dieser Fragestellung zu erzielen, wird hier ein funktionelles Modell der Entstehung von Gedächtnisstruktur aus Erfahrung mit natürlichen Stimuli vorgestellt und untersucht. Das Modell basiert auf einem hierarchischen, rekurrenten neuronalen Netzwerk. Das Netzwerk implementiert hypothetische Mechanismen der kortikalen Verarbeitung und Adaptation, welche für die Selbstorganisation der Verbindungsstruktur im Netzwerk verantwortlich sind. Die Architektur des Netzwerkes besteht aus zwei nacheinander geschalteten Schichten. Jede Schicht beherbergt eine Anzahl von verteilten Modulen, die sowohl vorwärtsgerichtet als auch rekurrent innerhalb und zwischen den Schichten miteinander vernetzt sind. Jedes Modul wird mit einem lokal umgrenzten kortikalen Cluster identifiziert, bestehend aus mehreren funktionell separaten Subnetzwerken von eng gekoppelten, exzitatorischen Neuronen.

Ein solches Modul ist imstande, aus den ankommenden Signalen unüberwacht zu lernen und eine geeignete Repräsentation für den lokal zugänglichen Eingaberaum zu bilden. Die fortlaufende Verarbeitung im Netzwerk setzt sich zusammen aus diskreten Fragmenten, genannt Entscheidungszyklen, die man hypothetisch mit den schnellen Rhythmen im gamma-Frequenzbereich in Verbindung setzen kann, wie sie im Kortex beobachtet werden. Die Zyklen sind synchronisiert zwischen den verteilten Modulen. Innerhalb eines Zyklus wird in Modulen lokal umgrenzt eine kompetitive, winner-take-all-ähnliche Operation durchgeführt, wobei die Kompetitionsstärke im Laufe des Zyklus anwächst. Diese kompetitive Operation wählt in Abhängigkeit von den ankommenden Signalen eine sehr kleine Anzahl von Einheiten aus der potentiell verfügbaren Menge und verstärkt ihre Aktivität auf Kosten der anderen, um den aktuell präsentierten Reiz in der Netzwerkaktivität abzubilden.

Ausgestattet mit adaptiven Mechanismen der bidirektionalen synaptischen Plastizität und der homöostatischen Regulierung der neuronalen Aktivität, wird das Netzwerk mit natürlichen Gesichtsbildern von verschiedenen Personen konfrontiert. Die Bilder werden der unteren Netzwerkschicht, je ein Bild pro Zyklus, als Ansammlung von Gaborfilterantworten aus lokalen Gesichtslandmarken zugeführt, ohne jegliche Zusatzinformation über die Personenidentität zur Verfügung zu stellen. Im Laufe der unüberwachten Lernprozedur schafft das Netzwerk die Verbindungsstruktur so aufzubauen, dass die Gesichter aller vorgeführten Personen ihren Platz in Form von dünn besiedelten Gedächtnisspuren im Netzwerk finden. Hierzu formt das Netzwerk simultan Vokabulare aus den wiederverwertbaren, lokalen Gesichtselementen auf der unteren Netzwerkschicht, verknüpft assoziativ jene Elemente miteinander, die Teile derselben Gesichtsidentität kodieren, entwickelt Identitätssymbole für die verknüpften Kompositionen auf der höheren Netzwerkschicht und projiziert diese Information zurück auf die Vokabulare der unteren Schicht in generativer Weise. Dieses Lernen entspricht der gleichzeitigen Ausdifferenzierung von vorwärtsgerichteten (bottom-up) und rekurrenten (lateral, top-down) synaptischen Verbindungen innerhalb und zwischen den Schichten. Im ausgereiften Verbindungszustand werden die einzelnen Gesichter so als Komposition ihrer Bestandteile entlang der Netzwerkschichten abgespeichert. Dank der generativen Art der ausgebildeten Gedächtnisstruktur reichen schon allein das höhere

Identitätssymbol oder eine kleine Teilmenge von entsprechenden Gesichtselementen, um die vollständige kompositionelle Beschreibung der gespeicherten Gesichter in Form all ihrer verknüpften Bestandteile aus dem Gedächtnis abzurufen. In der Testphase stellt das Netzwerk erfolgreich seine Fähigkeit unter Beweis, sowohl die Identität als auch das Geschlecht von Personen aus vorher nicht gezeigten Alternativansichten ihrer Gesichter zu erkennen.

Eine bemerkenswerte Eigenschaft der entstandenen Gedächtnisarchitektur ist ihre Fähigkeit, ohne Darbietung von externen Stimuli spontan Aktivitätsmuster zu generieren und die im Gedächtnis abgelegten Inhalte in diesem schlafähnlichen öff-lineRegime wiederzugeben. Interessanterweise ergibt sich aus der Schlafphase ein direkter Vorteil für die Gedächtnisfunktion. Dieser Vorteil macht sich durch eine drastisch verbesserte Erkennungsrate nach der Schlafphase bemerkbar, wenn das Netwerk mit den zuvor nicht dargebotenen Alternativansichten von den bereits bekannten Personen konfrontiert wird. Die Leistungsverbesserung nach der Schlafphase ist umso deutlicher, je stärker die Alternativansichten vom Original abweichen. Dieser positive Effekt ist zudem komplett unabhängig von der synapsenspezifischen Plastizität und kann allein durch die synapsenunspezifische, homöostatische Regulation der Aktivität im Netzwerk erklärt werden.

Das entwickelte Netzwerk demonstriert so eine im Bereich der neuronalen Modellierung bisher nicht gezeigte Funktionalität. Es kann unüberwacht eine Gedächtnisdomäne für kompositionelle, generative Objektrepresentation durch die Erfahrung mit natürlichen Bildern sowohl im reizgetriebenen, wachähnlichen Zustand als auch im reizabgekoppelten, schlafähnlichen Zustand formen und verwalten. Diese Funktionalität bietet einen vielversprechenden Ausgangspunkt für weitere Studien, die die neuronalen Lernmechanismen des Gehirns ins Visier nehmen und letztendlich deren konsequente Umsetzung in technischen, adaptiven Systemen anstreben.

Über Selbstorganisation der hierarchischen Gedächtnisstruktur für kompositionelle Objektrepräsentation im visuellen Kortex
(Zusammenfassung in deutscher Sprache)

Das Kapitel bietet eine deutsche Zusammenfassung der vorgelegten Arbeit. Die Originalunterteilung in die Hauptkapitel wird beibehalten. Die feinere Unterteilung in Unterabschnitte wird zum großen Teil verändert, um die Lesbarkeit der Kurzfassung zu verbessern. Englische Fachbegriffe, die auch eine feste Verwendung in der deutschen Sprache finden, werden unverändert übernommen.

1 Einführung und Motivation

Unter den zahlreichen ungelösten Rätseln des Gehirns nimmt die Fähigkeit, aus Erfahrung zu lernen und das Gelernte langfristig im Gedächtnis aufzubewahren, eine besondere Stellung ein. Jeden Tag wird unser Gedächtnis mit neuen Inhalten gefüllt, während manches in Vergessenheit gerät. Dies geschieht oft ohne besondere Anstrengung, wie wenn wir einer bisher unbekannten Person begegnen, ihr Gesicht behalten und es später wiedererkennen. Dabei ist es nicht nötig, explizite Instruktionen zu erhalten, wie man das Lernen zu bewerkstelligen hat, es vollzieht sich größtenteils automatisch. Die grundlegenden Formen von Lernen und Gedächtnis sind auch bei primitiven Organismen wie der Meeresschnecke vorhanden und können auf neuronaler Ebene festgemacht werden. Bei neuronalen Mechanismen des Lernens handelt es sich offenbar um eine sehr alte Vorrichtung, die sich bereits früh im Laufe der Evolution entwickelte, um den Organismus an die Begebenheiten der sich ständig verändernden Umwelt anzupassen.

