Visual Recognition with a Large Scale Network of Dynamical Systems

by

Socrates Dimitriadis

B.Sc. in Computer Science, University of Ioannina, Greece, 1999

M.Sc. in Computer Science, University of Crete, Greece, 2002

M.Sc. in Cognitive Science, Brown University, Providence RI, 2006

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Cognitive, Linguistic and Psychological Sciences at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2011

This thesis by Socrates Dimitriadis is accepted in its present form by the Department of Cognitive, Linguistic and Psychological Sciences as satisfying the thesis requirements for the degree of Doctor of Philosophy

Date_____ _____

James Anderson, PhD, Advisor

Date_____ _____

Fulvio Domini, PhD, Reader

Date_____ _____

Thomas Serre, PhD, Reader

Approved by the Graduate Council

Date_____ _____

Peter M. Weber, PhD, Dean of the Graduate School

AUTHORIZATION TO LEND AND REPRODUCE THE THESIS

As the sole author of this thesis, I authorize Brown University to lend it to other institutions or individuals for the purpose of scholarly research.

Date May 1, 2011 

Socrates Dimitriadis, Author

I further authorize Brown University to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Date May 1, 2011 

Socrates Dimitriadis, Author

Curriculum Vitae

Personal Data

Last name	Dimitriadis
First name	Socrates
Father name	Pantelis
Mother name	Ekaterini
Date of birth	August 11, 1976
Nationality	Greek
Website	www.socrates.name
e-mail	email@socrates.name

Education

2003-2006	M.Sc. in Cognitive Science – Brown University, USA
2000-2002	M.Sc. in Computer Science – University of Crete, Greece GPA 9.28/10.0 Major in Computer Vision and Robotics Minor in Software Engineering and Information Systems
1995-1999	B.Sc. in Computer Science – University of Ioannina, Greece 1st in class '99, GPA 8.13/10.0

Foreign Languages

English	1990, First Certificate in English 2002, T.O.E.F.L. (CBT 260)
French	1992, Certificat de langue francaise
German	1993, Zertifikat, Deutsch als fremdsprache
Italian	1995, Diploma di lingua Italiana

Athletics

Long Distance and Marathon Runner (participation in more than 100 races)
Bronze Medal in 1500m, National Student Championship 1998
Entry to six International Marathons (PR: 3.05.57)

Awards and Distinctions

2007-2008	Brown University Corinna Borden Keen Research Fellowship
11/2006	OTEnet Prize in competition *Innovation 2006*
2004-2009	Brown University Teaching and Research Assistantships
2003-2004	Brown University Fellowship
2001-2003	ICS-FORTH Research Assistantship
2001-2003	University of Crete Teaching Assistantship
2000-2001	University of Crete Fellowship
1999	Hellenic Telecommunications Organization Scholarship (student with the highest GPA in the 8th semester of studies)
1998	Hellenic Telecommunications Organization Scholarship (student with the highest GPA in the 7th semester of studies)
1998	Greek State Scholarships Foundation (IKY) Prize
1998	Scholar of the Greek State Scholarships Foundation
1995-1999	Voyatzis Foundation Scholarship

Research Projects

2008-2009	Brown University Laboratory for Engineering Man/Machine Systems Visual Change Detection in Satellite Images
2004-2010	Brown University Ersatz Brain Project Computational Cognitive Neuroscience Modeling
2001-2003	University of Crete & Institute of Computer Science – FORTH Content Based Image Retrieval
2001-2002	Institute of Computer Science – FORTH Computational Vision and Robotics Laboratory Development of a Robotic Soccer Platform for Robocup
2000	Institute of Cultural and Educational Technology, Xanthi Test and Evaluation of Teleconference-Teleducation Platforms
1998-1999	University of Ioannina EU Program EPEAEK Distance Learning Course for Neural Networks

Refereed Publications and Presentations

Socrates Dimitriadis, "**Visual processing as a large scale integration of associative neural networks**", *4th Computational Cognitive Neuroscience Conference, Boston MA, USA, November 2009*

Socrates Dimitriadis, "**Cortically-inspired parallel processing**", *13th SIAM conference on Parrallel Processing for Scientific Computing, Atlanta GA, USA, March 2008*

Socrates Dimitriadis and Fulvio Domini, "**Bayesian inference of visual depth based solely on primitive visual variables**", *30th European Conference on Visual Perception, Arezzo, Italy, August 2007 & Perception Vol.36*

James A. Anderson, Paul Allopenna, Gerald S. Guralnik, David Scheinberg, John A. Santini, Socrates Dimitriadis, Benjamin B. Machta, and Brian T. Merritt, "**Programming a Parallel Computer: The Ersatz Brain Project**", *In Wlodzislaw Duch and Jacek Mandziuk (Eds.), Challenges for Computational Intelligence, Vol. 63, Springer: Berlin, 2007*

Socrates Dimitriadis, Kostas Marias and Stelios Orphanoudakis, "**A Multiagent Platform for Content Based Image Retrieval**", *Journal of Multimedia Tools and Applications,* Special Edition: *Distributed Adaptation, Representation and Processing of Multimedia Information, Vol. 33, No. 1, 57-72, April 2007*

Socrates Dimitriadis and James Anderson, "**A network of networks model of pop-out phenomena in visual attention**", *2nd Computational Cognitive Neuroscience Conference, Houston TX, USA, November 2006*

John Moustakas, Kostas Marias, Socrates Dimitriadis and Stelios Orphanoudakis, "**A Two-level CBIR Platform with Application to Brain MRI Retrieval**", *IEEE International Conference on Multimedia & Expo (ICME), Amsterdam, The Netherlands, July 2005*

Socrates Dimitriadis, Kostas Marias and Stelios Orphanoudakis, "**A Versatile Image Retrieval Platform Based on a Multiagent Architecture**", *6th International Conference on Visual Information Systems, Miami FL, USA, September 2003*

Workshops

Parallel Programming, Brown University, Providence, RI, October 18-19, 2004, *Organized by Brown University's Technology Center for Advanced Scientific Computing and Visualization (TCASCV)*

Nonlinear Methods in Psychology, George Mason University, Fairfax, VA, October 24-25, 2003, *Organized by National Science Foundation (NSF)*

Other Publications

Socrates Dimitriadis et al., "**2006 Review of Greek Information and Communication Technologies**", Hellenic Informatics Union, February 2007

Socrates Dimitriadis, "**A Network of Networks Approach to Computational Modeling of Pop-out Phenomena in Visual Attention**", MSc. Thesis, Department of Cognitive and Linguistic Sciences, Brown University, May 2006

John Moustakas, Socrates Dimitriadis and Kostas Marias, "**A Cognitive Architecture for Semantically Based Medical Image Retrieval**", *ERCIM News No. 62, July 2005*

Socrates Dimitriadis, Kostas Marias and Stelios Orphanoudakis, "**Retrieval of Images based on Visual Content: A Biologically Inspired Multi-Agent Architecture**", *ERCIM News No. 53, July 2003*

Socrates Dimitriadis, "**An Image Retrieval Platform Based on a Biologically Inspired Architecture**", MSc. Thesis, Department of Computer Science, University of Crete, November 2002

Socrates Dimitriadis, "**Human-Computer Interaction with Gestures**", BSc. Thesis, Department of Computer Science, University of Ioannina, June 1999

Teaching Experience

2004-2009 Teaching Assistant
Brown University, Department of Cognitive and Linguistic Sciences
Courses: Quantitative Methods in Psychology (fall 2004), Neural Modeling Laboratory (spring 2005/2006/2007/2009), Visualizing Vision (fall 2005), Cognition (fall 2006)

2002-2003 Teacher of Computer Science
6th School of Technical Education of Heraklion, Greece
Courses: Multimedia, Operating Systems-UNIX, Computer Applications

2000-2002 Teaching Assistant
University of Crete, Department of Computer Science
Courses: Computer Networks, Computational Vision, Intelligent Systems

2001 Teacher of Computer Science
Vocational Training Institute of Kilkis, Greece
Courses: Programming in C, Sound Processing and Synthesis

Other Working Experience

since 2006	Founder and owner of Greek-Movies.com, a search engine and directory for Greek videos
1999-2001	Hellenic Army, Reserve Officer of Communications
1994	Water Supply and Sewerage Company of Kilkis
1993-1994	Radio Producer in Akroama Radio Station and Euro Channel

Memberships

2007-2009	Graduate student representative at the Computing Advisory Board of Brown University
2004-2005	Graduate student representative at the Board of the Department of Cognitive and Linguistic Sciences of Brown University
since 2002	Member of the Hellenic Informatics Union
1998-1999	Undergraduate student representative at the Senate of the University of Ioannina
1997-1999	Undergraduate student representative at the Board of the Department of Computer Science of University of Ioannina

Inspired by the words of an unknown poet

When things go wrong, as they sometimes will
When the road you're trudging seems all uphill
When the funds are low and the debts are high
And you want to smile, but you have to sigh
When care is pressing you down a bit
Rest, if you must, but don't you quit

Life is queer with its twists and turns
As every one of us sometimes learns
And many a failure turns about
When he might have won had he stuck it out
Don't give up though the pace seems slow
You may succeed with another blow

Often the goal is nearer than
It seems to a faint and faltering man
Often the struggler has given up
When he might have captured the victor's cup
And he learned too late when the night slipped down
How close he was to the golden crown

Success is failure turned inside out
The silver tint of the clouds of doubt
And you never can tell how close you are
It may be near when it seems so far
So stick to the fight when you're hardest hit
It's when things seem worst that you must not quit

Acknowledgments

First of all, I would like to thank my advisor, Jim Anderson, for believing in me and for giving me the opportunity to work my own way in this research. He has always been very supporting and infused me with confidence.

A great thanks goes to the rest of the members of the Ersatz Brain Group and in particular to Paul Allopenna, John Santini, and Gerry Guralnik. The endless yet fruitful discussions we had in our weekly meeting all these years helped me shape many aspects of this thesis.

I'm really grateful to the readers of my dissertation, Fulvio Domini and Thomas Serre, for the excellent comments and the valuable feedback they provided. I couldn't hope for a more expert reviewer than Thomas and I wish he had come earlier to our department. Fulvio Domini was a lot more than a reader, a true friend, and I will miss the running we did together.

My research wouldn't be possible without the equipment and the technical support of the Center for Visualization and Computation and I'm grateful they let me use their systems free of charge for many years. I would also like to thank Joe Mundy for the research assistantship he offered to me in his lab for one semester.

Last but not least, I would like to thank all my friends that helped me keep my sanity and my happy mood all these times – and there were plenty of them – that research didn't work the way I was expecting it to. My life in Providence wouldn't be the same without Dimitris' jocularity, Yannis' stresslessness, Babis' clumsiness, and Foteini's love. I will also never forget the fun we had with Christos, Panayotis, Vasilis, Olga, Nikos, Dimitra, Aris, Jeff, Arnold, Aggeliki, Adrian, Amit, Justin, Jae-Yung, Raj, Elena, Naomi, Giorgos, Maria, Danny, Vita, Katerina, Sophocles, Menia, Carmen, Qun, Paris, Stavroula, Kostas, Andrea, Sakis and Basilis, to name just a few!

Contents

List of Figures

1. Introduction

Visual perception is one of the most fundamental abilities of our cognitive system. For the visual world to be perceived the way it is tasks like visual attention, pattern recognition, and visual cognition have to be seamlessly integrated and coordinated all the time. Visual processing, although seemingly effortless, is the result of massive and complex distributed computations that involve millions of neurons and a disproportionate amount of our brain circuitry. Joint research in many fields over the past few decades has generated a lot of discoveries for this marvelous cognitive modality that is probably the most thoroughly studied today. Numerous secrets, from visual neurobiology to high level vision, have been revealed experimentally so far. However, despite the huge number of details we already know about vision, we still lack a larger theoretical picture that will put all the disparate pieces together and connect the dots (Olshausen 2005). Most findings are dissociated from each other and most theories and models have a very restricted scope. The visual system is still a complex puzzle.

The primary goal of this thesis is to begin to articulate a computational theory that will provide a neuronal-level explanation of the macro-level function of the visual system. In order to bridge the gap between neurons and behavior we have to shed the light not only on the pieces themselves but on how the puzzle assembles these pieces as well. Therefore we choose to experiment with a large scale integration of basic visual constituents instead of modeling certain details of the components of the visual system. Our belief is that understanding the fusion

of the information is sometimes equally, or even more, important than the information itself. And in order to understand the multifarious encapsulation of information that takes place in our brain we have to decipher the way that neuronal assemblies and network structures are formed. Studying the details without having a theory of how things may be put together is hopeless. Besides, from what we know about brain's robustness and adjust-ability, there is a strong indication that the integration process should be equally crucial and to some extent independent of the details of the individual components themselves. If this is indeed the case then our lack of knowledge for certain details at the neuronal level of processing, and the subsequent simplifications and assumptions we would have to inevitably make during this integration, should not be critical for the primary goal of this study.

The second goal of this thesis is to propose and defend a computational framework that is not only biologically plausible and computationally feasible, but it also provides explanations at multiple levels. The proposed approach is able to provide the mechanistic details of a large scale neuronal integration while also accounting for a variety of behavioral phenomena at the level of visual cognition. It reconciles the notion of statistics on large image sets with the evidence from large numbers of distributed action potentials. We call it a framework rather than a model because different instantiations of the same architecture can exhibit diverse behaviors. Formally, it is a network of nonlinear dynamical systems with a well described mathematical formulation that is simple and easy to understand and analyze. Visually, the framework consists of a large network of orderly interconnected neural networks that are arranged in a grid-like topography. This

topological arrangement is inspired by the cortical columnar organization (Hubel and Wiesel 1974, Mountcastle 2003). Each constituent network has a specific function that abstracts the functionality of a cortical column, or minicolumn, depending on the level of abstraction we choose. The human cortex has powerful associative properties (Hebb 1949), hence we use associative neural networks as building blocks in our system. Each of these attractor networks learns to auto-associate visual stimuli in a local receptive field and forms stable point attractors that correspond to learned patterns of activity. In addition to internal module activity, networks are also interconnected and form hetero-associations of their internal states. This hybrid association drives the formation of network-wide attractors that correspond to pattern assemblies. Therefore, attractors that tend to co-occur cause a reciprocal excitation – conversely an inhibition – and facilitate the recognition of incomplete or noisy visual patterns. In the presented version of this framework there is only one layer of the proposed architecture, but in the general case there should be multiple layers organized hierarchically similarly to the visual association areas (Felleman and Van Essen 1991).

The choice of the domain of visual perception for the experimentation with the proposed framework is not imperative. The overall approach could be equally suitable for other modalities like auditory perception or for integrating multimodal sensory information as well. Based on the principles of associativity and interconnectivity that are ubiquitous in the brain, an appropriate encoding of the stimuli along with a customization of the distribution of the dynamical systems network, should in principle be able to provide a diverse behavior to the proposed framework.