Im Gegensatz zu den biologischen Organismen haben die aktuellen technischen Systeme fundamentale Schwierigkeiten, durch Lernen ihre Funktion an die veränderlichen Aufgabenanforderungen anzupassen. Vor allem das Lernen im unüberwachten Modus stellt für die technischen Systeme eine große Herausforderung dar. Diese Umstände verhindern den entscheidenden Fortschritt in vielen klassischen Bereichen der künstlichen Intelligenz, wie in der Verarbeitung der natürlichen Sprache oder in der maschinellen Objekterkennung. Im Falle der visuellen Objekterkennung beschränken sich die maschinellen Ansätze zumeist auf sehr eng definierte Szenarien. Ein solches Szenario ist zum Beispiel gegeben bei der visuellen Kontrolle der bestimmten Bauteile, die nacheinander auf dem Laufband ankommen. In einer derartigen Situation werden die Mehrdeutigkeiten bei der Interpretation der aufgenommenen Bilder auf ein Minimum reduziert, da die Bildaufnahmebedingungen, die Position und die Art der Objekte streng kontrolliert sind.

Die streng kontrollierten Bedingungen sind jedoch eher eine Ausnahme für die komplexen Wahrnehmungssituationen, mit denen sich die technischen Systeme auseinander setzen müssen bei solchen Aufgaben wie zum Beispiel der Steuerung eines Fahrzeuges im Stadtverkehr. Die Vielzahl der Objekte, die Variationen in ihrer Erscheinung und die Vielfalt der möglichen Szenen, in denen sie potentiell auftauchen, verlangen dabei nach einem System, das weitgehend unüberwacht die günstigen Repräsentationen für die relevanten Objekte anhand der Beispiele aus den Bilddaten zu lernen vermag. Das Lernen aus Beispielen ist genau die Stärke des Gehirns. Insbesondere der visuelle Kortex ist darauf spezialisiert, die Identität eines Objektes aus nur wenigen Beispielen zu lernen. Es wird entsprechend vermutet, dass dort die visuelle Erfahrung gesammelt und im Langzeitgedächtnis aufbewahrt wird. Dabei ist es aufschlussreich, welcher Strategien sich der visuelle Kortex bedient, um die visuellen Objekte zu repräsentieren.

Eine Hypothese besagt, dass die Objekte im visuellen Kortex innerhalb kürzester Zeit entlang der hierarchischen Verarbeitungspfade in ihre Bestandteile zerlegt werden. Ein komplexes Objekt kann so auf die Komposition vieler Elemente von weitaus niedrigerer Komplexität als das Objekt selbst reduziert werden. Für eine solche effiziente Dekomposition ist die vorangegangene Erfahrung mit Vielzahl von visuellen Stimuli notwendig. Die visuellen Elemente von verschiedener Komplexität werden dabei vermutlich den universellen Vokabularen entnommen, die lokal entlang der hierarchischen Verarbeitungspfade geformt wurden. Die Elemente sind in einem solchen Schema wiederverwertbar, das heißt, sie können von verschiedenen gespeicherten Objekten geteilt werden. Diese Wiederverwertbarkeit verleiht der Repräsentation eine kombinatorische Ausdruckskraft, da ein beliebiges, auch bisher unbekanntes Objekt durch eine kleine Untermenge von vielen Elementen dargestellt ("kombiniert") werden kann. Diese Kombination liefert auch dann umgehend eine Gedächtnisspur für das neue Objekt. Der wichtige Bestandteil vom Lernen ist dabei wohl auch die Erstellung von Verknüpfungen zwischen den visuellen Elementen. Solche Verknüpfungen halten die statistischen Relationen höherer Ordnung auf verschiedenen Beschreibungsebenen fest, wie zum Beispiel die Zugehörigkeit von Elementen zu einem bekannten Objekt oder deren häufig vorkommende räumliche Anordnung.

Der Vergleich von der kompositionell-hierarchischen Repräsentation, die wahrscheinlich vom visuellen Kortex verwendet wird, mit den Repräsentationen, die in maschinellen Sehsystemen Verwendung finden, offenbart eine sogenannte *Repräsentationsarmut* auf der maschinellen Seite. Diese Art von Defizit besteht, wenn die verwendete Objektrepräsentation von Anfang an fundamentale Einschränkungen beinhaltet, die eine flexible Verarbeitung der Objekte gemäß ihrer kompositionellen Natur nicht mehr zulassen. Eine Reihe von aktuellen Ansätzen zur Objekterkennung aus dem Bereich der probabilistischen graphischen Modelle schlägt eine vielversprechende Richtung ein, um die Repräsentationsarmut weitgehend zu beheben. Es wird ebenfalls intensiv untersucht, wie solche hierarchischen graphischen Modelle aus natürlichen Bilddaten unüberwacht gelernt werden können. Immer noch in ihren Anfängen stecken die Ansätze zur Abbildung der in den graphischen Modellen verwendeten Operationen auf die Mikroschaltungen und Netzwerke im visuellen Kortex.

Das Ziel dieser Arbeit ist es, die hypothetischen neuronalen Mechanismen des Lernens im visuellen Kortex in einem funktionellen Modell zu implementieren, welches einen weiteren Fortschritt im unüberwachten Lernen der kompositionell-hierarchischen Objektrepräsentation erzielen soll. Hierzu wird eine neuronale mehrschichtige Netzwerkarchitektur vorgestellt, die allein durch Erfahrung mit natürlichen Gesichtsbildern eine Gedächtnisdomäne für die präsentierten Personen aufbauen muss. Im Konkreten, folgende Funktionalität wird vom System gefordert:

- Das System muss die kompositionell-hierarchische Repräsentation für die gespeicherten Gesichtsobjekte verwenden. Das heißt, es muss gleichzeitig die Vokabulare von wiederverwertbaren Gesichtselementen lernen, diese Elemente je nach Zugehörigkeit zum selben Gesicht assoziativ verknüpfen und die höheren Symbole für die Personenidentität aufgrund dieser Verknüpfungen bilden.

- Die Repräsentation sollte zudem eine generative Natur besitzen. Es sollte möglich sein, allein anhand des höheren Identitätssymbols oder der kleinen Untermenge der entsprechenden Elemente die vollständige Beschreibung eines gespeicherten Gesichtes als Komposition all seiner Teile aus dem Gedächtnis abzurufen.

- Das System sollte unüberwacht lernen. Die Bilder sollten ohne Zusatzhinweise über die Identität der abgebildeten Personen präsentiert werden.
- Das Netzwerk sollte gleichzeitig alle Verbindungen - vorwärtsgerichtete bottom-up und rekurrente laterale und top-down - innerhalb und zwischen den Schichten ausbilden. Dabei dürfen nur neuronal plausible Mechanismen verwendet werden.
- Das unüberwachte Lernen sollte selbst-stabilisierend und permanent sein. Es sind keine per Hand definierten Haltbedingungen für das Lernen erlaubt, die ab einem bestimmten Zeitpunkt das Lernen im Netzwerk einfach einfrieren würden.
- Das System sollte dazu imstande sein, ein im Gedächtnis abgelegtes Gesicht innerhalb einer kurzen Zeit abzurufen. Die Abrufzeit sollte vergleichbar sein mit den Vorhersagen der psychophysiologischen Experimente bezüglichen der ultraschnellen Objekterkennung.

2 Modell eines elementaren kortikalen Moduls

In diesem Kapitel wird ein Modell vom elementaren kortikalen Modul vorgestellt und auf seine Tauglichkeit zum unüberwachten Lernen in verschiedenen Aufgabenstellungen getestet. Ein solches Modul wird mit einem lokal umgrenzten kortikalen Cluster identifiziert, bestehend aus mehreren funktionell separaten Subnetzwerken von eng gekoppelten, exzitatorischen Neuronen. Solche Cluster wurden in einer Reihe von neurophysiologischen Experimenten ausgemacht. Das Modul ist primär darauf ausgerichtet, als Knoten in der beabsichtigten Netzwerkarchitektur zu agieren. Entsprechend dieser Rolle zielt der Entwurf darauf ab, dem Modul zwei wesentliche Funktionen zu verleihen. Zum einen sollte das Modul die aus verschiedenen Stufen der Netzwerkhierarchie lokal eintreffenden Signale zum Lernen der synaptischen Verbindungsstruktur und somit zum Aufbau der Gedächtnisspuren verwenden können. Diese Verbindungsstruktur sollte das Vokabular von wiederverwertbaren Elementen für den lokal zugänglichen Eingaberaum formen und gleichzeitig die Verknüpfung von Elementen aus verschiedenen verteilten Vokabularen ermöglichen. Zum anderen sollte das Modul imstande sein, die lokal verfügbare Information über den globalen Stimulus in Abstimmung mit anderen Modulen des Netzwerkes schnell zu interpretieren.

2.1 Neuronale Mechanismen für unüberwachtes kompetitives Lernen

Das Modul implementiert eine Form des unüberwachten Lernens, die als eine erweitere Version vom kompetitiven Lernen verstanden werden kann. In der klassischen Version vom kompetitiven Lernen wird aus einer Menge von Einheiten, die mit dem selben Eingaberaum beschäftigt sind, für je eine Eingabe eine Gewinner-Einheit bestimmt, die auf die aktuelle Eingabe am stärksten reagiert. Die Gewinnereinheit wird aktiv auf Kosten der anderen, die deaktiviert bleiben. Diese Gewinnereinheit darf dann auch ihre synaptischen Gewichte entsprechend der Eingabe modifizieren. Das Modul besteht ebenfalls aus einer Anzahl von Einheiten, die den hypothetischen exzitatorischen Subnetzwerken zuzuordnen sind. Die Einheiten sind selbst-erregend, teilen den gleichen afferenten Eingaberaum und sind durch eine gemeinsame laterale Hemmung gebunden.

Die Interaktion zwischen der Selbsterregung und der lateralen Hemmung bestimmt den kompetitiven Charakter der Aktivitätsdynamik von Moduleinheiten. Grundlegend für die Operation des Moduls ist die Aufteilung der fortlaufenden Verarbeitung in diskrete periodische Fragmente, die hypothetisch den Zyklen von schnellen kortikalen Rhythmen im gamma-Frequenzbereich (40Hz) entsprechen. Im Laufe eines solchen Fragments, hier auch Entscheidungszyklus, gamma-Zyklus oder einfach Zyklus genannt, wachsen sowohl die Stärke der lateralen Hemmung zwischen den Einheiten als auch ihre Selbsterregung, die am Zyklusanfang beide niedrig sind. Die anwachsende Hemmung führt an einem bestimmten kritischen Punkt innerhalb des Zyklus dazu, dass nur eine einzige Einheit aktiv bleiben darf, während die übrigen deaktiviert werden. Der übrig bleibende Gewinner erreicht eine vom aktuellen Input abhängige Aktivitätsstufe. Eine Abfolge aus mehreren Zyklen entspricht dann einer sich wiederholenden Operation der Auswahl und der graduellen Verstärkung einer Gewinnereinheit pro Zyklus.

Die vom Modul verwendete Kodierung des eintreffenden Inputs kann so innerhalb eines Zyklus als eine 2-Phasen Winner-take-all (2-Phasen WTA) Kodierung angesehen werden. Die erste Phase findet in der frühen Zyklushälfte statt. Dort können mehrere Einheiten den ankommenden Input mit ihrer abgestuften Aktivität abbilden. Dies entspricht einer soft Winner-take-all (sWTA) Operation. In der zweiten Phase, die den späten Zyklus einnimmt, wird der aktuelle Input allein durch die abgestufte Aktivität des Gewinners repräsentiert, was wiederum einer graduellen hard Winner-take-all (hWTA) Operation entspricht. Die zeitliche Anordnung und der Verlauf der beiden Phasen kann potentiell nach Belieben durch die Anpassung der Rhythmen moduliert werden. Eine solche flexible WTA Operation bietet die Grundlage für eine gelungene Kommunikation zwischen den Modulen im Netzwerk, das im nächsten Kapitel vorgestellt wird. Hierbei wird auch zum Tragen kommen, dass die eintreffenden afferenten Signale je nach ihrer Herkunft aus der Netzwerkhierarchie im Modul getrennt werden und auf verschiedene Weise die Aktivität der Einheiten beeinflussen.

Die eingesetzten adaptiven Mechanismen nutzen ebenfalls die Vorteile der 2-Phasen WTA Kodierung, um das Modul an die Statistik der eintreffenden Signale anzupassen. Diese Mechanismen sind die bidirektionale synaptische Plastizität und die homöostatische Regulation der neuronalen Aktivität. Die synaptische Plastizität modifiziert die synaptischen Gewichte in Abhängigkeit von der pre- und postsynaptischen Aktivität. In Anlehnung an bekannte Regeln wie BCM oder ABS wird die bidirektionale Modifikation bestimmt durch die gleitenden Schwellen, welche die entsprechenden Modifikationszonen postsynaptisch definieren. Je nach der Stärke der postsynaptischen Aktivierung bleibt die Synapse dann entweder unverändert, wird verringert (Depression, LTD) oder verstärkt (Potentierung, LTP). Synaptische Strukturbildung erhält durch die Kombination dieser Regel mit der WTA Operation ebenfalls einen kompetitiven Charakter. Der kompetitive Effekt erlaubt es, auch zwischen sehr ähnlichen Mustern zu differenzieren und sie separat zu speichern. Dies wird möglich, da die Synapsen, die diskriminative Merkmale tragen, in ihrem Wachstum begünstigt werden, während die Synapsen, die gemeinsame Merkmale tragen, eher gedämpft werden.

Während die synaptische Plastizität für die Entstehung der Verbindungsstruktur verantwortlich ist, ist die homöostatische Regulation der Aktivität dazu da, die Einheiten möglichst gleichmäßig an der Strukturentstehung teilnehmen zu lassen. Die Regulation wird bewerkstelligt, indem die intrinsiche Erregung der Einheit herauf- oder herunterreguliert wird, wenn die Einheit zu wenig oder zu viel aktiv wird. Anders interpretiert, je nachdem wie oft eine Einheit der Zyklusgewinner wird, wird ihre Gewinnwahrscheinlichkeit so angepasst, dass im Durchschnitt über einen längeren Zeitraum aus mehreren Zyklen jede Einheit gleich oft der Gewinner sein kann. Im Vorausblick auf die Netzwerkfunktion soll dieser Mechanismus dann zur der ausgeglichenen Beteiligung aller Einheiten an der Bildung der Gedächtnisspuren führen.

2.2 Unüberwachtes Lernen mit einzelnen Modulen

Um seine Fähigkeit zu testen, durch unüberwachtes Lernen eine geeignete Basis für den lokal zugänglichen Eingaberaum zu bilden, wird das Modul mit zwei Aufgaben konfrontiert. Die erste Aufgabe besteht darin, ohne Überwachung aus den natürlichen Gesichtsbildern verschiedener Personen ein Vokabular von lokalen Gesichtselementen zu lernen. In der zweiten Aufgabe wird verlangt, aus Beispieleingaben die Basiskomponenten, aus denen diese Eingaben zusammengesetzt sind, unüberwacht zu extrahieren.

Für die erste Aufgabe wurden die Gesichtsbilder von verschiedenen Personen aus der Datenbank entnommen, um als Lernmaterial zu dienen. Die Gesichter sind dabei durch ein automatisiertes Verfahren mit Landmarken versehen, die bezüglich des Ortes auf dem Gesicht über Personen hinweg konsistent sind. In der Trainingsphase werden die Gesichtsbilder einem oder mehreren Modulen im unüberwachten Modus dargeboten, das heißt, ohne die Hinweise auf die Personenidentität zur Verfügung zu stellen. Die Darbietung erfolgt inkrementell, es wird je ein Bild pro Zyklus als Ansammlung von Gaborfilterantworten aus entsprechenden Landmarken präsentiert. Ein Modul hat dabei Sicht auf je eine Landmarke. Im Fall von mehreren Modulen unterhalten diese keine Verbindungen zueinander, sind also voneinander isoliert.