In the following chapters we first going to give a concise review of the trends that exist in the literature regarding the modeling of visual pattern recognition. We'll put an emphasis on a comparative analysis rather than on the individual details. This will give us an overview of the current situation as well as an understanding of how we got here and where we are heading to. Next, we're going to give a comprehensive description of the proposed framework, provide the reasoning behind its architecture, and discuss the implications of such an approach. Following that, we will put our framework side by side to the most popular models and examine qualitatively the points of agreement, but mostly, the ones of disagreement. There are important differences, conceptual and pragmatic, that have significant implications not only in how we perceive a model of vision but also in the scope and the limitations of the process of modeling itself. Last, we will present our experimental methodology and the results we got. We will discuss the findings and see how close they are to the ones we were expecting, as well as the predictions they make. We will conclude with some general thoughts, ideas for potential improvements, and future directions.

2. Review of visual recognition models

The literature in the domain of visual pattern and object recognition is very large and rich. Since we argue for a meso-level approach that is able to account for high level behavioral phenomena while using low level mechanisms, it is necessary to examine a wide range of models that have a focus spanning from the neurophysiology of vision to visual cognition. Although this review will be far from comprehensive the main goal is to give a brief overview of the most influential models in the literature and examine the details of their operation, the constraints they work under, and the assumptions they make. This will provide an insight to their explanatory and predictive power that will help us make a comparison with our approach later.

A review of the literature reveals a classification of the visual recognition models mainly according to the following basic dimensions: the representation of the object, the encoding of the stimulus, the architecture of the model, and the flow of processing. Based on these principal criteria we see distinctions between an object-centered and a view-based representation, a generic versus a class-specific encoding, a monolithic versus a modular architecture, and a feed-forward versus a feedback processing (Logothetis and Sheinberg 1996, Tarr and Bulthoff 1998, Wallis and Bulthoff 1999, Riesenhuber and Poggio 2000b). These distinctions may not always have an obvious implication on the performance of the various models but they have a profound significance on their plausibility and feasibility as models of human behavior. Next, we will review several models based on their stance towards a variety of these principles.

2.1. Representation

One of the fundamental questions in visual recognition concerns the representation of the object or pattern that is recognized. On this aspect there are two main schools of thought. The first one argues for an object-centered representation that holds information about the full object in 3D space while the second one for a view-based representation that is limited by the number of views it has collected. In the first case we talk about a structural description that is formed by extracting visual information and building a view invariant representation. Once stored, the representation is uniquely defined, independent of the viewing conditions (viewpoint, lighting, etc.), and can be tested against previously stored or new objects. This idea was first introduced by Marr and Nishihara 1978 who argued that a shape representation for recognition should use an object-centered coordinate system and should include volumetric primitives of varied sizes. In this direction they suggested that the process of deriving a shape description such that must involve a means for identifying the natural axes of a shape and a mechanism for transforming viewer-centered axis specifications to specifications in an object-centered coordinate system. Moreover, regarding the process of recognition itself, they assumed a number of indexes that allow a new description to be associated to the stored ones. Clearly, the whole approach was influenced by the computer metaphor of the time.

Few years later Biederman proposed an object-centered theory of visual recognition named *Recognition by Components* (Biederman 1987, Hummel and Biederman 1992). The components in this proposal are geometrical volumes, called *geons,* that correspond to regions of deep concavity and constitute the

building blocks for representing objects in 3D space. Their basic assumption is that these generalized cones can be derived from only five properties of a 2D image (curvature, collinearity, symmetry, parallelism and co-termination) which are invariant over viewing position and image quality. Therefore, since geons are reusable and can be freely assembled to form a variety of visual configurations, a modest set of them can theoretically be adequate for the representation of a far larger set of objects. Depending on the richness of this geon bank and the limitations of the structural description the recognition by components can thus be very powerful. This theory, however, lies in a very high level of description without providing a formal account for the recognition of the geons themselves. Moreover, the extraction of a structural description is dependent on the number of disparate views the observer would be able to collect from an object. Therefore, the so claimed view-invariant recognition of an object-centered approach cannot be completely different from a view-based approach.

On the other side, the view-based recognition argues for a representation that consists of a collection of image views. A *view* doesn't necessarily correspond to a two-dimensional image but it can be any kind of viewpoint-dependent extraction of visual features, including depth and stereo information. Apparently, this type of representation is subject to noise caused by illumination and the viewing conditions in general. Moreover, the efficacy of this representation greatly depends on the number of disparate views the observer had the chance to collect in the course of time. If sampling of the scene is restricted to certain viewpoints only, extrapolation to new ones will probably be difficult. Although, this approach seems less powerful than the object-centered one and, in a sense,

counterintuitive to our own visual recognition abilities, experimental evidence favors the hypothesis that visual recognition in both humans and animals is viewpoint dependent (Logothetis et al. 1994, Tarr 1995, Tarr 1999, Tarr et al. 1998; but see Bar 2001 for a different interpretation). Psychophysical experiments demonstrate that humans are able to recognize unseen object views by interpolating views they previously learned (Bulthoff and Edelman 1992). This is also supported by model predictions (Poggio and Edelman 1990) as well as neurophysiological data showing that the majority of inferior temporal neurons are tuned to single views of the training objects with only a very small number of neurons being view invariant (Logothetis et al. 1995).

Most models in the literature adhere to a view-based approach and regardless of the many other differences they have they all get trained on a set of image views coming either from within the same class of objects, if the purpose is the identification, or from different classes, if the purpose is the categorization. Some of the most known approaches in this direction are the models by Poggio and Edelman 1990, Perrett and Oram 1993, Neocognitron (Fukushima 1988), SEEMORE (Mel 1997), VisNet (Wallis and Rolls 1997), and HMAX (Riesenhuber and Poggio 1999).

2.2. Encoding

The encoding of the input stimulus is of fundamental importance and probably one of the most critical issues, not only for visual recognition but for visual processing in general. There exists a great variety of proposed methods for encoding the input such as, intensity values of the raw image, color distribution,

shape information (e.g. different moments), texture, spatial frequencies, templates, etc. More important, though, the encoding of these features can either have a local scope (e.g. the color value of a particular pixel or receptive field) or a global one (e.g. the color distribution of the whole image given by the histogram). In either case, the goal of the encoding is to reduce the dimensionality of the highly variable input signal in order to facilitate the categorization or identification during visual recognition (Field 1994). By using a rich set of generic features with a local scope it is easier to flexibly encode a larger input space for a variety of objects from different classes. This, of course, comes at the expense of a higher complexity during the process of integrating these features across the image. On the other hand, the use of wide scope features, that result either from the encoding of the whole image or from a segmentation of it into regions of specific interest, has several limitations but with fewer problems during the information fusion. This latter approach reduces the high complexity arising from the full joint distribution over the features but is restricted to certain configurations or classes of objects. So, there is a clear trade-off and most models of visual recognition choose one of these two approaches. They either start with a local scope encoding of the two-dimensional input signal and ascend the path of integration by assembling these features, or they assume some kind of preprocessing that provides a wide scope class-specific image encoding that eliminates most of the input variability and focuses on higher-level relations within the image.

An example of the latter is the eigenfaces recognition (Turk and Pentland 1991) that takes a holistic approach and describes each face as a combination of eigenvectors produced by a high-dimensional space of face images. The role of

the encoding here is simply to extract the principal components (eigenfaces) of a covariance matrix formed by concatenating the training images. The approach is global in nature and doesn't take into account any of the features that make a face what it is. The most well known model of visual recognition that works with high-level information is probably the fragment-based recognition (Ullman and Sali 2000, Ullman et al. 2002). The model doesn't use absolutely global image information, though, but it partitions the images into fragments that are associated with a certain class of objects. The fragments have a varying size that depends on the class of objects they target and their selection is made with the goal of optimizing their mutual information with a particular class. Recently, the model has been extended to incorporate a hierarchy of fragments that argues for a possible framework for categorization, recognition and segmentation in human vision (Ullman 2006).

All these approaches emerged primarily from a computer vision background and the need for automated systems that can borrow some of the human abilities. Apparently, by working with more abstract visual information one can avoid the issues that arise from processing the low-level highly-variable input signal, and possibly achieve a better recognition performance. However, the assumptions that these models make have no biological ground and so far seem pretty implausible. Although, it is true that in the higher association areas of the visual cortex (e.g. V4) the concept of *features* becomes very complex; and it is also true that there exist selective cells in the inferotemporal cortex that respond to complete object views (Logothetis et al. 1995, Tanaka 2003), there is no evidence so far that these complex features or object views correspond to image

fragments of arbitrary sizes. Even if in the future such an encoding proves to be correct, and therefore justifies the approach that these models take, the fundamental question (in computational terms) of how do we arrive at this level of encoding remains unanswered by these approaches.

Contrary to these approaches, most of the models in the literature adopt some variation of a local-feature encoding. Some of them use receptive fields with predetermined position and size (Fukushima 1988, Wallis and Rolls 1997, Riesenhuber and Poggio 1999). Some others use receptive fields with random positions and varying size (Edelman 1993, Edelman 1996), while other models use local patches without committing to the concept of receptive fields (Nelson and Selinger 1998, Amit and Geman 1999). As for the feature extraction itself, most models apply some form of filtering on the raw signal that detects edges or orientation bars, while others use Gabor wavelets which according to neurophysiological evidence model the filter response profiles of the simple cells in V1 (Daugman 1980, Marcelja 1980, Lee 1996). But regardless of the numerous details, the basic idea in most models is the same. The processing of the stimulus starts independently on an array of local fields and the information is incrementally fusing during the flow of the visual processing. Similar to what happens in the mammalian visual system, from the retina, to LGN, to V1 and the visual association areas, these models start with small pieces of encoded information and gradually fuse the information.

2.3. Architecture

Although we have a pretty good idea of the overall architecture of the visual cortex, the details within each cortical area and especially the details of their interconnection are not well known (Felleman and Van Essen 1991). This is the main reason why, contrary to the representation and the encoding of the stimuli, the architectures of the various visual recognition models demonstrate a great diversity. For some models, this differentiation is justified by the level of abstraction they are targeting at. For others, though, it's only an assumption that simplifies modeling. Although there is no agreed-upon classification of the various architectures we will attempt to give one that ties better with the rest of the model features we are reviewing in this section.

Given that the input to the visual system is a two-dimensional signal that is processed in subsequent stages, we can distinguish architectures in the horizontal and the vertical dimension. In the horizontal dimension we see differences among the architectures of the models in the way they compartmentalize the two-dimensional signal and the approach they later use to bind the processed pieces together. In the vertical dimension the most common variation regards whether a model is employing a single- or a multi-layer architecture. The two dimensions of the architecture that most models assume are usually independent. So for instance, we can see both monolithic and modular architectures with both single- and multi-layer models.

An example of a monolithic horizontal architecture is the model by Hinton et al. 2006. This model assumes a single sensory input for the whole image. Despite

the fact that it uses multiple layers to process the information in both a bottom-up and a top-down fashion, it never breaks the unity of the visual stimulus. Apart from being biological implausible, this type of horizontal architecture has difficulties in scaling since the dimensionality of the system increases with the square of the input (Geman and Bienenstock 1992). By contrast, most of the models of visual recognition assume some form of a modular horizontal architecture. In this case the compartmentalization can have several forms like, receptive fields (Fukushima 1988, Edelman 1996, Mel 1997, Wallis and Rolls 1997, Riesenhuber and Poggio 1999, Serre et al. 2007b), patches (Nelson and Selinger 1998, Amit and Geman 1999), attentional windows (Anderson and Van Essen 1987, Olshausen et al. 1993), or even fragments of the image (Ullman and Sali 2000, Ullman et al. 2002, Ullman 2006). However, more critical in a modular architecture is not as much the name or the type of the constituents but the approach the model is using to bind them together. This is where the so called vertical architecture kicks in.

In the vertical dimension most models have either some form of a layered structure or a single layer that incorporates an extra functionality. Characteristic examples of a vertically flat architecture are the models by Nelson and Selinger 1998, Edelman 1993, and Ullman and Sali 2000. These models do not focus on the biological plausibility of the architecture so they use a single layer on which they apply a series of computer vision processes without attributing them to a particular stage of visual processing. So typically, they use a monolithic vertical architecture but it is so complex that if we wanted to interpret it in biological terms we would probably couldn't do it without resorting to some form of a multi-layer

architecture. On the other side of the spectrum, the models that exhibit some form of modularity in the vertical dimension have either a stack of layers or a hierarchical architecture. For instance, the model by Anderson and Van Essen 1987 belongs to the first category. Its layers serve the purpose of routing the information from the bottom to the top instead of performing a hierarchical binding. By contrast, most of the models with a modular vertical architecture, like Neocognitron (Fukushima 1988, Fukushima 2005), VisNet (Wallis and Rolls 1997), and HMAX (Riesenhuber and Poggio 1999, Jiang et al. 2006), use the sequence of layers to gradually integrate the information and arrive to a final percept at the top layer.

A modular architecture, though, does not only mean an information binding that occurs between adjacent layers. The brain architecture implies a rich interconnection among layers as well as among neurons within a layer (e.g interneurons) that play a significant role in the associative nature of this complex system (Thomson and Bannister 2003). Although there are some models that make a first attempt to account for this kind of connectivity (e.g. Wang 2001) most models choose not to touch this aspect of the cortical architecture.

2.4. Processing flow

The brain architecture implies a system with multiple components that are interconnected and interdependent. In the visual system the graph of the various visual areas is densely connected with almost 40% of the cortices having links with all other areas and the majority of them being reciprocal (Felleman and Van Essen 1991, Braitenberg and Schuz 1998). This connectivity wouldn't be a big

problem if we were dealing with a directed acyclic graph. In that case the flow of processing would only depend on the point of entrance and the routing at each node. However, the brain circuitry involves many loops, both thalamo-cortical (Mumford 1991) and cortico-cortical ones (Mumford 1992). Hence, the flow of processing is not merely a function of the input and the routing, but a function of the time too. The more time we allow for the activity in the system to propagate across the network the more interactions among the areas will occur. Thus, the behavior that emerges is not the result of a simple input-output system but the outcome of a dynamical process that may or may not converge to a final state. And this, of course, is a matter of dynamics – the time evolution of physical processes. Therefore, *time* is a critical parameter that constraints the flow of processing and has significant subsequent implications on the types of behavior a model can exhibit. The flow of processing as a direct consequence of the temporal dynamics is probably one of the major differences among the various models of visual recognition.

Behavioral models work in a more abstract level and do not usually provide a detailed account for the flow of processing. Computational models, on the other hand, provide a more formal description of the processing flow but they take a dichotomous position towards *time*. On one side we have models that decide to ignore it at all and adopt some kind of feed-forward processing, while on the other side, we have models that incorporate time in various forms usually as feedback or recurrent activity. The reasoning behind fully dropping the temporal component of a dynamical system like brain is based on the argument that visual recognition is too fast for feedback loops to have a significant effect in the

processing. The feed-forward models do not claim, of course, that feedback connectivity does not play some role in visual recognition, just that the speed of processing imposes a constraint to the modeling and makes feedback an unnecessary complexity. The fact that experimental evidence from EEG studies (Thorpe et al. 1996) have shown that humans can perform object recognition withing 150 ms, which is comparable to the latency (150-270 ms) of responses of IT units (Fuster 1990), is interpreted by some researchers (e.g. Riesenhuber and Poggio 2000b) as a strong indication for an *immediate* object recognition. However, they agree that this doesn't rule out the use of feedback processing. In a series of experiments that Serre et al. 2007a did, they found out that a feed-forward mechanism could only model the performance of humans when the possibility of feedback and top-down effects was eliminated with the use a stimulus onset asynchrony that was less than 50 ms. Apparently, the role of feedback connections is not to be neglected if a more complete account of processing is what we seek.