Den natürlichen Gesichtsbildern ausgesetzt, formen die Module für je eine Landmarke ein lokales Vokabular von Gesichtselementen. Diese entstandene Repräsentation kann dazu verwendet werden, die Identität und das Geschlecht der gezeigten Personen zu erkennen. Die Fehlerrate sinkt gegen Null auf dem Trainingsdatensatz sowohl für die Erkennung der Identität als auch für die Erkennung des Geschlechts der Personen. Ferner lässt es sich beobachten, dass dieselben Vokabularelemente von mehreren Personen gleichzeitig geteilt werden und zwar so, dass die Nutzlast für die verfügbaren Einheiten ausgeglichen ist. Die Wiederverwertung der Elemente für die Repräsentation von gelernten Gesichtern geschieht also auf eine wohl balancierte Weise.

In der zweiten Aufgabe werden zwei Szenarien behandelt. In dem einen Szenario werden dem Modul fortlaufend 10×10 Pixel große Ausschnitte aus den Grauwertbildern von Naturszenen dargeboten. Die Bilder werden im voraus durch DOG-Filterung (Difference of Gaussians) bearbeitet. Das zweite Szenario ist eine Instanz von einem sogenannten Balken-Test (bar test). Das Modul bekommt dabei 4×4 Pixel große Bilder dargeboten, wo jedes Pixel entweder schwarz oder weiß ist. Jedes dieser Bilder ist das Ergebnis einer Überlagerung von einer Anzahl an horizontalen oder vertikalen Balken von der Breite eines Pixels.

In beiden Szenarien schafft das Modul, eine geeignete Basis für die Darstellung der jeweiligen Eingaberäume zu finden. Für die Ausschnitte aus natürlichen Naturszenen wird ein übervollständiger Satz von Filtern gebildet, von denen einige an orientierte Bandpassfilter erinnern, die Gaborfunktionen ähneln, und andere wiederum lokalisierte DOG-Formen ohne spezifische Orientierung aufweisen. Rezeptive Felder mit solchen Profilen sind gut dokumentiert für die Neuronen im primären visuellen Kortexareal V1. Im Balken-Test gelingt es dem Modul, die einzelnen Balken, aus denen die Eingaben zusammengesetzt wurden, zu extrahieren.

Im Zustand nach dem Lernen kann man eine interessante Einsicht in die vom Modul verwendete Kodierung erhalten, wenn man die Antworten der Einheiten auf einen angelegten Stimulus über mehrere Zyklen beobachtet. Während innerhalb der einzelnen gamma-Zyklen die Gewinner hohe Aktivitätsniveaus erreichen, ist die Durchschnittsaktivität von den Einheiten, die an der Kodierung des Stimulus teilhaben, über mehrere Zyklen hinweg gering. Die Auswahl von einer Einheit, die für eine der Komponenten im Stimulus verantwortlich ist, zur Gewinnereinheit in einem Zyklus kann so als probabilistische Kodierung des dargebotenen Stimulus interpretiert werden. Die Durchschnittsaktivität einer Einheit über mehrere Zyklen gibt dabei Aufschluss über die Gewinnwahrscheinlichkeit der Einheit in einem Zyklus gegeben den Stimulus.

3 Ein selbstorganisierendes hierarchisches Gedächtnisnetzwerk für kompositionelle Objektrepräsentation

Dieses Kapitel stellt die zentralen Ergebnisse der Arbeit vor. Ausgehend von dem Modell des elementaren kortikalen Moduls aus dem vorangegangenen Kapitel, wird eine neuronale mehrschichtige Netzwerkarchitektur entworfen, die allein durch Erfahrung mit natürlichen Gesichtsbildern verschiedener Personen eine Gedächtnisdomäne für kompositionell-hierarchische Repräsentation der dargebotenen Gesichter aufbauen soll. Das implementierte Gedächtnisnetzwerk stellt erfolgreich seine Funktionalität unter Beweis, indem es die Identität und das Geschlecht von den zuvor unüberwacht gelernten Personen aus verschiedenen Ansichten ihrer Gesichter wiedererkennt. Es wird ebenfalls auf einige bemerkenswerte Eigenschaften der Verarbeitung im entstandenen Gedächtnisnetzwerk eingegangen.

3.1 Unüberwachtes Lernen der Identität und des Geschlechts aus natürlichen Gesichtsbildern

Das Netzwerk besteht aus zwei aufeinander folgenden Schichten, der unteren Vokabularschicht und der oberen Identitätsschicht. Jede Schicht enthält eine Anzahl von Modulen, die sowohl vorwärtsgerichtet als auch rekurrent innerhalb der Schichten und zwischen den Schichten vollständig miteinander verbunden sind. Die Entscheidungszyklen (siehe Abschnitt 2.1), die in den verteilten Modulen ablaufen, sind im gesamten Netzwerk synchronisiert. Die Module der unteren Schicht (insgesamt $M = 6$ Module, $N = 20$ Einheiten pro Modul), im folgenden Vokabularmodule genannt, erhalten den sensorischen Input aus an den Gesichtsbildern vordefinierten Landmarken ($L = 6$ entsprechend der Anzahl der Vokabularmodule), wie im vorigen Kapitel bereits beschrieben. Das Modul auf der oberen Schicht, im folgenden Identitätsmodul genannt, besteht aus $N = 40$ oder $N = 120$ Einheiten, je nach der Anzahl der Personen, für die explizit eine höhere Identitätseinheit reserviert werden soll. Im Initialzustand ist die Netzwerkkonnektivität ohne jede spezifische Struktur, alle rezeptiven Felder der Einheiten sind gleich.

Zusätzlich zu der Standardkonfiguration, wird eine weitere, rein vorwärtsgerichtete Netzwerkarchitektur verwendet, um nach der Lernphase einen Vergleich der beiden Architekturen bezüglich ihrer Erkennungsleistung anstellen zu können. Die rein vorwärtsgerichtete Konfiguration (im folgenden VGK) setzt keine rekurrenten lateralen und top-down Verbindungen zwischen den Modulen, ansonsten ist sie identisch mit der Standardkonfiguration (im folgenden SK). Beide Konfigurationen durchlaufen die gleichen Trainings- und Testphasen. In der Trainingsphase werden die Gesichter der Personen in Originalansicht dargeboten. In der Testphase werden Alternativansichten gezeigt, um die Erkennungsleistung mit der Generalisierungsfehlerrate zu messen.

Den natürlichen Gesichtsbildern ausgesetzt, zeigt sich die SK imstande, jedes einzelne im Laufe der unüberwachten Lernprozedur dargebotene Gesicht als Komposition seiner assoziativ verknüpften lokalen Elemente im Gedächtnis abzulegen. Die Gesichter hinterlassen in der Verbindungsstruktur des Netzwerkes die entsprechend gestalteten Gedächtnisspuren. Die strukturelle Basis für diese Gedächtnisspuren entsteht, indem die vorwärtsgerichteten (bottom-up) und die rekurrenten (laterale und top-down) Verbindungen gleichzeitig gelernt werden. Dabei entstehen auf der unteren Schicht die Vokabulare von wiederverwertbaren lokalen Gesichtselementen, die für je eine Gesichtslandmarke eine Basis für das lokale Erscheinungsbild repräsentieren. Zugleich werden die Elemente aus verschiedenen Vokabularen miteinander über die lateralen Verbindungen assoziativ verknüpft. So werden jene lokalen Elemente als relevante Kompositionen explizit festgehalten, die Teile derselben Gesichtsidentität kodieren. Diese relevanten Kompositionen werden ebenfalls von den Identitätseinheiten der höheren Schicht erfasst und in ihrer bottom-up Verbindungsstruktur festgeschrieben. Dieselbe Information wird von den Identitätseinheiten über die top-down Verbindungen wiederum zurück auf die Vokabularschicht projiziert.

Im ausgereiften Verbindungszustand werden die individuellen Gesichter also auf die kompositionell-hierarchische, generative Weise als dünn besiedelte (sparse) Gedächtnisspuren in der Netzwerkstruktur gelagert. Dieselben Vokabularelemente können dabei von mehreren Gedächtnisspuren, oder anders ausgedrückt, von mehreren gespeicherten Personen geteilt werden. Wenn man nämlich die Selektivität der Einheiten der Vokabularschicht untersucht, stellt man fest, dass jede Einheit von mehreren Personen benutzt wird und zwar so, dass die Nutzlast auf alle Einheiten in etwa gleich verteilt ist. Die Einheiten beteiligen sich also im etwa gleichen Maße an den durch die Erfahrung entstandenen Gedächtnisspuren. Diese ausgeglichene Beteiligung ist der homöostatischen Regulation der Aktivität im Netzwerk zu verdanken. Manche Vokabulareinheiten entwickeln Selektivität nicht nur für die Identität, sondern auch für das Geschlecht der Personen. So lässt sich feststellen, dass einige Einheiten bevorzugt nur auf weibliche oder nur auf männliche Gesichter reagieren.