Wallis and Rolls 1997 argue for a feed-forward visual recognition on the basis of a speed of processing account too. Their argument is based on neurophysiological studies by Tovee et al. 1993 that examined the responses of face selective neurons in the temporal lobe of rhesus macaques. Information theoretical analysis on neuronal spike trains in this study showed that a firing period of as little as 50 or even 20 ms is sufficient to give a reasonable estimate of the firing rate of the neuron. By doing a principal component analysis they showed that the information available in a short window of only 50 ms of firing rate can be as high as 84% of the information available in a firing period of over

400 ms. Even a window as small as 20 ms could account for 44% of the total information. In general, this analysis provided evidence that the start of the neuronal response provides a reasonable proportion of the total information that would be available if a long period of neuronal firing were utilized to extract it. There is also some evidence from visual backward masking experiments that a cortical area can perform the necessary computation for visual recognition in as little as 20-30 ms (Rolls and Tovee 1994). All this evidence combined led Wallis and Rolls to conclude that a feed-forward processing should be sufficient for visual recognition. However, a more recent model (Deco and Rolls 2004) makes a heavy use of both feed-forward and feedback connections, so it seems that they value more the role of feedback connections after all.

In either way, none of the above evidence regarding the speed of processing is strong enough to support the idea that feedback processing is not participating in visual recognition, most probably the opposite. An information content of 40 to 80 percent in the first milliseconds is far from a complete 100%. It may be true that the initial response is more critical but that doesn't mean that the rest of the processing is useless or purposeless. For instance, the more time we allow for processing the more complex tasks we are able to solve. As for the latencies in the visual areas, given that feedback activity cannot be easily isolated when running experiments, it's quite tricky to assume that the measurements are free of the effects of the cortico-cortical loops. To the contrary, it's very likely that the latencies reported include the contribution of the feedback connections as well. Besides, there exist recurrent connections within visual areas that are involved in feedback processing without having to travel back to distant visual areas. And

last, there are analyses that suggest that even the local recurrent neocortical circuits can produce very rapid dynamics with small latencies (Treves 1993). Hence, feed-forward processing is not necessarily the only way to achieve the observed speed of recognition.

Moreover, most models of visual recognition separate the process of recognition from learning and attention. It's hard to imagine how a model without a feedback process or a top-down influence could engage in visual learning, attention, storage and recall of stimuli. This may simplify things and help analyze independently the various visual processes but is biologically implausible and unrealistic. Even if it works for simple cases its limitations will uncover in the course of time. In this direction there is evidence from a complexity level analysis of visual search that speaks against a bottom-up processing (Tsotsos 1990). Architectural constraints suggest that attentional mechanisms that perform top-down tuning and selection of the stimulus are necessary during visual processing.

Overall, the flow of processing causes a major split in the models of visual recognition because it differentiates them substantially from each other. The majority of models probably assumes a feed-forward processing with the most representative examples being Poggio and Edelman 1990, Perrett and Oram 1993, Neocognitron (Fukushima 1980, Fukushima 1988, Fukushima 2005), SEEMORE (Mel 1997), VisNet (Wallis and Rolls 1997), HMAX (Riesenhuber and Poggio 1999, Jiang et al. 2006), and Ranzato et al. 2007. On the other end, some of the most representative models that apply a feedback processing are

Mumford 1992, Olshausen et al. 1993, Rao and Ballard 1997, Deco and Rolls 2004, and Hinton et al. 2006. The feed-forward models have a sequential processing which is simpler and more intuitive so one can find many similarities among them. By contrast, in the case of feedback models, there is no general agreement on how the recurrence should work (within cortical areas, between cortical areas, both, etc.). Each one implements its own approach and one can see many differences among them.

2.5. Learning

Learning in visual recognition is not a domain of debate or dichotomy among the various models. Most of them choose an approach that better suits their own architecture and not some general criteria. This is mostly due to the fact that the macroscopic properties of learning are not fully understood. We do know many microscopic details regarding synaptic plasticity and long term potentiation but when it comes to visual recognition which involves large groups of neurons and complicated circuitry it is hard to make any strong claims. Moreover, since all models of visual recognition work with neuronal spike rates that are easier to simulate and avoid the intricacies of spike-timing dependent plasticity (Markram et al. 1997, Zhang et al. 1998), any debate regarding the actual neural basis of learning would be highly speculative.

In a more abstract, computational, level of description where most of the learning of the visual recognition models takes place, the main classification of learning mechanisms is between supervised and unsupervised methods. Supervised methods are easier to control but are not biologically plausible since they require

training examples that are labeled. Unsupervised or adaptive learning (Kohonen 1982), on the other hand, seems much more plausible but is of limited use to the task of visual recognition since it works as a method of clustering or dimensionality reduction that requires extra assumptions in order to account for recognition. An alternative approach is a hybrid method called semi-supervised learning (Blum and Mitchell 1998) that combines these two approaches. It first uses labeled examples to generate the best possible classifier and then gradually incorporates into the model its best predictions on the unlabeled data. A different type of learning that has both a biological underpinning and a computational realization is Hebbian learning (Hebb 1949). It is a form of associative learning that works in a distributed self-supervised way and, although it's not as powerful as the other methods, it has been used successfully for constructing content-addressable associative memories (Hopfield 1982, Anderson 1993).

All these types of learning have been used in diverse ways in the various models of visual recognition. For example, Neocognitron (Fukushima 1988) is composed of layers of simple (S) and complex (C) cells that are arranged alternately in a hierarchical network. In this model only the layers with the S cells have their input connections modified through learning and the training can be either supervised or unsupervised (both methods have been proposed). A similar approach is followed by HMAX (Riesenhuber and Poggio 1999) which also has alternate layers of simple and complex cells. In this case the S cells show a Gaussian-like tuning and can be trained in an unsupervised fashion with a variation of the Hebbian rule while the C cells perform a nonlinear max operation and therefore their synaptic weights are all uniform and fixed to 1. A more recent modified

version of the same model (Jiang et al. 2006) is using an extra layer on top of the hierarchy in order to perform a supervised categorization of the features extracted from the view-tuned units.

In a rather different model Hinton et al. 2006 use multiple layers of Boltzmann machines (Hinton and Sejnowski 1983) to create a deep belief network. The top two layers form an associative memory while the remaining hidden layers are trained with an unsupervised method called wake-sleep algorithm (Hinton et al. 1995). Although the learning algorithm is unsupervised it can also be applied to labeled data by learning a model that generates both the label and the data. Another example of a layered architecture (Ranzato et al. 2007) is having two stages, one for the encoding of the stimuli and one for the decoding, where learning proceeds in an EM-like fashion by using a gradient descent algorithm in a supervised mode. A more complex model with a multilayer architecture (Deco and Rolls 2004) is using both feed-forward and feedback connections, as well as intra-layer lateral connectivity. The feed-forward connections in this case are trained with Hebbian learning while the back-projections, which are symmetric and reciprocal in their connectivity with the forward connections, are set to a fixed (single parameter) fraction of the strength of the forward connections. As for the intra-modular local competition it is implemented as a lateral inhibition in a local neighborhood of neurons with a Gaussian-like weighting that is a function of distance.

Apparently, there is a rich variety of learning methods and, unlike the other principles we reviewed so far, there seems to be no consensus on the approach

to be taken. Furthermore, it is obvious that learning is not a standalone issue but is highly related to decisions regarding the model's architecture and the flow of processing. For example, Hebbian learning is more suitable for bidirectional connectivity and recurrent processing while a supervised learning is easier to apply to omni-directional links and feed-forward processing.

3. The proposed approach

In this thesis we present a computational approach to visual recognition that is based on a large scale dynamical system with a hierarchical architecture that has modular layers. The system employs a recurrent processing scheme applied directly to low-level visual information and generates a view-based distributed representation of visual patterns. The proposed framework is in agreement with the concepts of topographical representation and hierarchical processing that most models advocate but introduces a new paradigm regarding the flow of processing and the intra-layer communication and learning. Overall, it puts an emphasis on the information fusion that occurs within a particular stage of processing, an aspect that is less explored and usually overlooked despite being ubiquitous in the visual system and the brain in general.

The key idea behind the proposed approach is coming from the work on Network of Networks (Anderson et al. 2007, Anderson and Sutton 1995). Conventional neural networks do not scale easily to larger dimensions, so the workaround is to create large scale networks by integrating small ones. At an abstract level, this idea is quite simple and there is plenty of evidence that the brain circuitry is following it too. However, there are many details (e.g. coupling of the networks, architecture, etc.) that can take many different forms which makes the proposal quite challenging. In the Ersatz Brain Group at Brown University we have worked on this approach for many years, mainly in the domains of language and cognition. This thesis was an attempt to apply the concept of Network of Networks to the domain of visual perception. It started a few years ago with the

modeling of pop-out phenomena in visual attention, and through various modifications over the time it evolved to the current form. One thing it showed is that the overall approach is quite generic and can be used equally well to model various modalities. But the major contribution of this thesis to the work on the general concept of Network of Networks, is not as much the domain of application. What makes it more valuable is the experimentation with a fully dynamical system. Due to the high computational demands of the proposed idea, it was quite difficult in the past to test very large structures without resorting to some sort of simplification (e.g. discretization of attractors, scalar couplings, substitution of automata for dynamics, and others). In the current work, with the use of High Performance Computing, we were able to fully implement all aspects of dynamics and of parallel distributed processing for very large scales, even up to millions of networks. But most important, we had to resolve numerous issues that arise during large integrations of dynamical systems that may seem trivial when a simplification is applied but are quite crucial when the complex dynamics comes into play.

Next, we present the approach we take in the proposed framework regarding all major dimensions of the visual recognition models as we discussed them in the previous chapter.

3.1. Representation

The proposed framework relies on a view-based representation and makes no assumptions about the structure and the configuration of the visual patterns. The visual stimuli correspond to two-dimensional images taken as snapshots from a

particular viewpoint and they are generated from a high resolution sampling and subsequent digitization of the raw image signal. There is no preprocessing in the form of clustering or segmentation that could extract more abstract visual information like object fragments or image components that facilitate a higher level non-pictorial representation. This absence of any internal representation of the visual patterns to be learned entails that the performance of the system is highly dependent on the richness and the complexity of the visual stimuli. This is also true for the capacity of the visual recognition as well as the invariance of the system towards viewing conditions. Everything will depend on the number and the content of the views that the system has the chance to be trained on. On the other hand, this type of representation has no bias towards a particular visual configuration or class of patterns in either 2D or 3D space. In other words, the system is not coming with any pre-built knowledge about the visual representation of objects, and whatever representations it will form they will be the mere result of the training and the statistics of the visual world.

3.2. Architecture

Choosing an architecture is probably the most difficult decision to be made since it will set the scaffold for everything else to be based upon. If we go after an extremely detailed approach that is modeling properties at a level as low as the molecular one (e.g. the Blue Brain Project, Markram 2006) then we run the risk of never being able to interpret the findings in a more abstract *computational* level of description. If, on the other hand, we take a serial discrete view of behavior biased by the computer metaphor, we run the risk of not being able to tell

whether the findings have anything to say about the actual brain. Keeping the balance between a highly distributed system and a system that is devoid of the unnecessary complexities of the neuronal level is the primary goal of the proposed architecture. In order to bridge the gap between neural activity and behavior we need a multifaceted architecture that provides explanations at multiple levels. The architecture must retain its low-level biological plausibility while at the same time being able to account for high-level cognitive phenomena. In such a meso-scale approach we cannot neglect the principles dictated by any level of abstraction nor can we dissociate these levels like Marr did in his three-level analysis by claiming an independence between computation and implementation (Marr 1982).

To face all these challenges the proposed framework is based on a multi-layer modular architecture with a hierarchical organization. In the horizontal dimension each layer has a spatial modularity with the constituent units exhibiting a massive communication that produces a highly hetero-associative network. In the vertical dimension the constituent modules of neighboring layers have the same functional interaction that they have within their layers but they are organized in a hierarchical fashion that produces different scales of abstraction. The most critical issue that arises from this approach is not as much the hierarchical organization as the specifics of the within-layer modularity and organization. This is the point where we'll put our focus both in theory and in the experiments. First, we will describe the structure of a single processing unit, then the organization of a whole layer, and finally, how multiple layers can stack on top of each other to form the full system.

3.2.1. The basic processing unit

Recent evidence from neuroscience favors the argument that neural computation involves a lot more than just an aggregate of individual neurons. The neuron doctrine is under continuous scrutiny and the experiments lately put a great effort on exploring the coordinated behavior of groups of neurons (by using several techniques like multiple-cell recordings, fMRI, etc.). Moreover, it has been known for a long time that there exist in the cortex columns with a discrete anatomical organization (Mountcastle et al. 1955). These columns seem to have a nested structure that contains about 50-100 minicolumns with each minicolumn having roughly 80 neurons. More important, though, is that these cortical columns exhibit a functional specialization and operate as discrete processing units. This is also suggested by their connectivity which is dense internally (i.e. along the vertical dimension of the cortical sheet) and more sparse externally (i.e. along the horizontal dimension in their communication with other columns).

In the proposed framework, the basic processing unit is a single cortical column and not a neuron. Our goal is to depart from the organizational level of networks of neurons (i.e. neural networks) and reach the level of network of networks. Although the full details of the organization of the cortical columns are not yet known, their functional role in some cases has been clearly revealed. One such example is the orientation selectivity columns discovered by Hubel and Wiesel 1962. In this case it is easy to abstract the functionality of an otherwise complex column with a simple neural network that performs a specific function. In our case this neural network will take the form of an auto-associative attractor network. The reason we take this approach is that we want a unit that is able to

learn its function directly from the stimuli, and we want this to happen in an unsupervised fashion too. Brain is a powerful associative system that learns with several variations of the Hebbian rule, so the choice of an auto-associative attractor network should be sufficient to model an abstracted version of a simplified visual cortical column. The idea that neural activity behaves like dynamical attractors has been suggested in the literature (Amit 1989) while there exists increasing evidence that many areas in the brain, like hippocampus (Wills et al. 2005) and olfactory bulb (Niessing and Friedrich 2010), form attractor representations.

An attractor network (Figure 1) is a fully connected neural network with no input or output layers and no need for mapping patterns to labeled responses. A set of neurons is interconnected with Hebbian synapses and learns to associate a set of stimuli with themselves. Intuitively, it's a content addressable storage that is able to retrieve a partially completed or corrupted memory.

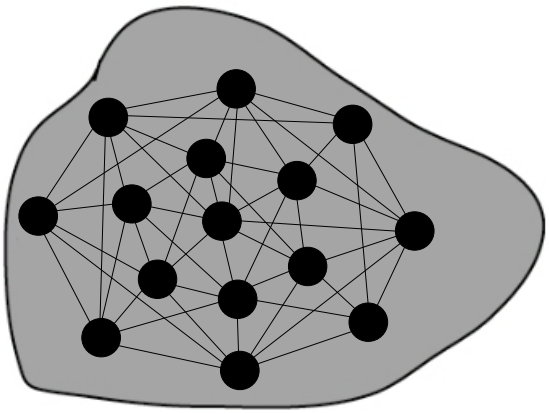


Figure 1: Abstract view of an attractor network

Formally, it's a dynamical system that is able to form stable attractors in a hyper-dimensional space that correspond to learned stimuli of the respective

dimensionality. For example, Figure 2 depicts the energy landscape that an attractor network has formed after learning four different patterns. The fixed points (in red) will attract any network activity that lies within their basin of attraction (depicted as dotted circle for one of them). If a corrupted or noisy variation of a learned stimulus is given as input to the system the attractor network will be able to retrieve the original stimulus by driving the network activity to the closest attractor.

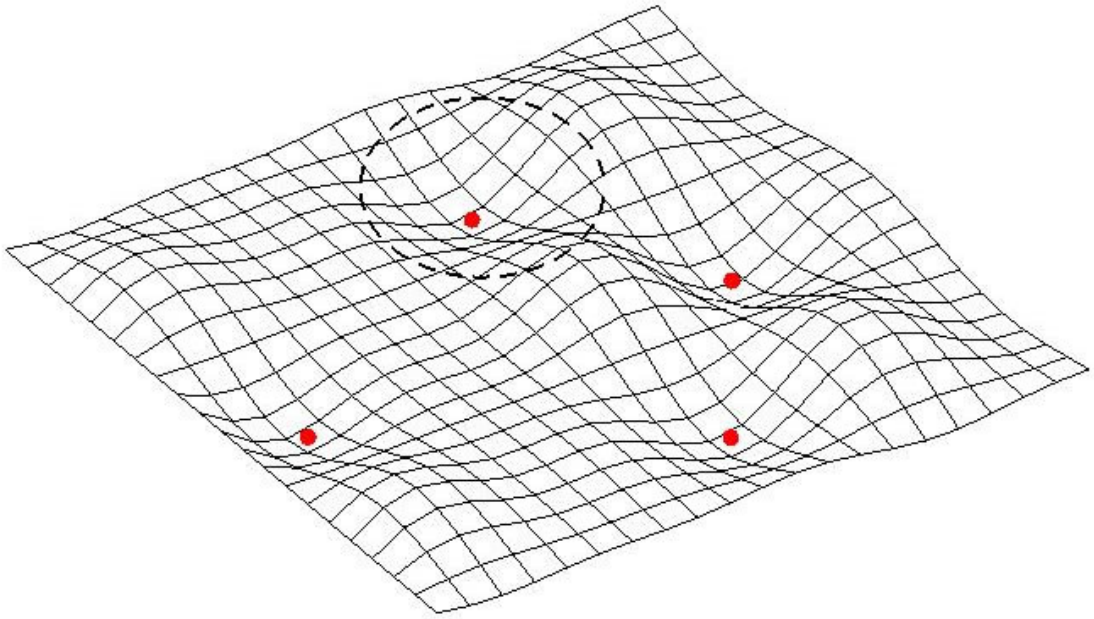


Figure 2: Energy landscape of an attractor network with four fixed points

3.2.2. The organization of a layer

Now that we have our basic building block, the attractor network, we have to see how we put many of these processing units together to form a single layer of the proposed framework. Figure 3 depicts an overview of the organization of this network of attractor networks (Dimitriadis 2008). Within a layer, the attractors

form a rectangular grid and connect to other attractors in their vicinity. For illustration purposes, in this visualization we only see a partial connectivity for only two of the attractor networks (depicted with dark grey color). The omni-directional arrows represent the axonal projections, from one cortical column to another, that are bundled in single links. The synaptic strengths of these grouped connections are assembled into matrices and depicted in the figure with small squares.

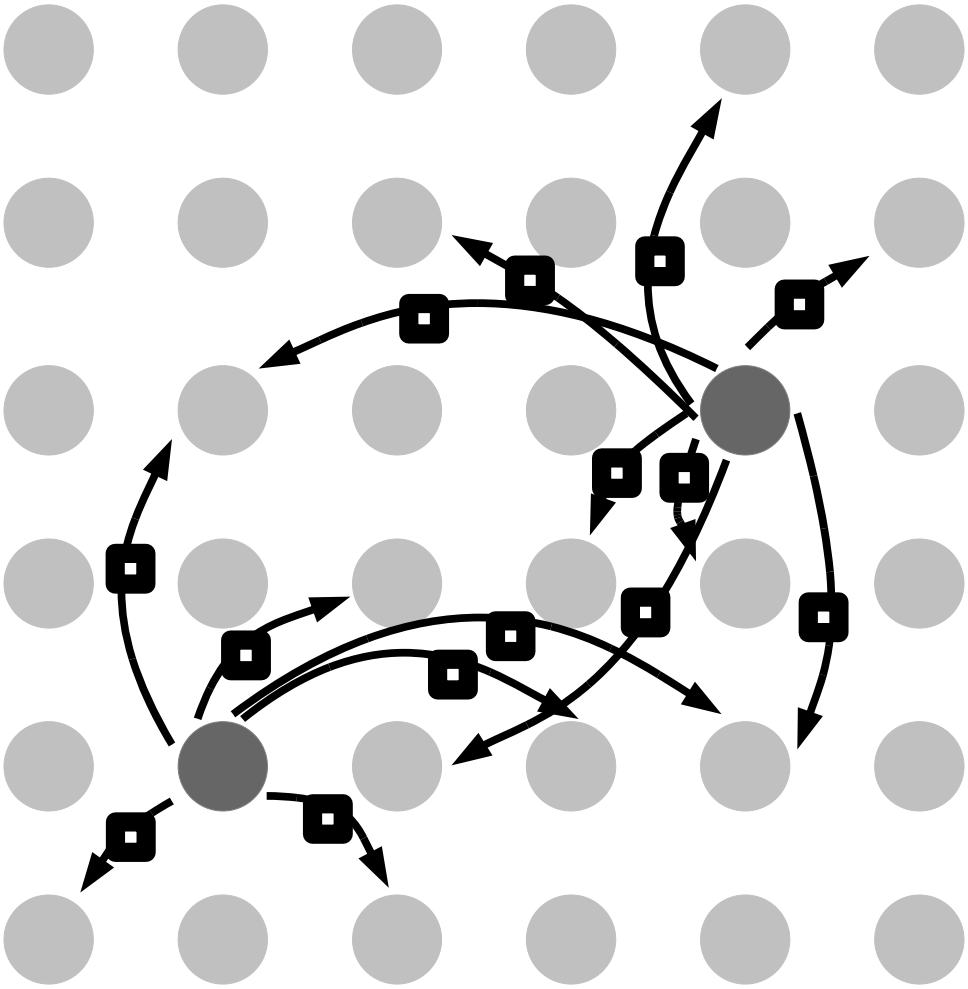


Figure 3: A network of attractor networks

If a stimulus is repeatedly presented to the system, the synaptic matrices will essentially get tuned to the neural activity of the corresponding cortical columns and will learn to hetero-associate the states of the respective attractor networks. So, the attractor networks have both internal and external dynamics. Internally, they work as local auto-associative systems that recognize small visual patches extracted from their receptive fields. Externally, they interact through hetero-association with a number of columns in order to fine-tune both their own and the distal network's dynamic states. This dual dynamics occurs simultaneously and continuously for all the networks in a given layer. Therefore, we have a hetero-associative network built from auto-associative attractor networks.

3.2.3. Multi-layer organization

This dynamical system can be extended in the vertical dimension by creating a stack of layers on top of each other. Each of the layers in this case will have the exact same internal organization that we've seen before. Regarding the inter-layer connectivity, the associations between columns in different layers will follow the same principle of hetero-associativity that we've seen for the intra-layer organization. Hence, the attractor networks will not only form horizontal hetero-associations within their layer but vertical hetero-associations that span across layers as well. All the hetero-associations in this approach, both horizontal and vertical, operate with the same principles. The only thing that changes is the content of these associations and their spatial distribution, two issues that we will discuss further later. As for the number of layers, it is a variable that depends on the scales of abstraction that we want to implement and therefore it's a function of the specifics of the encoding that we apply to the stimulus. However, the

power of the proposed framework can be shown with as little as one layer only.

Figure 4 gives an illustration of the full-fledged implementation of the framework. In the first layer each of the attractor networks (green columns) processes its own local receptive field (yellow cone) and constructs through auto-association a set of attractors that correspond to the set of the recognized stimuli the respective network can handle. At the same time, the whole bottom layer, in a collective way that we described in Figure 3, is forming a hetero-associative network that binds together the internal states of the attractor networks forming by that an assembly of networks instead of simple neurons. An example of such

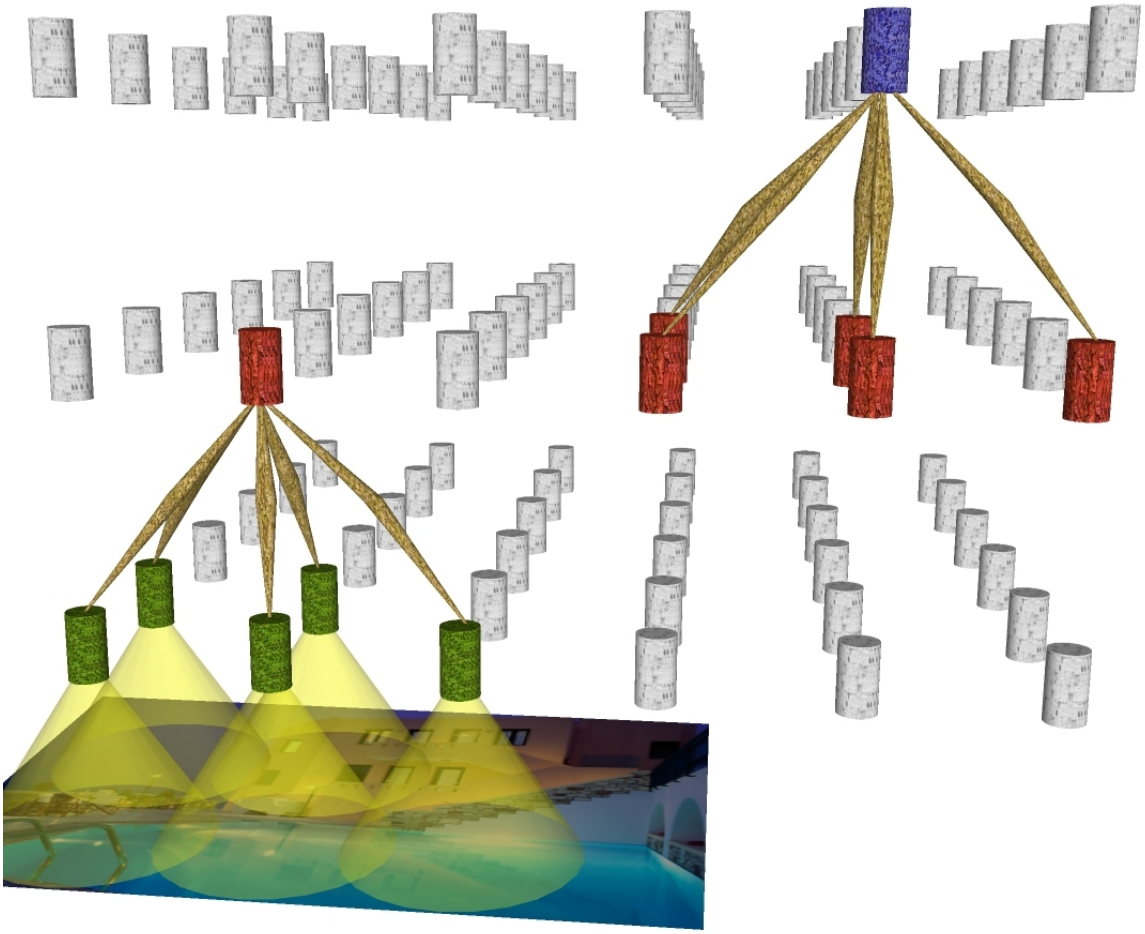


Figure 4: Hierarchical organization of the hetero-associative network

an assembly might be the five green columns depicted in the bottom layer of the scheme in Figure 4. These attractor networks might have responded to some coherent configuration that existed in the stimulus (e.g. an object part) and formed an assembly of their internal states. In the vertical dimension, if we assume that the layers are organized in a hierarchical fashion as depicted in the figure with the many-to-one connections between layers, as we ascend the levels of the hierarchy we would expect assemblies of a larger scale to gradually be formed. So, for instance, the red columns in the second layer pool several hetero-associations of the layer beneath (green columns) together in order to form an assembly of a higher level of abstraction. In a similar way the blue columns in the third layer might pool different sets of hetero-associations from the second layer (red columns) and form an even higher assembly of attractor states, and apparently, this can go on and on to higher levels of abstraction and scale. As we reach higher levels of this hierarchy the attractor networks presumably encode larger parts of a visual stimulus. This also means that in the higher levels we would expect to see assemblies that consist of a small number of hetero-associative networks or even of a single column that encodes a whole object in a very localized way. Evidence for this kind of representation is coming from physiological data in the inferotemporal cortex, one of the later stages of processing in the ventral visual stream (Tanaka 2003).

In the current version of the framework we put the focus of our experimental work on a single layer of the proposed architecture in order to acquire a comprehensive analysis of all the details pertaining to such a complex dynamical system. Although we won't have the resources to demonstrate the behavior of an

immense network of networks that is the result of a vertical integration of many such layers, we believe that the essence and the power of the proposed organization would nevertheless become apparent. Moreover, even with a single layer is still possible to demonstrate some of the effects of the multitude of abstractions by experimenting with a variety of stimulus resolutions and scales.

3.2.4. Connectivity map

In the previous sections we described the general architecture of the system and specifically the structure of the basic processing unit, the organization of the layer, as well as the integration of a set of layers. What remains to complete a full description of our architecture is the interconnectivity of the units. A major scientific question towards this direction is the genesis of the synaptic connections in the brain. The resonance hypothesis, probably the first theory on this issue, suggested that neuronal connections during stages of early development were nonspecific and the cortical circuits were formed by the activity-dependent rewiring of the initially random connections. However, a series of experiments on the frog's visual system held in the 1940's gave rise to a contrasting hypothesis called chemoaffinity (Sperry 1963) which is the most widely accepted theory of neuronal wiring today. According to this theory, growing axons recognize chemical markers produced by target cells and by that they connect to specific neurons or groups of neurons. Thus, connections are not initiated randomly but are predominantly predetermined into the identity of the differentiated cells. This means that in the early stage of development circuits of the brain are largely hardwired.