Das entstandene Gedächtnisnetzwerk kann die Gesichter aus dem Speicher während eines einzelnen gamma-Zyklus abrufen. Wird dem Netzwerk ein Gesichtsbild dargeboten, werden im Laufe des Zyklus eine Reihe von lokalen Entscheidungen in den verteilten Modulen getroffen, die zusammen die kompositionelle Identität des Gesichts wiedergeben. Dabei erzeugt das Netzwerk sehr spärliche (sparse) Aktivität. Nur sehr wenige Einheiten, üblicherweise eine pro Modul, erreichen im Zyklus eine hohe Aktivität, während der Rest sich auf einem sehr niedrigen Aktivitätsniveau einfindet. Die wenigen hochaktiven Gewinnereinheiten repräsentieren die Teile des abgerufenen Gesichtes auf der Vokabularschicht und seine Identität auf der Identitätsschicht. Diese Repräsentation geht nicht nur sparsam mit Einheiten um, sondern ist zudem einfach auszulesen und zu interpretieren.

Wichtig ist, dass die lokalen Entscheidungen, die zur Bildung vom Gewinner-Ensemble führen, nicht isoliert voneinander getroffen werden. Die Module tauschen Signale über die lateralen und top-down Verbindungen aus, die während des Lernens entstanden sind. So wird bei der Interpretation der lokalen Bildinformation aus den jeweiligen Gesichtslandmarken auf den globalen Kontext der bereits gespeicherten kompositionellen Gesichtsidentitäten zurückgegriffen. Wird dem Netzwerk die Ansicht einer bereits bekannten Person dargeboten, so ist der gelungene Abruf des entsprechenden gespeicherten Ensembles sehr wahrscheinlich, da die Einheiten in diesem Ensemble über starke Verbindungen zueinander verfügen. Über diese Verbindungen kann eine starke Kooperation stattfinden, die den Einheiten bei der Entscheidung über die Zusammensetzung des Gewinner-Ensembles gegenüber anderen potentiellen Kombinationen einen eindeutigen Vorteil beschert. Die lokalen Entscheidungen werden im ausgereiften Netzwerkzustand also nicht nur durch die lokale Kompetition zwischen den Einheiten in Modulen getroffen, sondern maßgeblich von den globalen, kontextabhängigen Kooperationseffekten koordiniert.

Beim Abruf lässt sich der Grad der Kooperation zwischen den Gewinnereinheiten am Grad der Übereinstimmung der afferenten Signale ablesen, die lokal an Gewinnereinheiten konvergieren. Diese lokale Signalübereinstimmung ist eine zuverlässige Signatur für einen gelungenen Abruf, wo das bereits bekannte Gesicht korrekt in Form des entsprechenden Gewinner-Ensembles erkannt wird. Wenn der Abruf hingegen misslingt und ein Erkennungsfehler vorliegt, ist die Signalübereinstimmung sehr niedrig, was den Zusammenbruch der Kooperation zwischen den Gewinnereinheiten widerspiegelt. Mit dem Maß der lokalen Signalübereinstimmung erhält man also einen Indikator für die Qualität des stimulusinduzierten Gedächtnisabrufs im Netzwerk.

Im ausgereiften Konnektivitätszustand kann für beide Netzwerkkonfigurationen VGK und SK die Erkennungsfehlerrate auf den Original- und Alternativansichten von den gelernten Personen gemessen werden. Die Fehlerrate wird separat für die Vokabularschicht und für die Identitätsschicht bestimmt. Sowohl bei der Erkennung von Identität als auch bei der Erkennung vom Geschlecht lässt sich auf

Originalansichten, die während der Trainingsphase verwendet wurden, kein signifikanter Unterschied in den Fehlerraten zwischen der VGK und der SK feststellen. Auf den Alternativansichten hingegen besteht ein solcher Unterschied zugunsten der SK. Bemerkenswerterweise ist der Vorteil der SK umso deutlicher, je stärker sich die Alternativansicht von der Originalansicht unterscheidet. Dies deutet darauf hin, dass die SK über eine bessere Fähigkeit verfügt, auf veränderte Ansichten der bereits bekannten Gesichter zu generalisieren. Die bessere Generalisierungsfähigkeit verdankt die SK den rekurrenten lateralen und top-down Verbindungen. Offenbar können diese Kontext vermittelnden Verbindungen die Erkennung auch in der kurzen Zeit eines einzigen gamma-Zyklus vorteilhaft unterstützen. Dieser Vorteil kommt insbesondere dann zum Tragen, wenn die lokale Bildinformation mehrdeutig ist und alleine nicht ausreicht, um die korrekte Interpretation der lokalen Gesichtselemente und der gesamten Gesichtsidentität vorzunehmen.

3.2 Besonderheiten der Verarbeitung im Gedächtnisnetzwerk nach dem Lernen

Nach der Trainingsphase zeigt das entstandene Gedächtnisnetzwerk einige weitere bemerkenswerte Verarbeitungseigenschaften. Das Netzwerk ist beispielsweise imstande, die stimulus-induzierte Aktivität auch nach der Entfernung des Stimulus über mehrere Zyklen aufrechtzuerhalten. Dieser sogenannte Sperrzustand (Locking) kann durch eine simple Modulation der laufenden Rhythmen herbeigeführt werden. Dieser Mechanismus kann als Grundlage für die Arbeitsgedächtnisfunktion verstanden werden, da er erlaubt, die dem Stimulus entsprechende Aktivität über eine Zeit im Netzwerk zu behalten und sie vor dem Überschreiben durch eventuelle Ablenkungsreize zu schützen.

Ferner kann das Netzwerk dank der generativen Art der ausgebildeten Gedächtnisspuren die vollständige kompositionelle Beschreibung der gespeicherten Gesichter abrufen, ohne die Gesichtsbilder sehen zu müssen. Der Abruf kann entweder identitätsgetrieben (top-down) oder teilgetrieben (bottom-up) erfolgen, je nach der Art der verfügbaren Hinweise. Beim top-down Abruf reicht eine leichte Voraktivierung (Bahnung) einer der Identitätseinheiten, um diejenigen Vokabulareinheiten auf der unteren Netzwerkschicht vollständig zu reaktivieren, die den Bestandteilen der gespeicherten Gesichtsidentität entsprechen. Beim bottom-up Abruf ist die Bahnung einer kleinen Teilmenge aus der Gesamtheit der Bestandteile eines gespeicherten Gesichtes ausreichend, um die restlichen Teile und die dazugehörige Identitätseinheit abzurufen.

Experimente mit der Deaktivierung von den lateralen oder top-down Verbindungen zeigen, dass jede dieser Verbindungsstrukturen eigenständig die generative Funktion erfüllen kann. Diese generative Funktion kann auch im Sinne der kompetitiven Aufmerksamkeit gedeutet werden. Wird ein Teil der bestehenden Gedächtnisspur durch Voraktivierung begünstigt, bekommt die gesamte Spur durch die bestehenden Verbindungen zwischen den Einheiten einen Vorteil im Wettbewerb gegen die anderen, potentiell abrufbaren Spuren. Die top-down Voraktivierung kann so als eine objektbasierte Aufmerksamkeit, die bottom-up Voraktivierung als teilbasierte Aufmerksamkeit interpretiert werden.