However, this only answers the question of *how* the genesis of the connections is realized and not *what* kind of connectivity pattern is produced. Despite the large body of statistical data regarding neuronal connectivity (Braitenberg and Schuz 1998), our knowledge in this matter is mostly quantitative. We know for instance that cortex has a dense connectivity with roughly 40% of the areas having reciprocal connections with other areas but we don't really know any topographic details regarding these connections other than which areas connect with each other. Therefore, there is a need to devise and test a set of plausible and feasible connectivity patterns. As the chemoaffinity hypothesis suggests these patterns will be fixed for any given test. However, we will be experimenting with several patterns in order to study a variety of properties that will possibly help us discern the suitability of the various connectivity maps for certain situations.

The simplest connectivity map is probably a fully connected graph of all the attractor networks. Although this kind of map has been already proposed for classic neural networks and has been shown to work quite well for small-scale practical applications (Kohonen 1982), it is biologically implausible and computationally extremely expensive for a large scale implementation of the size of the proposed framework. A more plausible map would suggest a full connectivity only for the units within a certain radius R[1]. If we assume a rectangular grid (as the one we described earlier) with each node having 8 neighbors within its layer and 26 neighbors in the multi-layer structure, the degree of connectivity would scale according to $4R^2+4R$ for a single layer, and

1 The unit of radius R in this case is the distance between two adjacent nodes in the hetero-associative network and not the real distance of the units in space.

according to $8R^3+12R^2+6R$ for a non-hierarchical multi-layer organization. Given the rate of growth of these numbers and the fact that cortical neurons connect to an average of 10^3-10^4 other neurons, it is obvious that this type of connectivity can only be feasible for a hetero-associative connectivity of a rather small scale. Of course, with a more plausible hierarchical organization connectivity constraints would reduce the cubic growth but even in this case the degree of connectivity would retain a quadratic form which is highly unrealistic for a full connectivity among the attractors in a neighborhood that exceeds a small range.

A more efficient connectivity pattern can have a Gaussian, or a similar distribution, where the density of the connections tapers as a function of the distance *R* from the column. One advantage of this pattern is that the number of hetero-associations can grow in a fully controlled way by manipulating the mass and the width of the distribution of the connections. More important, though, is that it allows the attractor networks to form associations at much longer distances than the previous connectivity patterns would permit with the given limitations. Of course, this ability comes at the expense of a connectivity that becomes sparser as we depart from the origin. However, since the target columns are not isolated processing units but they form hetero-associations of their own *neighborhood* as well, it should theoretically be possible for a column to have an adequate interaction with a distant area without preserving a full connectivity with all possible target units in it. Due to the dynamics of processing an attractor network has the ability to receive the distant activities in a progressive way and this allows the target column to incorporate into their communication its gradually formed neighboring activity.

A connectivity pattern is not necessarily exclusively spatial but can combine a functional role too. Psychophysical evidence regarding lateral interactions between receptive fields indicates that the interaction of receptive fields with the same functionality is a lot stronger when they are situated along vertical and horizontal meridians as contrasted to random locations relative to each other (Polat and Sagi 1994). If we wanted to convert this finding to a proposal for a possible connectivity map, that would mean that the distribution of the hetero-associations cannot only depend on the spatial properties of the attractor networks but it should take into account their function too. But according to our initial hypothesis, the function of each attractor network is to recognize a set of forms in its receptive field. So, the connectivity of the networks cannot be static. It should either have a dynamic switch depending on the attractor that a network is into at every time, or we should assume a connectivity where the hetero-associative matrices contain functionally-inactive hetero-associations. From a biological point of view the first approach seems highly implausible whereas the second approach is much more realistic. Besides, as we will see later in the section of learning, there is no process of associative learning that could store in a single hetero-associative matrix all possible hetero-associations. So, the idea of having some of the hetero-associations set a-priori to be inactive serves both a statistical and a biological purpose.

3.3. Stimulus encoding

Regarding the stimulus encoding, we follow a local scope approach that is based on receptive fields with a uniform distribution over the input signal. Given that the

proposed framework is focused on the processes underlying the integration and the fusion of the visual information, it is necessary to make as few initial assumptions as possible. We choose to start with the most biologically plausible encoding, the one that resembles the receptive fields of the visual system, and do not rely on any kind of informative image segments or components that may facilitate processing but will defer the crucial issues and will leave a lot of important questions unanswered.

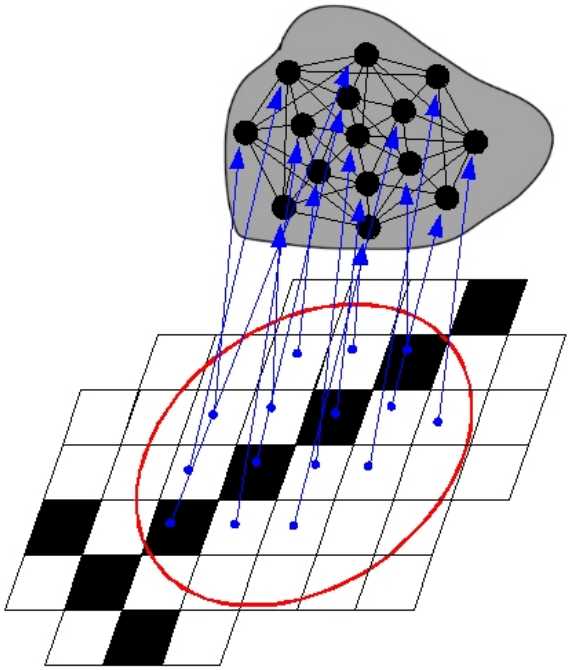


Figure 5: A processing unit over the respective receptive field

Figure 5 illustrates a single processing unit (i.e. the attractor network) and the respective receptive field that receives input from the underlying image patch. The signal in this case is a binary image that may result from an edge detection filtering process. This is a simplified depiction of a generic receptive field (RF)

based on the discoveries by Hubel and Wiesel 1962 and we will be using it throughout this document.

3.3.1. Geometry of the encoding

Figure 6 depicts the geometric and topographic characteristics of the receptive fields and the encoding of the visual stimulus. For illustration purposes the shape of the receptive fields in this figure is round. The proposed framework, however, allows for elliptical RFs of any variable size. For computational efficiency, and in order to simplify the vector calculations, we are going to approximate the ellipsoid of an RF having axes s_x and s_y with a rectangular patch of $s_x \times s_y$ (depicted with red frames in the figure).

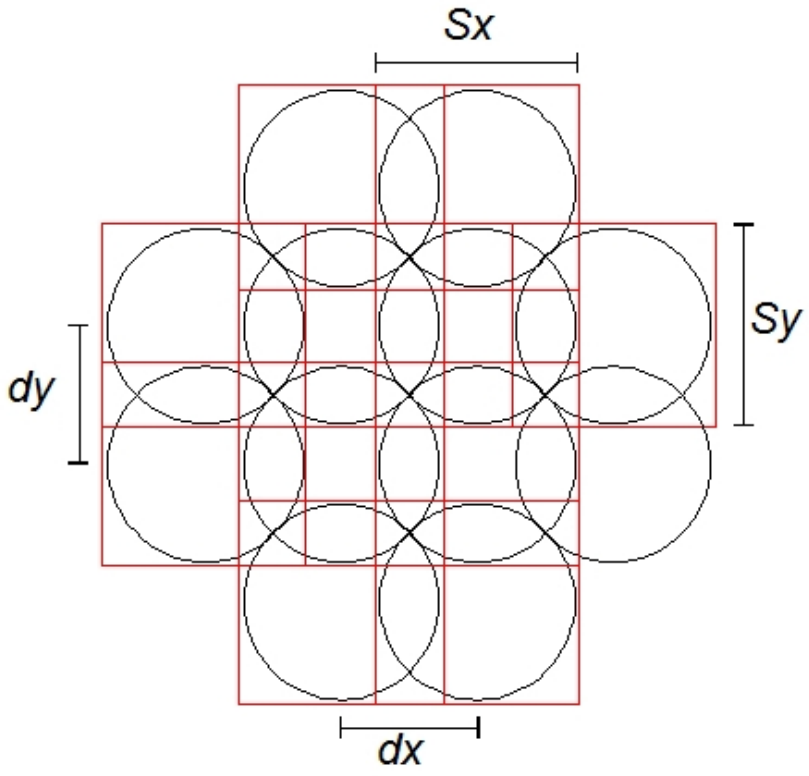


Figure 6: Geometric characteristics of the stimulus encoding

This may modify a bit the properties of the local processing but, overall, it shouldn't have any effect in the macro-scale processing. The receptive fields within a certain stage of processing in the visual system (e.g. V1) have a variable size that increases with the distance from the center of the projected visual field. However, for small visual angles (less than 5 degrees) the size of the receptive fields can be considered quite uniform across a visual map[2]. This is especially true for the part of a cortical area that undertakes the foveal visual field where most of the processing occurs. So it wouldn't be too risky to assume that for visual recognition – unlike visual attention and other mechanisms that recruit the peripheral parts of the visual field – a visual patch roughly the size of the fovea with receptive fields of equal size is sufficient for our modeling purposes.

Regarding the macro-scale properties of the topography of the encoding, we will be using a uniform grid of receptive fields with a variable but equal degree of overlap across the visual maps. The proposed framework allows for the quantities *dx* and *dy* (Figure 6) to vary depending on the purpose of each experiment but their value has to remain fixed for all RFs for any given test. This means that the receptive fields will always be forming a rectangular lattice. Although the topographical arrangement of the receptive fields in the cortex demonstrates a hexagonal pattern of local alignment this is almost certainly a result of a developmental biological constraint and there shouldn't be any reason for not adopting a technically simplified rectangular alignment. Last, we will

2 In the case of V1, in particular, the increasing size of the receptive fields as we depart from the point of fixation serves mostly as a compensation for the lower density of the ganglion cells in the periphery of the retina. So the actual number of cells per receptive field, even for bigger visual angles, is probably less variable in this preliminary stage of processing.

always assume some degree of overlap between adjacent receptive fields that will provide the necessary redundancy and robustness in the system. So the distance of the receptive fields will be, $0<dx<Sx$, and, $0<dy<Sy$.

3.3.2. Content of the encoding

Regarding the content of the encoding this varies depending on the level of processing. In the initial layer of the hierarchal network the encoding is directly related to the stimulus itself. In the layers above, though, the encoding in each level is related to the network assemblies that are formed in the neighboring layers and not the stimulus itself. So, the content of the encoding in the bottom layer consists of some feature of the visual signal but the content of the encoding in the other layers consists of the associations that are formed in the smaller or larger scales (lower or higher layers respectively). Therefore, the increasing complexity of the hierarchical network is not related to some integration of the stimulus itself, as many models suggest, but to some integration of the associations that are formed after the initial encoding of the signal. This means that the *code* in the initial layer will be some sort of representation of the stimulus information but the *code* in the layers above will correspond to an abstract and probably arbitrary representation. This may sound counterintuitive but the evidence from electrophysiological studies regarding neural correlates that become increasingly scrambled in the higher visual areas provides a strong support for it. In this phase, our primary goal is to demonstrate the formation of the assemblies in the initial stage of the encoding. Once these assemblies are formed, it is quite easy to imagine how a hierarchical system of associations can

build up on top of them. Therefore, in the current work, we are only going to explore the initial layer of this hierarchical encoding.

Regarding the processing of the stimulus itself, we'll be using a single feature map of image gradients. Edge detection provides a highly informative signal that has been shown to be extremely valuable for human vision (Shapley and Tolhurst 1973). Specifically, we preprocess the stimulus with the Canny edge detector (Canny 1986) that produces a binary image to be fed to the receptive fields of the initial layer. As we will discuss later, the receptive fields can either be trained on a set of natural images or on a set of artificial orientation bars with a predetermined bandwidth and angular resolution. The architecture of the framework allows also for multiple feature maps to associate with each other in a similar fashion that the hetero-associations between layers are formed. However, this goes beyond the scope of the current project so we're not going to experiment with any visual properties (e.g. color, texture, etc.) other than gradients.

3.4. Learning

The biggest challenge for learning in a visual recognition system is how to maintain a specificity while allowing for some variance. With a monolithic system this is impossible to achieve for practical situations because the variability of the stimulus grows extremely fast with the dimensionality of the input signal. With a modular system, though, it should in principle be feasible to maintain this balance and achieve the desired behavior. Moreover, if we want to preserve some biological plausibility, we would have to rely on unsupervised methods and

incorporate in one way or another the principles of Hebbian learning (Hebb 1949) that are prevalent in the domain of synaptic plasticity. So, what kind of learning could be used in this situation where the visual system consists of a large number of units that are organized diversely? The answer to this question, as we've seen in our review earlier, is very different from model to model, and in most cases there is no real effort to address it. Contrariwise, our goal is to propose a learning scheme that is biologically plausible and consistent with the possible variations of the architecture. It should also apply the same principles for all units and it should take into account the spatial and functional modularity which we believe is the core property of any visual recognition system.

If one accepts the hypothesis of a modular visual system (either as an aggregate of networks that is suggested here or of any other form) then one would expect that learning in such a system cannot be but modular as well. If learning shapes the structure of a system and conversely, the structure shapes its learning, then the system has to demonstrate some form of a hybrid learning that can account at the same time for both the intra-modular and the inter-modular plasticity. Learning, in other words, should be based on the same principles regardless of the scope and the function of the neurons it tunes.

In the proposed framework there are two distinct but interacting forms of learning. The first one, auto-association, takes place inside the processing units and its role is to generate a high-dimensional space of attractors for the local stimuli that are coming from the receptive fields of the respective attractor networks. The second one, hetero-association, takes place between the processing units and its

role is to generate a space of attractors for the pairwise relations between the attractors of the respective attractor networks. The two learning mechanisms operate with the exact same rule but on different inputs. Spatially, this means that every column holds an *internal* auto-association and a number of *external* hetero-associations with other columns. The auto-associations have a fixed seat but the hetero-associations span across multiple columns like the bundles of axons in the white matter. Functionally, the associations may have a discrete realization but their post-synaptic potentials, as we will see in a latter section, are integrated in the dynamics of the columns they are attached to. So, learning has a dual hybrid role that facilitates recognition at the local level while at the same time providing a connectedness to the whole network and thus, a coherence to the visual percept.

One can also consider the two learning mechanisms as a *passive* versus an *active* one. Local attractor networks through auto-association will *passively* learn a more or less universal set of small visual patterns (some kind of feature bank), while hetero-associations will learn to *actively* reinforce the most promising relations between these patterns (i.e. the ones that maximize the mutual information in the respective columns). By doing so, identification and categorization are not any more two different tasks of object recognition (Logothetis and Sheinberg 1996, Riesenhuber and Poggio 2000) but two different points on the spectrum of this hybrid learning. It is the trade-off between these two mechanisms that will define the capacity of the learning and hence the power of the recognition.

Formally, the auto-association is a square matrix ***A*** that performs a mapping of the *n*-dimensional input (coming from the receptive field) to itself. As for the hetero-association, it is a square matrix ***H*** that performs a mapping of the *n*-dimensional state of one column to the *n*-dimensional state of another. So, both matrices have the same size and can be trained with the same learning rule. The goal of learning is to adjust the weights of the association (i.e. the values of a matrix) in order to optimize the mapping of the input to the output. We have to note here that contrary to other approaches in visual recognition we don't assign any labels to the visual patterns. So, there is no need for a biologically implausible supervised learning that will map the visual signal to some form of symbolic output. Instead of learning the full joint distribution of visual patterns and their labels we learn to associate patterns with patterns. This form of associative memory constitutes the fundamental mechanism of Hebbian learning in the brain.

In practice, learning is the first process towards the formation of the visual recognition system. In order to train the system and generate the matrices ***A*** and ***H*** for all columns of the framework, we need to have a set of training images that will be fed to the system. This set of images is presented repeatedly until all the associative matrices have been sufficiently tuned and converged to a stable configuration. In computational terms, there are many algorithms that can adjust the values of such a matrix. Two of the most known variations of the generalized Hebbian learning rule are the Perceptron (Rosenblatt 1958) and the Widrow-Hoff (Widrow and Hoff 1960) algorithms. Their main difference is that the second one, also called the *least mean squared (LMS)* algorithm or *delta rule*, is using a linear activation function while Perceptron is using a threshold function. In biological

terms, the threshold activation function is more representative of the firing of neuronal action potentials but, in computational terms, the linear activation function gives a more powerful learning rule[3]. Therefore, as a matter of convenience, we will be training the associative matrices of our system with the *LMS* algorithm.

In the case of auto-associative learning the update of matrix ***A*** can be written as

$$\boldsymbol{A}(t+1)=\boldsymbol{A}(t)+\eta\,\boldsymbol{e}_i\,\boldsymbol{x}_i^T \tag{1}$$

$$\boldsymbol{e}_i=\boldsymbol{y}_i-\boldsymbol{x}_i \tag{2}$$

where $\boldsymbol{A}\in R^{n\times n}, \boldsymbol{x}_i\in R^{n\times 1}, \boldsymbol{y}_i\in R^{n\times 1}, i=1,\ldots,m$, η is the learning rate, m is the number of training patterns, ***x*** is the input to the algorithm, n is the size of it, and ***y*** is the output that the algorithm generates at time step t

In the case of hetero-associative learning the update of matrix ***H*** follows the exact same procedure with the only difference that Eq.2 becomes

$$\boldsymbol{e}_i=\boldsymbol{y}_i-\boldsymbol{k}_i \tag{3}$$

where $\boldsymbol{k}_i\in R^{n\times 1}$ is the vector state of the target column of the hetero-association.

3.5. Dynamics

The dynamical aspect of processing is undoubtedly the most distinctive characteristic of the proposed framework. Contrary to the other properties we

3 A nonlinear activation function, like the sigmoid, is even more powerful and allows for more sophisticated but less plausible (biologically) learning rules like the error back-propagation

described so far, most of which could probably have been treated with a variety of alternative approaches, the dynamics assigns some unique attributes to the proposed system. The incorporation of *time* into the distributed parallel processing and the integration of the spatial with the temporal aspects of visual processing produce a system that is hard to simulate with a discrete symbolic approach. Furthermore, dynamics provides a means for a seamless and inherent handling of *time* which is a fundamental variable in behavioral modeling.

Before getting into the details of dynamics let us see the general idea behind it and the reason we need it. What happens if we don't use the dynamics at all? In this case the whole network of networks will degenerate into an aggregate of small independent pattern recognizers. Each of them would operate independently on a small set of visual patches and they would try to recover (in a single step) the memories that matches best their testing inputs. This is what most feed-forward pattern recognition systems do. The only difference is that we put many of them next to each other in order to divide the high variability of a large input into small ones that are easier to tackle. But the visual world is not random. We can take advantage of the relations among neighboring or more distal stimuli and reinforce the recognition performance. In other words, the system will be more efficient if the individual recognizers are not independent. If they were able to interact with each other they would have the chance to correct their recognition mistakes and produce better results. And the longer this interaction is the more corrections it can give. Therefore, we need a system that not only provides the couplings among the constituent networks but it has a

formulation that allows the networks to interact for as long as needed as well. This is where the dynamics comes into play.

The proposed framework is a large dynamical system that consists of many smaller and interacting dynamical systems. First of all, this means that the response of the system is time-dependent. There is no single output as the result of the application of some function to the input. The system follows a trajectory of states as a result of an iterative process. Second, there is no guarantee that this iteration will eventually converge to some final state, let alone a desired one. So, most of the time we need to make a decision about when to stop the process. And third, when the system reaches some desired state, it's a question of mathematical stability whether it will remain in that state or not. This is not necessarily bad; brain as a dynamical system is constantly going from one state to another. However, since we're only modeling a small part of a complex behavior, it's not always easy to tell whether a short stay in an otherwise unstable state can be considered sufficient. These are only a few of the problems we're dealing with and they give us a small idea of what it takes for a complex behavior to be realized in any dynamical system, either a biological or an artificial one.

The hetero-associative network of attractor networks is built upon the idea of assembling a large number of attractor neural networks. These networks act like content-addressable associative memory systems and can be trained to learn a number of states in a hyper-dimensional space. Some of the most known examples of this type of networks are the Hopfield net (Hopfield 1982), the Boltzmann machine (Ackley et al. 1985), and the BSB (Anderson 1993). For a

number of reasons, in the proposed framework we will be using the BSB network. It has real instead of binary units, the training of the auto-association matrix can be done independently with a variety of algorithms, it has a deterministic operation, and it doesn't exhibit any serious scaling issues.

The state $\boldsymbol{x}$ of a single attractor network evolves in time according to the following equation

$$\boldsymbol{x}(t+1)=f\left[\alpha \boldsymbol{A}\,\boldsymbol{x}(t)+\gamma\,\boldsymbol{x}(t)+\delta\,\boldsymbol{x}(0)\right] \tag{4}$$

where $\boldsymbol{x}\in R^{n\times 1}$ and $\boldsymbol{A}$ is the auto-association matrix that is generated according to the learning rule we've seen in a previous section. In each iteration the new state (for time step t+1) integrates the previous states of the system. The three parameters, α, γ and δ take positive values and control the behavior of the dynamics by weighting the contribution of the three components of the summation respectively. The first component is the contribution of the auto-association, the second one is the contribution of the previous state (and recursively of all past states), and the third one is the contribution of the initial state at time 0, that is the input to the system. As for f, it is a non-linear clipping function that enforces the stability of the network by limiting the activity of a state within the boundaries of a hypercube of radius 1. We will explain later why this is a necessary component of the dynamics.

$$f_i(\boldsymbol{x})=\begin{Bmatrix} -1, & x_i<-1 \\ 1, & x_i>1 \\ x_i, & other \end{Bmatrix} \qquad \boldsymbol{x}=\left\{x_1, x_2, \ldots, x_n\right\} \tag{5}$$

Parameter δ is a binary switch that controls the participation of the input to the dynamics. Some times we need a constant reinforcement of the input (e.g. in a pattern completion task) and sometimes we don't (e.g. in a case of a distorted input).

Parameters α and γ are critical for dynamics because the convergence of the network to an attractor depends on their values. If we think of the auto-association matrix ***A*** as a linear transformation, the set of stimuli that the system was trained on during the learning phase will constitute a set of eigenvectors with certain eigenvalues. The magnitude of these eigenvalues in combination with the parameters α and γ will define the dynamics of the system (Eq.4), i.e. whether the network will converge to an attractor. If an eigenvalue is less than 1 the respective vector will shrink each time we apply the auto-associative transformation. If the eigenvalue is greater than 1 it will grow. Therefore, if we knew the eigenvalues of ***A*** we could set the parameters α and γ in such a way that the contribution of the transformation is balanced appropriately by the past activity in order for the system to converge to some attractor. Luckily, this is not that hard to know. Actually, one of the properties of the LMS algorithm that we use for the learning phase is that it generates a matrix ***A*** where all the eigenvalues corresponding to the learned eigenvectors are having a value of 1. Figure 7 depicts the magnitude of the eigenvalues of a 100-dimensional system after training with 13 different stimuli. In this case most of the eigenvalues have a random value that ranges from 0 to 3. However, the 13 eigenvalues (numbers 76-89) that correspond to the learned stimuli (i.e. the eigenvectors) have an exact value of 1.

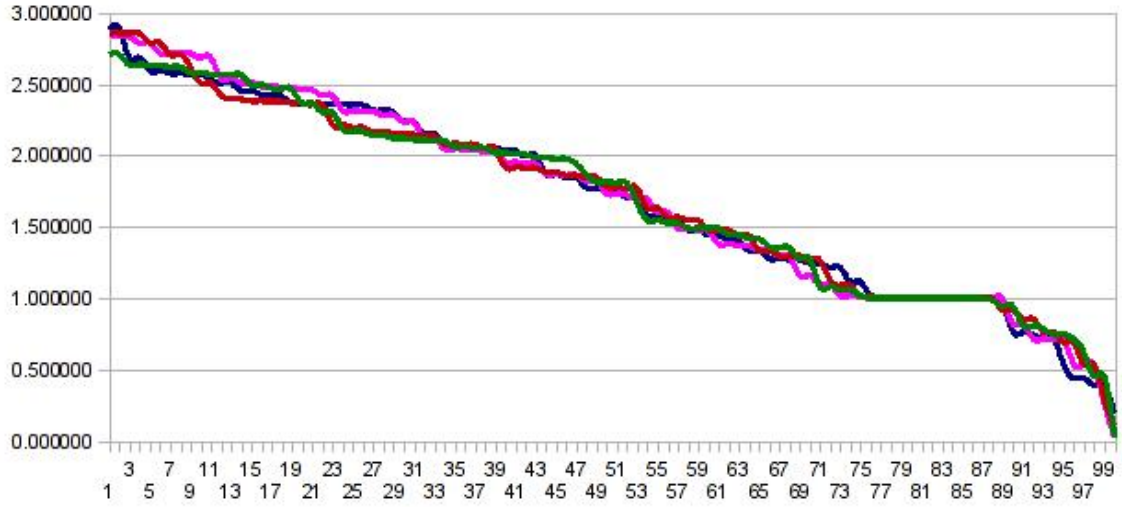


Figure 7: Magnitude of 100 eigenvalues

If parameters α and γ were free variables that we had to tune for each individual network then things would be very difficult for learning. This is not the case though. Given that all the important eigenvalues have a value of 1, all we have to do is to fix the two parameters to some values that have a sum greater than 1 (e.g. 0.2 and 0.9). By that, we guarantee that when the input to the network is close to one of its eigenvectors the vector state will gradually grow towards the direction of the eigenvector which leads to the corner of the hypercude that corresponds to the respective learned stimulus.

The role of the clipping function is more obvious now. If we didn't have such a limiting function there would be no bounds to the hyperspace and thus no corners for a network to converge to. The vector states would grow indefinitely and it would be extremely difficult, if possible, to know when to stop. In other words, the clipping function allows us to create a bounded hypercube and hence,

it provides for a means to discretize the visual stimuli and assign them to specific corners of a hyperdimensional space where dynamics is possible.

Regarding the dynamics of the full network of networks as a hetero-associative aggregate of the attractor networks, this is governed by the following equation

$$\boldsymbol{x_i}(t+1)=f\left(R_i(t)+\sum_j^N D_{j,i}(t)\right) \tag{6}$$

where the two components, *R* and *D,* are the recurrent (auto-associative) and dendritic (hetero-associative) activities of the constituent networks respectively

$$R_i(t)=f\left(\alpha A_i \boldsymbol{x_i}(t)+\gamma \boldsymbol{x_i}(t)+\delta \boldsymbol{x_i}(0)\right) \tag{7}$$

$$D_{j,i}(t)=f\left(\beta H_{j,i} \boldsymbol{x_j}(t)+\gamma \boldsymbol{x_i}(t)+\delta \boldsymbol{x_i}(0)\right) \tag{8}$$

N is the total number of networks (indices *j*) that each constituent network (index *i*) connects to. For the shake of simplicity the degree of outgoing connectivity is the same for all constituent networks. As for the incoming connectivity, in the case of fully- or statically-connected neighborhoods it has apparently the same degree with outgoing, but in the case of stochastic connectivity it exhibits some variance.

The dynamics of the full system has two parts. The first part, Eq.7, is identical with the auto-association dynamics of a single attractor network (Eq.4). The second part (Eq.8) is the contribution of the hetero-associativity and follows the exact same dynamics except that this time instead of the recurrent activity we integrate the denritic activity originating form the distal columns. Parameter β is

similar to α and it usually has the same scalar value. Matrices $\boldsymbol{H}_{j,i}$ correspond to the hetero-associative connections from networks i to j and are trained beforehand with the learning rule we've seen in a previous section.

Regarding the time evolution of the whole dynamical system, the iteration stops either when all constituent networks converge to some attractor or the states stop changing.

3.5.1. Alternative forms of dynamics

Equation 6 is not necessarily the best possible rule rather a particular formulation of a more abstract concept. One can think many alternatives that encompass the same idea of integrating dendritic activities from several linked systems. However, the actual behavior of the dynamical system will be highly dependent on the specifics of the evolution rule that is chosen. Even the minor detail or the simplest variation of the rule can cause a radically different evolution in time and the end result may have nothing in common with any of the alternative approaches. Before concluding to this particular formulation we experimented with other rules that produced mixed results, nevertheless they paved the way for fine-tuning the general idea and discovering the proposed rule.

The simplest, perhaps, approach to the binding of multiple synchronous attractor networks is the application of a scalar function *k(x,y)* that integrates uniformly the intermediate activations (states) of all the dendritic connections (Dimitriadis and Anderson 2006)

$$x_i(t+1) = f\left\{\alpha A x_i(t) + \gamma x_i(t) + \delta x_i(0) + \sum_j^N [k(x_i, x_j) x_j(t)]\right\} \quad (9)$$

Although this approach was capable of explaining some pop-out phenomena in visual attention it wasn't powerful enough to model more complicated processes in visual cognition. It was obvious that we would need a more versatile function in order to account for the multifarious synaptic interactions between neuronal groups. This led to the substitution of the scalar functions with full-fledged hetero-associations ***H*** that are formed with Hebbian learning inspired by the cortical neurophysiology (Dimitriadis 2008):

$$x_i(t+1)=f\left\{\alpha A\, x_i(t)+\gamma\, x_i(t)+\delta\, x_i(0)+\sum_j^N \beta\, \boldsymbol{H}_{i,j}\boldsymbol{x}_j(t)\right\} \tag{10}$$

Although this formulation was much more powerful and had the potential to handle a diversity of situations it proved extremely hard, almost impossible, to control. The reason is that the two main components of the summation, the auto-association ***A*** and the hetero-associations ***H***, are generating signals that belong to different spaces. Integrating tens or hundreds of these signals without first aligning them is in vain. So the obvious and simplest solution seemed to be the introduction of a normalizing function *g* that would project the vector signals on a common scale (Dimitriadis 2009):

$$x_i(t+1)=f\left\{\alpha A\, x_i(t)+\gamma\, x_i(t)+\delta\, x_i(0)+\sum_j^N \beta\, g\left(\boldsymbol{H}_{i,j}\boldsymbol{x}_j(t)\right)\right\} \tag{11}$$

This was indeed a great improvement and the results it produced made much more sense. However, the normalization of the signals was a memoryless process and was generating vectors with a temporal instability. As a result, the dynamics of the system was heading to the right direction for the first few iterations but was trapped prematurely in a state that couldn't improve further regardless of the time we allowed the dynamical system to evolve.