Eine weitere interessante Besonderheit des Netzwerkes ist seine Fähigkeit, spontane Aktivität in Abwesenheit vom externen Input zu erzeugen. Diese Fähigkeit ist das Produkt der Selbsterregung der Einheiten und der exzitatorischen Verbindungen, die sie während der Lernphase untereinander entwickelt haben. Werden dem Netzwerk keine Gesichtsbilder präsentiert, läuft es autonom weiter, wobei in jedem Zyklus Aktivitätsmuster erzeugt werden. Einige der so erzeugten Aktivitätsmuster entsprechen den im Gedächtnisnetzwerk abgelegten Gesichtern, was der generativen Art der entstandenen Gedächtnisspuren zu verdanken ist. Andere sind wiederum keine exakte Wiedergabe der bereits gespeicherten Inhalte. Solche Muster können als Phantasie-Gesichter aufgefasst werden, die vom Netzwerk als neue Kompositionen aus Gesichtselementen zusammengesetzt werden. Wie oft die Aktivitätsmuster einen tatsächlichen Gedächtnisinhalt wiedergeben, kann durch eine einfache Justierung der Rhythmen reguliert werden. Betrachtet man das Netzwerk als dynamisches System mit Attraktorzuständen, die durch die existierende Konnektivitätsstruktur gegeben sind, kann diese Wiedergabe als Lauf durch die transienten Attraktoren aufgefasst werden. Aus der Perspektive der generativen Modellierung kann diese Wiedergabe wiederum als Stichprobenentnahme aus der gebildeten generativen Gedächtnisstruktur interpretiert werden.

Im Zustand nach dem Lernen zeigt das Netzwerk zudem die Fähigkeit, die Erkennungsleistung nachträglich zu verbessern, ohne die synapsen-spezifische Plastizität zu bemühen. Das ist überraschend, denn für gewöhnlich wird die Modifikation der synaptischen Struktur als Voraussetzung für derartige funktionelle Verbesserung angenommen. Hier stellt sich heraus, dass die Darbietung von Blöcken aus Alternativansichten zum rapiden, starken Abfall der Fehlerrate auf ebendiesen führt und zwar auch bei vollständiger Deaktivierung der synaptischen Plastizität im gesamten Netzwerk. Als verantwortlich für diesen positiven Effekt wird die synapsen-unspezifische homöostatische Regulation der Aktivität ausgemacht. Derselbe Effekt kann dann auch ohne Datendarbietung erzielt werden, indem die Wirkung des homöostatischen Mechanismus - der Ausgleich der intrinsischen Erregung von Einheiten im Netzwerk - manuell durch die Gleichsetzung der Erregungsschwellen nachgebildet wird. Diese Untersuchung bildet die Vorstufe für die Experimente im nächsten Kapitel, wo die Funktion des homöostatischen Mechanismus im vorher beschriebenen, schlafähnlichen "off-line" Regime eine zentrale Rolle spielt.

4 Autonome Gedächtnisverarbeitung in einem schlafähnlichen Regime und die funktionellen Konsequenzen

Der Fokus in diesem Kapitel liegt auf dem Zusammenspiel von zwei bereits beschriebenen Eigenschaften des entstandenen Gedächtnisnetzwerkes. Die eine Eigenschaft ist die Fähigkeit des Netzwerkes, ohne Darbietung von externen Stimuli spontan Aktivitätsmuster zu erzeugen. Während dieser schlafähnlichen "off-line" Phase kann das Netzwerk autonom durch die gespeicherten und "phantasierten" Inhalte laufen. Die andere Eigenschaft ist die festgestellte positive Wirkung der homöostatischen Regulation der Aktivität auf die Erkennungsleistung des Gedächtnisnetzwerkes, die auch in Abwesenheit der synaptischen Plastizität besteht. Hier wird nun untersucht, welche funktionelle Konsequenz es für das Netzwerk hat, wenn es nach der on-line Lernphase Gelegenheit bekommt, in dem schlafähnlichen off-line Regime eine Zeit lang abgekoppelt von externen Stimuli zu verbleiben, während der homöostatische Mechanismus die Aktivität im Netzwerk reguliert.

Für das Testen werden die beiden Konfigurationen, VGK und SK, nach der Lernphase in das off-line Regime versetzt. Im off-line Regime ist die synaptische Plastizität deaktiviert, der homöostatische Mechanismus ist aktiv. Nach der off-line Phase wird die Erkennungsleistung der Netzwerke auf Original- und Alternativansichten erhoben. Wichtig ist, dass die Alternativansichten vor der off-line Phase nicht dargeboten wurden und also für die Netzwerke neue Daten sind. Die Erkennungsleistung nach der Schlafphase wird dann mit der Erkennungsleistung im ursprünglichen Zustand vor der Schlafphase verglichen, um entsprechende Schlüsse zu ziehen.

Während der off-line Phase lässt sich ermitteln, welche Wirkung der homöostatische Mechanismus auf den Zustand des Gedächtnisnetzwerkes hat. Diese Wirkung besteht hauptsächlich im Ausgleich der intrinsischen Erregungsschwellen, der im Netzwerk schichtweise über alle Einheiten der jeweiligen Schicht geschieht. Die Erregungsschwellen, die vor dem Schlafmodus auseinandergedriftet waren, konvergieren so bereits nach einer kurzen Zeit im off-line Regime zu einem gleichen Wert. Dieser Zustand der ausgeglichenen Erregungsschwellen ist genau derselbe, der im vorigen Kapitel für die Verbesserung der Erkennungsleistung als ursächlich befunden wurde.

Tatsächlich stellt es sich heraus, dass die Erkennungsleistung nach dem off-line Regime drastisch verbessert ist, vergleichbar mit dem positiven Effekt wie er im vorigen Kapitel festgestellt wurde. Zwei Punkte sind besonders erwähnenswert im Zusammenhang mit der beobachteten Verbesserung. Zum einen, die Verbesserung der Erkennungsleistung kommt deutlich stärker zum Ausdruck auf den Alternativansichten als auf den Originalen. Das bedeutet, dass die Verarbeitung im Schlafmodus insbesondere die Generalisierungsfähigkeit des Netzwerkes befördert. Zum anderen, obwohl der positive Effekt beide Konfigurationen betrifft, ist er jedoch deutlich stärker ausgeprägt für die SK. Die Verarbeitung im Schlafmodus scheint sich so besonders günstig auf die rekurrente Netzwerkarchitektur auszuwirken.

Diese Ergebnisse deuten klar darauf hin, dass die Verarbeitung im Schlafmodus über den Ausgleich der Erregungsschwellen zur nachfolgenden Verbesserung der Netzwerkfunktion führt. Die Differenzen zwischen den Erregungsschwellen, die während einer andauernden on-line Lernphase zustande kommen, sind offenbar hinderlich für die korrekte Wiedererkennung von gelernten Inhalten, vor allem wenn diese in einer vom Original abgewandelten Form dargeboten werden. Vermutlich kann dieses Phänomen als eine Art von Überanpassung (Overfitting) an die Trainingsdaten verstanden werden. Diese Überanpassung stört nach einer längeren on-line Lernphase die Erkennungsfunktion des Gedächtnisnetzwerkes. Diese Störung kann durch die Verarbeitung im Schlafmodus behoben werden, indem der Effekt der Überanpassung - die ausgeprägten Differenzen der Erregungsschwellen - durch den homöostatischen Mechanismus beseitigt wird. Da die rekurrente Netzwerkarchitektur von der off-line Verarbeitung am meisten zu profitieren scheint, sind die kontexttragenden lateralen und top-down Verbindungsstrukturen wohl besonders anfällig für die negative Wirkung der Überanpassung im on-line Regime. Wird der Effekt der Überanpassung im Schlafmodus beseitigt, können die Kontext vermittelnden Signale ihre Unterstützung der Erkennungsfunktion offenbar besser entfalten.

Die kausale Verbindung zwischen der Verarbeitung im Schlafmodus und der danach beobachteten Verbesserung der Netzwerkfunktion ist bemerkenswert angesichts der experimentellen Befunde über den positiven Effekt der Schlaf- und Ruhezustände auf die Gedächtnisleistung in den davor gelernten kognitiven Aufgaben. Diese Befunde lassen bisher unbeantwortet, was den positiven Effekt auf neuronaler Ebene verursachen mag. Es besteht vor allem die Schwierigkeit zu erklären, wie solche Zustände wie Tiefschlaf (SWS) oder sehr kurzer Schlaf von nur wenigen Minuten (nap) zur nachträglichen Verbesserung der Gedächtnisleistung führen, da die Rahmenbedingungen für die synaptische Plastizität dort ungünstig sind. Die hier erzielten Ergebnisse lassen die Vermutung zu, dass in solchen Phasen die homöostatische Regulation der neuronalen Aktivität für die nachträgliche Verbesserung der Gedächtnisfunktion verantwortlich sein kann.

5 Résume und Ausblick

Das Abschlusskapitel fasst die Ergebnisse zusammen und bietet einen Ausblick auf die Weiterentwicklung der vorgestellten neuronalen Netzwerkarchitektur. Die zu Beginn dieser Arbeit formulierten Anforderungen an ein funktionelles Modell für unüberwachtes Lernen der hierarchisch-kompositionellen Objektrepräsentation im visuellen Kortex wurden erfolgreich umgesetzt. Das vorgestellte neuronale Netzwerk war imstande, aus den natürlichen Gesichtsbildern unüberwacht eine Gedächtnisdomäne für mehrere Personen aufzubauen, wobei die einzelnen Gesichter als Komposition ihrer Bestandteile entlang der Netzwerkschichten abgelegt wurden. Folgende essentielle Funktionalität wurde vom Gedächtnisnetzwerk demonstriert:

- Das Netzwerk verwendete die kompositionell-hierarchische Repräsentation für die gespeicherten Gesichtsobjekte. Es lernte gleichzeitig die lokalen Vokabulare von wiederverwertbaren Gesichtselementen, verknüpfte diese Elemente assoziativ je nach Zugehörigkeit zum selben Gesicht und bildete die höheren Symbole für die Personenidentität aufgrund dieser Verknüpfungen.

- Die gebildete Gedächtnisstruktur hatte eine generative Natur. Es war möglich, allein anhand des höheren Identitätssymbols oder der kleinen Untermenge der entsprechenden Elemente die vollständige Beschreibung eines gespeicherten Gesichtes als Komposition all seiner Teile aus dem Gedächtnis abzurufen.
- Das Netzwerk lernte unüberwacht. Die Gesichtsbilder wurden ohne Zusatzhinweise über die Identität der abgebildeten Personen präsentiert.
- Das Netzwerk bildete gleichzeitig alle Verbindungen - vorwärtsgerichtete bottom-up und rekurrente laterale und top-down - innerhalb und zwischen den Schichten. Dabei wurden nur neuronal plausible Mechanismen verwendet.
- Das unüberwachte Lernen war selbst-stabilisierend und permanent. Es gab keine per Hand definierten Haltbedingungen für das Lernen, die ab einem bestimmten Zeitpunkt das Lernen im Netzwerk eingefroren hätten.
- Im reifen Zustand konnte das Netzwerk sowohl die Identität als auch das Geschlecht der gelernten Personen aus den vorher nicht dargebotenen Alternativansichten ihrer Gesichter wiedererkennen.
- Das Netzwerk war dazu imstande, ein im Gedächtnis abgelegtes Gesicht innerhalb einer kurzen Zeit von einem gamma-Zyklus abzurufen, was vergleichbar ist mit den Vorhersagen der psychophysiologischen Experimente bezüglich der ultraschnellen Objekterkennung.
- Das Netzwerk zeigte die Fähigkeit, spontane Aktivität in Abwesenheit vom externen Input zu erzeugen und die gespeicherten Inhalte im schlafähnlichen off-line Regime wiederzugeben. Diese autonome Verarbeitung im off-line Regime hatte einen direkten funktionellen Vorteil für das Gedächtnisnetzwerk. Die Erkennungsleistung war verbessert nach der off-line Verarbeitung, und zwar besonders stark bei den zuvor nicht gezeigten Alternativansichten. Der positive Effekt wurde überraschenderweise allein durch die synapsen-unspezifische, homöostatische Regulation der neuronalen Aktivität im Netzwerk herbeigeführt.

Dieses Funktionsspektrum wird derzeit von keinem anderen neuronalen Modell der visuellen Objekterkennung geboten. Nichtsdestotrotz kann der vorgestellte Entwurf nur als eine Basisarchitektur verstanden werden, die noch einer intensiven Weiterentwicklung bedarf, um eine universelle selbstorganisierende Gedächtnisdomäne für jede Art von visuellem Inhalt zu bieten. Auf dem Weg dahin muss eine Reihe von schwierigen Problemen gelöst werden, für die gegenwärtig noch keine zufriedenstellende Lösung existiert.

Ein sehr wichtiges Problem, das hier offen gelassen wurde, ist das Problem des Lernens von Transformationsinvarianzen. Die Frage dabei ist wie das Lernen von Merkmalen, die die lokale Erscheinung des Objektes wiedergeben, und das Lernen von Transformationen, die auf die Erscheinung einwirken können (wie beispielsweise Translation, Skalierung, Drehung, und so weiter), miteinander zu kombinieren sind. Die Erkennung eines Objektes aus seinem eventuell transformierten Abbild bedeutet dann nicht nur die invariante Erkennung seiner Identität, sondern ebenfalls die Erkennung der Transformationen, die seine Erscheinung im Bild verändert haben. Eine Erweiterung für die translationsinvariante Erkennung würde es dem aktuellen System beispielsweise ermöglichen, die Gesichtslandmarken automatisch zu lokalisieren und die Position von den Gesichtselementen im dargebotenen Bild explizit in das Ergebnis des Erkennungsvorganges zu integrieren.

Für die Behandlung der Translationsinvarianz existiert ein bereits erprobter Ansatz, der nahtlos in das bestehende System integriert werden kann. Dieser Ansatz gründet auf der sogenannten Dynamic Link Architecture (DLA) und verwendet eine zusätzliche Netzwerkschicht, in der die zulässigen topologischen Anordnungen von lokalen Teilen zu einem ganzen Objekt gespeichert sind. Die Module dieser topologischen Schicht können Korrespondenzen zwischen den Orten auf dem Eingabebild und den entsprechenden Teilen in der Gedächtnisdomäne herstellen und zugleich die bestimmten Positionen mit ihrer Aktivität explizit signalisieren. Sind von Anfang an weder die topologischen noch die erscheinungsrelevanten Merkmale im System vorverdrahtet, besteht das zu lösende Lernproblem darin, die beiden Gedächtnisdomänen, die topologische und die für die Erscheinung zuständige, gleichzeitig zu bilden. Für die Behandlung von weiteren Invarianzen, wie der Größen- oder Drehungsinvarianz, sind in diesem Konzept weitere entsprechende Subsysteme notwendig, für die die ersten Studien bereits erfolgversprechend verlaufen sind.

Ein weiteres klassisches Problem, das eine besondere Aufmerksamkeit verdient, ist die Frage nach der langfristigen Verwaltung der gebildeten Gedächtnisstruktur. Diese Verwaltung muss sicherstellen, dass sowohl neue Inhalte ins Gedächtnis schnell integriert werden können als auch die alten relevanten Inhalte nicht überschrieben werden und verloren gehen. Eine derartige Verwaltung kann hypothetisch durch die Einführung eines vollständigen Schlaf-Wach-Zyklus implementiert werden. Während die Wachphase für den Erwerb der neuen Inhalte in Form von schwachen Gedächtnisspuren benutzt werden kann, könnte in der Schlafphase die Verstärkung und Stabilisierung dieser schwachen Spuren mittels der selbst-generierten Wiedergabe der Inhalte stattfinden. Die nicht relevanten, störenden Spuren könnten dabei eliminiert werden. Dieselbe Wiedergabe im Schlafzustand könnte auch dazu verwendet werden, die alten relevanten Inhalte aufzufrischen, um sie vor der Löschung zu bewahren.