Equation 6 has overcome all these issues and as we will show later in this thesis it seems to work exactly as the proposed idea of neuronal binding at the abstract level dictates. Although it's impossible to exclude further improvements this equation is quite simple yet powerful. One could also argue that it's more intuitive

than Equations 9-11 but, no surprise, this only becomes apparent when one overcomes a problem. Most important, though, is the fact that all these equations look so much alike yet they produce behaviors that have almost nothing in common. This shows how crucial is for cognitive science to have computational models with detailed descriptions and actual implementations.

3.6. Discussion

3.6.1. Class of models

The proposed approach is more of a computational framework rather than just a model. By framework we mean a class of models with the same principles but with multiple possible realizations that can account for different situations (Anderson et al. 2007). This versatility of the proposed framework is concealed behind the various parameters and architectural decisions we described earlier in this section. We have to note, though, that none of these parameters is the subject of learning. We're used to think of variables as a burden for learning – the more variables a model has the harder it is to train it. This is not the case in the proposed framework though. We're not talking about variables but parameters. Parameters help us customize the framework towards the needs of the problem we have to model. For example, the size of the receptive field and the overlap between adjacent RFs are two of these parameters. The system can work equally well with a variety of values for these two parameters. However, by adjusting them appropriately we have the ability to control the resolution and the granularity of the visual perception. This can be quite useful if, for instance, the training images are prohibitively large.

Two major parameters that can differentiate the possible realizations of the framework is the number of hetero-associations and their distribution over the processing sheet. In the experiments to follow we will experiment with both of these parameters and we will see that they play a significant role in the system's ability to model behavioral data. Of course, there are more parameters we could experiment with (e.g. number of layers, receptive field properties, learning properties, etc.) but we're not going to go into this direction here.

Versatility is a ubiquitous property of the brain. There are many cortical networks that are quite similar but perform quite dissimilar processes. It seems that what we should be looking for is a generic architecture that has the ability to adjust easily to different modalities and tasks. We believe that this framework is heading to this direction. The next goal is to see what kind of customizations we need in order for this associative architecture to process stimuli from other modalities or perform some cross-modal integration.

3.6.2. A network of networks is not simply a large network

To anticipate a common misconception, we have to make a clear distinction between a large neural network and a network of networks (Anderson and Sutton 1995). In a traditional neural network approach the dimensionality, and therefore the variability, of the network increases as a function of its input (see Hinton et al. 2006). If we think of the visual input as an array of the size of the retina the complexity of a large neural network that will deal with this kind of variability is theoretically and practically unbounded. Contrariwise, in a network of networks approach the dimensionality is not any more a curse. The increase of the input

doesn't entail an exponential growth of the system. At the level of an attractor network the variability can only scale up to the size of a receptive field, while at the macro-level the variability depends on the number of possible pairwise hetero-associations which, to a great extent, is independent of the size of the visual input. There are two reasons for this. First, because the attractor networks do not have to maintain a full connectivity in order to interact with all possible locations of the processing sheet. The sparseness of the hetero-associations combined with the dynamics of processing allow to the system to grow linearly, and not exponentially, to the input. The second reason is that the statistics of the visual world only favors certain associations between parts of the stimuli. For example, an attractor that corresponds to a T-junction is very likely to be strongly hetero-associated with a neighboring attractor that corresponds to a straight line and not an L-junction or another T-junction. Hence, contrary to what a large neural network would do, a network of networks can capture most of the visual statistics and can generate the necessary associative power by only preserving a dense connectivity at the local level and a sparse one at the global level. By that, a network of networks serves also an answer to the criticism on traditional neural networks regarding their lack of structure. In the proposed hetero-associative network the structure is intrinsic in the connectivity and the communication of the constituent attractor networks.

3.6.3. Why do we need a dynamical systems approach

Given that the individual attractor networks act like local pattern recognizers, a question that arises very easily is why do we need a dynamical systems approach? Why can't we simply replace the local attractor networks with some

other mechanism of pattern recognition like a neural network, a support vector machine, a Bayesian classifier, etc.? Although this would be a technically valid substitution, the use of symbolic (or sub-symbolic) components instead of ones with temporal dynamics would eliminate the inherent advantages of the proposed approach. This is because the interaction of the attractor networks (i.e. the hetero-associations) do not simply occur *after* the completion of the attractor's operation (i.e. the convergence to a pattern) but *during* its temporal processing. It's not the result of the attractor network's convergence that is hetero-associated with the results of its linked networks, but all the intermediate states that occurred in the course of this convergence. Attractor networks function iteratively and this allows them to constantly interact by incorporating each other's ongoing state into their own course of time. So, the hetero-associated networks will continuously interact and influence each other before reaching their final state that corresponds to the recognized pattern. This kind of processing and interaction cannot be achieved with a non-dynamical approach.

Besides, if we were able to substitute the whole dynamical process with the mere output of some instantaneous pattern recognition then we should also have been able to substitute patterns of distributed activity with unique symbols and labels. One can easily realize that this kind of regression will eventually call for a pure symbolic system with no need for temporal dimension and distributed information. And brain is definitely not this kind of system!

3.6.4. Perception without action

The convergence of a dynamical system generally depends on the existence of fixed points in the state space where the system will settle. However, in a brain-like dynamical system new input would constantly be arriving and attractor networks would never have enough time to fully converge to a fixed point (Spivey and Dale 2006, Spivey 2007). So activity would never stop. As state trajectories would get closer to an attractor's basin, enough *evidence* would be provided for an *action* to be taken. And this action would generate a new (motor) behavior that would provide new percepts to the system which would cause a change in the unfinished trajectory and a new processing flow that would lead to a new attractor. However, our framework is only about perception and it doesn't model but a tiny part of the actual human behavior. Since we don't couple it with a motor behavior, or any other kind of cognitive *action*, we cannot expect but the system to iterate and converge to a fixed point in state space. This is kind of unrealistic from a biological point of view but it's a necessary simplification that serves the purpose of our experimentation.

4. Qualitative comparison with other approaches

Although there is no single point of agreement among the various models of visual recognition, most of them share some characteristics with each other. This is true for our approach as well. Most important, though, are the differences, even the minor ones, which combined with the rest of a model's properties have sometimes the potential to convey a uniqueness. In this section we will examine these distinctive characteristics that differentiate our approach from the most known models we reviewed earlier. We will focus on four aspects of the proposed framework that summarize the essence of the comparison. Specifically, the role of the architecture towards the specificity/variance problem, the constraints in processing imposed by the dynamics of visual perception, the types of learning that are necessary for recognition, and the limitations of serial processing to account for biologically plausible computations. The comparison in this section will only be qualitative.

4.1. Specificity and variance

One of the major issues that any recognition model has to resolve is the balance between specificity and variance. A holistic approach to the horizontal architecture will have a good specificity but a small variance while a very modular horizontal architecture will have a high variance with a small specificity. So there is a need to keep the right balance and combine in some way an appropriate degree of local specificity with the desired global variance. Most models tackle this problem with the use of a multi-layer vertical architecture where each layer integrates information of the layer beneath. The idea is that the lowest layer will

only have to face the specificity-variance problem within a local receptive field (Geman and Bienenstock 1992) and as we ascend the hierarchy we gradually increase the variance by pooling together information of the layers beneath. The problem with this approach is that in order for two distant points in an image to be associated and perceptually fused their receptive fields would have to ascend a certain number of layers before they meet; and this number is proportional to their distance in the visual field. This means that in order to face the invariance-specificity problem and perform a recognition with these models we would have to optimize the number of layers according to the particular class of objects under consideration. Moreover, we would have to appropriately select and optimize the pooling mechanisms. Apparently, this is very restrictive and counterintuitive because, for an architecture as generic as the visual system seems to be, no fixed number of layers or pooling mechanisms would be able to account for the recognition of every possible class of objects. So, there must be some powerful mechanism of fusing the visual information that works within each layer and in parallel to whatever integration occurs in the ascending pathway. Without such a mechanism, that extends the local scope of the receptive field and its projected locus, the information fusion can only be limited and the recognition process hopeless.

This is where, according to our opinion, most of the models are lacking an essential property that any visual recognition architecture should have. They put all their focus on the projections from layer to layer and neglect the more crucial processing that takes place within the layers. They focus on the visual features (e.g. size, shape, color, etc.) and pay less or no attention to the binding of this

information. From our perspective, the big question is not as much how to model the locality of the visual processing (e.g. with neural networks, pattern templates, cellular automata, etc.) or what kind of building blocks to use (e.g. receptive fields, fragments, segments, etc.), but what kind of communication and fusion can we employ to bind all this information together. Towards this direction, the model by Wang 2001 is using lateral connections that excite specific neurons within each layer. However, these connections are only excitatory, their pairing is hardwired, and their strength is not tuned by learning. A very similar approach has been tested before by Dimitriadis and Anderson 2006 and it showed that it has a limited application and it's quite hard to control. By contrast, the current framework provides the flexibility that a system needs in order to combine both local and global processing, both within and across layers. The proposed horizontal architecture is not only modular with respect to the fragmentation of the stimulus array, as suggested by the majority of the models in the literature, but is also modular with respect to the processing that occurs at each of these layers. The information fusion is not the mere result of a local pooling mechanism that occurs at the interface of two layers. It is the result of a repetitive interaction that allows any two parts of the stimulus array to associate themselves before the fused percept advances to the next level in the hierarchy. By doing so, the role of what other models call *layers* is partly incorporated into the hetero-associative processing. Therefore, the number of layers is dissociated from the process of fusion itself and is mostly related to the levels of complexity that the visual representation may have.

In our approach, the auto-association in the constituent networks of any layer works similar to what other models do with recognition or pooling mechanisms in local receptive fields. But in the global image-wide level, these networks learn to hetero-associate their internal states in order to form assemblies and fuse their information without relying to ad hoc number of layers as other models do. The local specificity through the recurrent processing takes into continuous account the global variance dictated by the interconnectivity of the constituent networks that evolve and take shape as a result of the statistics of the environment. With this hybrid horizontal processing the system is able to retain its specificity while tolerating some degree of variance.

4.2. Time

There is a lot of evidence in the literature that visual recognition is not immediate but has various time dependencies. For example, in case of scenes with complex backgrounds and multiple occluded objects that require visual attention (Treisman and Gelade 1980, Wolfe and Horowitz 2004), or in the case of rotated objects (Shepard and Metzler 1971), or inverted faces (Valentine 1988), recognition is dependent among other things on the time the visual system is allowed to process the stimuli. The fact that the system's response time is variable has direct implications on the architecture of the visual recognition. Previously, we've seen that in order to account for the specificity/variance problem we would have to optimize the number of layers for every different class of objects. Now we see that even with an optimal number of layers we would still be unable to account for differences in behavioral response times. This is

because the processing time in a purely feed-forward system would be fixed regardless of the stimulus configuration. For instance, the HMAX model (Riesenhuber and Poggio 2000) which involves a series of Gaussian tunings and nonlinear max operators would always have the same response time regardless of the content of the stimulus. In this case it only depends on parameters like the number of layers and the density of the receptive fields but not on the stimulus itself. And although this may be a highly desired property for a computer vision system, for the purpose of behavioral modeling it doesn't tell much about the temporal aspect of processing.

In order to account for the time dependency in the visual recognition we need an architecture that incorporates delays. The simplest way to do that is through feedback loops because the one thing we know for sure is that there are plenty of them in the visual system (Felleman and Van Essen 1991, Callaway 2004, Douglas and Martin 2004). A massive number of back-projections in the cortex convey information from higher areas back to the lower ones, while feed-forward connections constitute only a small fraction of the total connectivity. The significance of feedback loops has been shown not only for the cortico-cortical connections (Mumford 1992) but also for the thalamo-cortical ones (Mumford 1991). As high as 80 percent of input to the LGN is coming from the primary visual cortex which is also the major synaptic target of the LGN (Bear et al. 2001). In a sense, lower and higher areas are in a continuous communication until they agree upon the pattern to be recognized. As Mumford nicely put it, “the whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top-down, bottom-up loops”. This

synchronous activity is the critical component of recognition with feedback providing the opportunity to use top-down knowledge and task dependent expectations.

In spite of all this evidence most computational models we reviewed earlier focus on a constant time feed-forward processing and neglect or underestimate the crucial role of feedback connectivity. This is a major difference for the proposed framework that focuses exactly on the simultaneity that Mumford was describing. Contrary to what serial processing suggests, in a dynamical system we can allocate time as needed and we can expect different results depending on the time constraints we impose. So, behavioral time response is inherently built into the system while the temporal dimension is not any more the mere function of the number of layers or stages of visual processing but the product of a dynamical process. Moreover, this dynamical process provides the gateway not only to link the various visual modules together but also to associate them to other cognitive functions. In a pure feed-forward approach it would be extremely difficult, if not impossible, to bind information from different sources without explicitly resolving the numerous time dependencies. Contrariwise, in a dynamical systems approach this is not an issue because the system will have its time to interact and align its modules even if the initial interface is not the right one.

We do not claim, of course, that the recurrent processing of the attractor networks can account for the whole complexity of the feedback loops in the neuronal circuitry. Even if we accept the abstract idea that each of these attractor networks is mimicking the role of a minicolumn, there is still a lot of complexity

that a single network couldn't explain. But it is definitely a starting point. It is preferable to make a simplified first attempt to account for the dynamics of processing than neglect it at all. The only thing for sure if we use a feed-forward processing and neglect the top down effects and the feedback loops is that we will never explain anything but the simplest cases. So, this shouldn't be an option for any model of visual recognition. We better start with a dynamical approach – even the simplest one – than take a dead end path that will never explain the temporal aspects of visual recognition.

4.3. Learning

As we've seen in the literature review earlier there is no general agreement among the various models regarding the principles of learning in visual recognition. Since the details of visual learning are not fully understood, most models seem to pay less or no attention on this aspect of modeling. So they employ a learning approach that is mainly dictated by their architecture and their processing scheme rather than by the neural basis of learning. Contrariwise, our approach not only puts a great emphasis on learning but proposes a scheme where learning has a crucial central role.