Um Grundlage für aktives Sehen und Lernen zu schaffen, muss eine andere wichtige Form vom Lernen einbezogen werden, nämlich das Lernen durch Verstärkung (reinforcement learning). Hierzu kann der verwendete Mechanismus der bidirektionalen synaptischen Plastizität erweitert werden, um zusätzlich die Verstärkungssignale zu berücksichtigen, die im Gehirn in Form der dopaminergen (DA) Neuromodulation kommuniziert werden. Die DA-Signale könnten die Zonen der Potentierung und Depression in der synaptischen Regel je nach ihrer Intensität so verschieben, dass geringe Mengen von DA die Depression und große Mengen von DA hingegen die Potentierung wahrscheinlicher machen würden. Eine solche verstärkungsabhängige Regel würde Zuwendung hin zu den Inhalten fördern, die verhaltensrelevant sind, und dadurch aktive Auswahl der Inhalte ermöglichen, die vorzugsweise im Gedächtnis abgelegt werden sollten.

Das entwickelte Netzwerk hat so nicht nur eine im Bereich der neuronalen Modellierung bisher nicht gezeigte Funktionalität demonstriert, sondern stellt auch offene Schnittstellen für eine konsequente Weiterentwicklung zur Verfügung. Der vorgestellte Entwurf bietet daher einen vielversprechenden Ausgangspunkt für weitere Studien, die sowohl die neuronalen Lernmechanismen des Gehirns ins Visier nehmen als auch letztendlich deren konsequente Umsetzung in technischen adaptiven Systemen anstreben.

Danksagung

Während der Arbeit an der vorliegenden Dissertation hatte ich das große Glück, die Unterstützung von vielen Menschen und Institutionen zu bekommen, ohne die diese Arbeit so nicht zustande gekommen wäre. Das Mindeste, was ich tun kann, ist mich an dieser Stelle bei all denen zu bedanken und ganz klar zum Ausdruck bringen, dass diese Arbeit als Produkt gemeinsamer Anstrengungen zu verstehen ist.

Die größte tragende Kraft, die mir den Antrieb gewährte und die Inspiration einflößte, entsprang meiner Familie. Ich kann gar nicht genug betonen, wie viel von dieser Arbeit deshalb meiner Frau Catherine und meiner Tochter Cailie Mia zu verdanken ist. Auch meine Eltern ließen mich stets spüren, dass sie bei allem, was ich tue, bedingungslos hinter mir stehen. Diese Hilfe war für mich vom unschätzbaren Wert.

Ihre vorhandene Gestalt konnte die Arbeit nur dank einer besonderen Betreuung annehmen, die mir am FIAS zuteil wurde. Hervorheben möchte ich die Art der Zusammenarbeit mit meinem langjährigen Mentor, Prof. Christoph von der Malsburg. Er schaffte für mich eine kreative Arbeitsatmosphäre, in der ich immer auf seinen Rat hoffen konnte, und gewährte mir dabei den größtmöglichen Freiraum, um eigene Ideen, wie extravagant diese auch sein mochten, zu verfolgen. Mein Dank an dieser Stelle gebührt auch Prof. Jochen Triesch, Prof. Rudolf Mester, Prof. Wolf Singer, Prof. Gaby Schneider, Dr. Danko Nikolic, Dr. Gordon Pipa, Dr. Thomas Burwick und Dr. Junmei Zhu, die alle durch fruchtbare Diskussionen zu der Qualität der Arbeit maßgeblich beigetragen haben.

Sehr wichtig für mich waren die Freundschaften, die während der Zeit entstanden sind. Unser Office war die unbestrittene Oase und der Knotenpunkt des sozialen Lebens, errichtet von gemeinsamen Bemühungen von Hasnaa Fatehi (beste Lebensweisheiten), Cristina Savin (tolle Kuchen), Maneesh Mathew (unerschütterliche Ruhe) und Adilah Hussien (halb legal dabei, Kaffeeflecken auf dem Teppich, ganz großes Herz). Nicht minder wertvoll waren die vielen, nicht nur im wissenschaftlichen Sinne aufschlussreichen Momente, die ich mit Urs Bergmann, Jan Scholz, Christian Keck, Jörg Bornschein, Daniela Pamplona, Yasuomi D. Sato, Phillipp Wolfrum, Martha Havenith und Andreaa Lazar erlebt und geteilt habe.

Nicht unerwähnt bleiben darf hier das kongeniale Wesen von Stefan Bölingen. Seine seltene Gabe, die schlechte Stimmung und die verdeckten Fehler zu vertreiben, kam nicht nur dieser Arbeit zugute. Einen ganz besonderen Dank möchte ich auch Francoise Lubbers aussprechen, die mich in allen möglichen Lebenslagen mehrfach gerettet hat.

Schließlich möchte ich für die finanzielle Unterstützung im Rahmen des EU Projekts DAISY (Förderungsnummer: FP6-2005-015803) und des BMBF Bernstein Netzwerks (Förderungsnummer: 01GQ0840) danken.

Curriculum Vitae

Personalien

Name	Evgueni (Jenia) Jitsev
Adresse	Hermann-Gmeiner-Str. 20, 53229 Bonn
E-mail	jitsev@gmail.com
Geburtsdatum	23.07.1979
Geburtsort	Smolensk, Russland
Familienstatus	Lebensgemeinschaft
Kinder	Cailie Mia Lubbers, geb. 18.05.2002

Studium

04.2008	Aufnahme in Otto Stern School for Integrated Doctoral Education (OSS, nachfolgend GRADE)
SEIT 12.2007	Doktorand an der Johann Wilhelm Goethe Universität Frankfurt
10.2000 - 04.2006	Studium der Informatik mit Nebenfach Psychologie, Rheinische Friedrich-Wilhelm Universität Bonn

Beruf

SEIT 05.2010	Wissenschaftlicher Mitarbeiter am Max-Planck-Institut für neurologische Forschung, Köln
07.2009 - 03.2010	Wissenschaftlicher Mitarbeiter im Bernstein Focus: Neurotechnology Projekt, FIAS
06.2007 - 06.2009	Wissenschaftlicher Mitarbeiter an der Johann Wilhelm Goethe-Universität Frankfurt
07.2006 - 05.2007	Wissenschaftlicher Mitarbeiter an der Ruhr-Universität Bochum, Institut für Neuroinformatik
07.2006 - 06.2009	Wissenschaftlicher Mitarbeiter im EU-Projekt DAISY: "‘Neocortical Daisy Architectures and Graphical Models for Context-Dependent Processing"’

Publikationen (peer-reviewed)

Jitsev, J. and von der Malsburg, C. Off-line memory reprocessing following on-line unsupervised learning strongly improves recognition performance in a hierarchical visual memory. In *International Joint Conference on Neural Networks (IJCNN), Special session on Organic Computing, IEEE World Congress on Computational Intelligence (WCCI), Barcelona, Spain*, 1–8, 2010

Jitsev, J. and von der Malsburg, C. Experience-driven formation of parts-based representations in a model of layered visual memory. *Frontiers in Computational Neuroscience, Special Issue on "Complex Systems Science and Brain Dynamics", 2009, 3:15*

Sato, Y.D., Jitsev, J. and von der Malsburg, C. A visual object recognition system invariant to scale and rotation. *Neural Network World (ICANN Special Issue), 19(5), 529–544, 2009.*

Sato, Y.D., Jitsev, J. and von der Malsburg, C. A visual object recognition system invariant to scale and rotation. *In Proc. ICANN 2008, LNCS 5164(1), 991-1000, 2008*

Vorträge auf Einladung

07.2010	International Joint Conference on Neuronal Networks (IJCNN), World Congress on Computational Intelligence (WCCI), Barcelona, Spain *Special Section on Organic Computing, Chair : Rolf Würtz*
02.2010	Max-Planck-Institut für neurologische Forschung, Gleueler Str. 50, 50931 Köln *Abteilung "Kortikale Netzwerke". Institutsleiter : Yves von Cramon*
12.2009	Die Eidgenössische Technische Hochschule (ETH) Zürich, Institut für Neuroinformatik, Winterthurerstrasse 190, CH-8057 Zürich, Schweiz *Institutsleiter : Roudney Douglas*
12.2009	Abteilung für Kognitive Neurophysiologie, Klinik für Epileptologie, Sigmund-Freud-Straße 25, 53105 Bonn *AG Fell/Axmacher, Cortical Oscillations Lab. Klinikleiter : Christian E. Elger*