In the proposed framework, learning is not simply a component of a modular system but a widespread property upon which everything else is built. Contrariwise, most models see learning as a technical necessity rather than a biological property. For instance, the HMAX model (Riesenhuber and Poggio 1999) is only using learning in alternating layers of the feature-extraction hierarchy while the rest of the layer interfaces remain hardwired. Moreover, in the

top layer, learning has to take a whole different form (a supervised classification) in order to solve part of the problem (Jiang et al. 2006). A different mechanism with a similar concept is followed by Hinton et al. 2006. Learning here is trying to glue together the monolithic layers of the feature-extraction hierarchy while the top layer is performing a categorization task with a learning mechanism that is different from the one in the layers beneath. An engineering approach is used in Ranzato et al. 2007 where learning (supervised gradient descent) is merely an encoder/decoder in a pipeline processing. But even in the case of more biologically plausible architectures like the model by Deco and Rolls 2004, learning (in this case Hebbian) is only applied to the feed-forward connections while feedback and lateral connections are based on preset, not learned, criteria like reciprocity and distance.

In all these approaches learning is a necessary measure and not a fundamental widespread principle. But this is not how learning is supposed to be. Learning is tuning the whole system from bottom to top and from one side to the other. Learning is supposed to modify all synapses in all the layers with no exclusions. This doesn't mean it has to be the exact same process across the system but it has to follow the same underlying principles of neural plasticity. And this is exactly the approach we take. We propose an associative learning mechanism that works under the same principle, simultaneously and for all scales, both local and distal, both within and across layers. There is no segmentation or discontinuity. There is no need to employ different mechanisms for different neuronal groups or layers. There are no synapses left behind or set by hand. We

let the statistics of the environment shape the network and make the strongest most significant connections emerge and prevail.

4.4. Parallel processing

Although the computer metaphor had a profound effect in initiating the field of cognitive science we now realize that the brain performs a very different processing than the serial one computers do. When neurons seem to be inactive it is not because they're waiting for some program counter to instantiate them but because they receive a weak input. In a distributed system like that, where constantly all the units work in parallel, it is hard to imagine how the serial routines proposed by the computing paradigm can operate. Many models of visual recognition, despite their seemingly parallel architecture, have essentially a serial processing. The proposed framework, on the other hand, has a pure parallel approach that distributes time and processing evenly among the units. To see this difference take for instance the fragment-based model (Ullman and Sali 2000, Ullman 2006). In order to select the optimal shape and size of the fragments, we have to make an exhaustive (brute-force) search of the full joint distribution of shape, size and object class. This may improve considerably the recognition performance but is not feasible without a central executive that coordinates not only the serial search but the storage and retrieval of the fragments as well. But even in the case of the more biologically plausible hierarchical neural networks (Fukushima 1988, Wallis and Rolls 1997, Riesenhuber and Poggio 1999, Jiang et al. 2006), the processing is performed in serial steps where consecutive layers operate with a temporal order. If this was

the case for neural computations then we would have to explain what's the mechanism controlling the brain's clock, and what the layers do when, for most of the time, it is not their turn to operate. Within a serial processing scheme both of these questions are very hard to answer. But within a parallel processing scheme these questions do not even exist. The massive connectivity proposed in our framework eliminates the need for biologically unrealistic serial computations. And although this connectivity would probably never achieve a performance as optimal as a brute-force controlled approach suggests, it will, nevertheless, conclude to a solution that better balances between efficiency and performance.

4.5. Contribution to other approaches

So far we focused our analysis on the differences between the proposed approach and the other models and we didn't discuss any similarities or common ground that may exist. For instance, there are certain aspects of this framework that under a different interpretation could be compatible with what other models suggest. One such example is the relation between our approach and Ullman's fragment-based model (Ullman and Sali 2000, Ullman et al. 2002). The basic principle in that model is to maximize the mutual information between fragments and object classes in order to select the most informative class-fragments. From a computational point of view this is a greedy search that explores all possible pairwise combinations between fragments and object classes. From a biological point of view, though, it is implausible not only because it is using a supervised learning mechanism but also due to the sequential nature of the search. If there was a biologically feasible mechanism that provided the means for an information

theoretic interpretation of this search, it would help us bridge the two levels of abstraction. In the proposed framework we do suggest such a mechanism that approximates the maximization of the mutual information and avoids both the supervised learning and the exhaustive search. The hetero-associative learning among the modules is an unsupervised process that due to its limited capacity, like every learning mechanism, will end up picking up the associative links with the highest mutual dependence. After a prolonged period of unsupervised training on a set of stimuli that exceeds the capacity of the synaptic matrices, the hetero-associative connections will only hold the most informative associations among the receptive fields (see Experiment 4). The fragments, of course, encapsulate more information than the receptive fields, but in a hierarchical system we can think of the fragments as the result of assemblies formed by mutually dependent receptive fields at various scales. As for the indirect search performed by our framework, it may not cover the whole visual field as a brute-force approach would do, but given a relatively rich and strategically distributed connectivity of attractor networks this approach could be a sufficient and plausible implementation of Ullman's greedy algorithm.

Another model that could benefit form the proposed approach is the Deep Belief Network (Hinton et al. 2006). Although the two architectures may seem quite incompatible they both employ an associative memory that can be used as the basis for sharing some features. The Deep Belief Network is a multilayer model with bottom-up and top-down links but has no connections within the layers. This makes it a monolithic structure in the horizontal dimension and prevents it from taking full advantage of all possible associations among its units. On the other

hand, the power of our approach lies exactly in the heart of associativity. If the Deep Belief Network assumed a more modular architecture in the horizontal dimension it could exploit the use of lateral interactions in order to improve its performance and scale easier to larger inputs. Moreover, it would reduce considerably the learning time since the individual modules would have a much smaller dataset to learn in parallel. This is not a suggestion that pertains to the Deep Belief Network only. Any kind of monolithic network that has the structural potential to break in smaller parts or modules could benefit by the lateral connectivity we propose in this thesis.

It is often the case that seemingly different approaches share a common abstract idea but they choose to implement it in a different way or level of detail. In this case, one can complement the other and provide a new insight or simply an alternative implementation. It is our belief that the proposed approach provides a coherent framework that could make many contributions to other models, particularly in providing neuronal level explanations for many abstract concepts that most models share.

5. Experimental results

5.1. Methodology

The visual processing proposed in this thesis involves a large network of dynamical systems with a very large number of associative links among them that create a complex distributed system which exhibits some abstract similarities with the complexity of the brain circuitry. The constraints of the implementation of such a system are undoubtedly bounded by the availability of computational resources. To give a rough idea of the computational resources that are necessary to construct the proposed framework we will give a small example of visual processing. Let us assume a square visual patch of 100x100 points which is roughly the size of the fovea, and receptive fields that, for the shake of computational simplicity, have a square size of 10x10. For achieving the high acuity of the fovea we should also assume the largest possible overlap between adjacent receptive fields, i.e. RFs that are positioned 1-point apart from each other. An instantiation of this system with our framework will have 10,000 auto-associative attractor networks with each one having 100 neurons and 10,000 recurrent connections. Furthermore, if we make the conservative assumption that each network is hetero-associated with 10 other networks only, this will give a system with 1 million neurons and 1.1 billion synapses. If we want to be more realistic about the cortical organization and the number of neuronal connections then the total number of synapses could be 100 times higher than that (Braitenberg and Schuz 1998). In terms of computational resources these figures translate to tens of GFLOPS of processing power and tens of GBs of working

memory. Obviously, this is not something that a personal computer can currently handle efficiently.

In order to face this immense computational demand we use a high performance computing cluster provided by the Center for Computation and Visualization of Brown University. The nucleus of the system is an IBM 166-node Linux cluster with each node having two Intel Xeon 5540 (2.53 GHz) quad-core Nehalem processors and 24 Gigabytes of DDR-3 memory (1333 GHz). A 40 Gb/s Quad-Data-Rate (QDR) infiniband messaging interface provides the communication among the 1328 processing units and the 3984 Gigabytes of the total distributed memory. We program the proposed framework in C++ and we use the Message Passing Interface (MPI) for the communication among the nodes. Our approach to concurrency is based on data parallelism and the neurobiology of vision. The visual stimulus is split into many regions and processed in parallel by different CPUs with a single-instruction multiple-data (SIMD) paradigm. Each processing unit undertakes a certain number of cortical columns and executes the same code on different data following the processing paradigm and the topographic organization of the visual cortex. The signal transmission among the cortical columns that are processed by the same CPU is mediated by the use of the node's local memory, while the distal axonal projections are using the infiniband interface and the MPI (Dimitriadis 2008).

Last, regarding practical considerations and in order to limit the use of computational resources to the least possible, our experiments will be based on stimuli with a small size ranging from 100x100 to 200x200 points. A typical

number of processing units for these experiments will be 64 cores. As for the running time of an experiment it can vary widely, usually in the range from 1 to 60 minutes, depending mainly on the dimensions of the stimuli, the size of the training set, and the number of hetero-associations among the columns.

5.2. Stimuli

The proposed framework will be trained and tested mainly on natural images with uncluttered backgrounds. This approach provides a realistic trade-off between the high complexity of fully natural scenes and the low plausibility of artificial stimuli. In the current version, as we explained earlier, we're experimenting with a single-layer architecture that encodes one visual feature only. The visual map we choose for our experiments is the result of an edge detection filter that resembles the function of the simple cells in the visual cortex (Hubel and Wiesel 1974). Figure 8 depicts the stages of pre-processing we apply to a set of stimuli before we feed them to our system. In the left column we see the original images. The middle column is the result of the first stage of processing after discarding the color information and empirically scaling the images to an approximately similar size (100x100 pixels each). In the second stage of pre-processing (third column) we encode the stimulus into a binary vector that has a suitable format to be processed by the attractor networks. The filtering is based on the Canny edge detection algorithm (Canny 1986) and is using the same fixed procedure and parameters for all the stimuli and all the experiments (no optimization applied). In most of the experiments we will be using this set of 11 tables which is randomly selected from publicly available images on the Internet.

Figure 8: Input stimuli and pre-processing

5.3. Visualization

Scientific visualization plays a critical role in any kind of computation. This is more so in our case where the dimensionality of the whole system far exceeds the dimensionality of the input stimuli. The processing of a typical stimulus of 100x100 points will involve millions of variables changing over time. If we assume a receptive field with a diameter of 7 points and a maximum overlap with adjacent RFs, we will have 10,000 attractor networks with a *49*-dimensional state each. Is it necessary to visualize all this information? Not really. We just need to deploy some method that maps the state of each dynamical system to a single pixel. Apparently, in a mapping like this, a lot of information will be discarded. However, with the right methodology it is possible to provide an intuitive and informative picture of the internal states of all the attractor networks (and hence, the system's overall state) without undermining our ability to interpret the results of visualization. This is because the information we mostly care about is not the content of the whole state (i.e. the whole receptive field) but the central location only. The overlap between RFs provides a lot of useful information for the processing but is not necessary for the visualization. Therefore, we can discard it and visualize the content of the central components of the state only. In case of a maximum overlap between adjacent RFs the remaining central location will be a single point. Otherwise, we will have to pool together (e.g. averaging) the values of two or more central components of the state vector. In either case, the extracted value will range in [-1,+1] and a direct mapping to a grayscale value for visualization purposes is straightforward.

In practice, the grayscale value of the visualized pixel will depict the performance of the respective attractor network. A network performing well on the testing stimulus will converge to one of the binary attractors so the visualized pixel will have a value of either 0 (black) or 255 (white). A network that hasn't yet converged to a corner of the hypercube will generate a visualization pixel that ranges somewhere between 0 and 255. The closer this gray value is to 0 or 255 the closer the respective attractor network is to convergence. Of course, there is also a probability that a network may reach a corner that doesn't correspond to any attractor. In this case the visualized pixel will have a random value of either 0 or 255. However, this won't be hard to detect in the visualization because random pixels pop easily out of the image. In the experiments to follow we will see that this situation is very rare.

For each time step of the dynamics the visualized pixels from all attractor networks are put together in a mosaic and presented as an image. The more time we allow the system to converge, the closer this grayscale image should get to a binary image (all the attractor visualizations be either black or white). Ideally, when the criteria for convergence have been met, the final image should be identical to one of the binary training stimuli. In practice, though, there are always some discrepancies so we will be comparing the final image with all the training stimuli and provide the pairwise correlations in order to examine how well the system performed.

Last, one thing we have to note here is that, although the visualized output will have the form of an actual image, this doesn't imply by any means that the

internal representations are isomorphic to the output. There is no *image* into the system. There only exist states in a hyper-dimensional space. The proposed framework doesn't use any kind of labels or symbolic representations in order to process the visual information. The same holds for the outputs, either intermediary or *final* ones. Although they will definitely look like images to us, the system itself does not have a holistic percept of these images. It doesn't assign any kind of label either. This is very different from most traditional image processing methodologies (statistics, neural networks, AI, etc.) that assume that computation has to rely on some form of symbolic representation.

5.4. Study I – Sensitivity and specificity

The most basic challenge that any visual recognition system faces is noise and variability. In order for a system to be successful it must have the ability to tolerate variability while preserving its specificity. This is not an easy task by any means. Because the more variable the system becomes the less able to retain its specificity it gets. And the opposite. The more biased to certain stimuli it becomes the less able to generalize it will be. Moreover, it is often the case, especially in visual cognition, that one cannot discern between noise and signal. What is noise for a certain situation it might be a signal for a different one. Noise and signal are often context dependent and this poses a great problem for visual learning.

In this study we are testing the system's sensitivity and specificity to inputs with variability and noise. In the first experiment we test the tolerance to various degrees of random white noise. In the second experiment we go a step further and test the system's ability to reconstruct ablations of large contiguous image regions. In the third experiment we test the specificity of the system with regard to previously unseen stimuli.

Ideally, the system should exhibit a strong robustness to stimuli with random noise and a somewhat diminished strength for the case of ablated image regions. This is because in the case of random noise there will still be some weak but randomly distributed signal that will drive the activity of the system via the associative links. However, in the case of ablated blocks, the only available signal will lie far apart so the system should have a greater difficulty recovering

the unknown information. Regarding specificity, although the training of the system is building a strong bias in the attractor networks this shouldn't be sufficient to drive an unseen input stimulus to one of the learned ones. The output in this case would probably consist of the overlap between the training and the testing stimuli.

5.4.1. Experiment 1

In this experiment we test the system's tolerance to random noise. The training set consists of the 11 stimuli presented in Figure 8. The actual input to the system is the result of the edge detection as depicted in the third column. The testing stimuli consist of noisy variations of the training stimuli where 70% of the image points have been randomly deleted (i.e. set to zero instead of ±1). For the training process we use the Widrow-Hoff learning algorithm with a stopping criterion being a vector similarity (cosine) of 0.9999 or 200 iterations, whichever comes first. Regarding the hetero-associative connections, all attractor networks are statically connected to their 8 adjacent neighbors. The training of these connections is also based on the Widrow-Hoff learning algorithm. As for the parameters of the dynamics, they are all fixed to the values α=0.9, γ=0.2, and δ=1. During the dynamics we let the whole system iterate for 50 steps and we visualize the output for each discrete time step. The processing time for 64 CPUs was 3min. Figure 9 presents the results for iterations 1 through 5, and 8. After that point no significant change takes place and most attractor networks have converged to a pretty stable state.

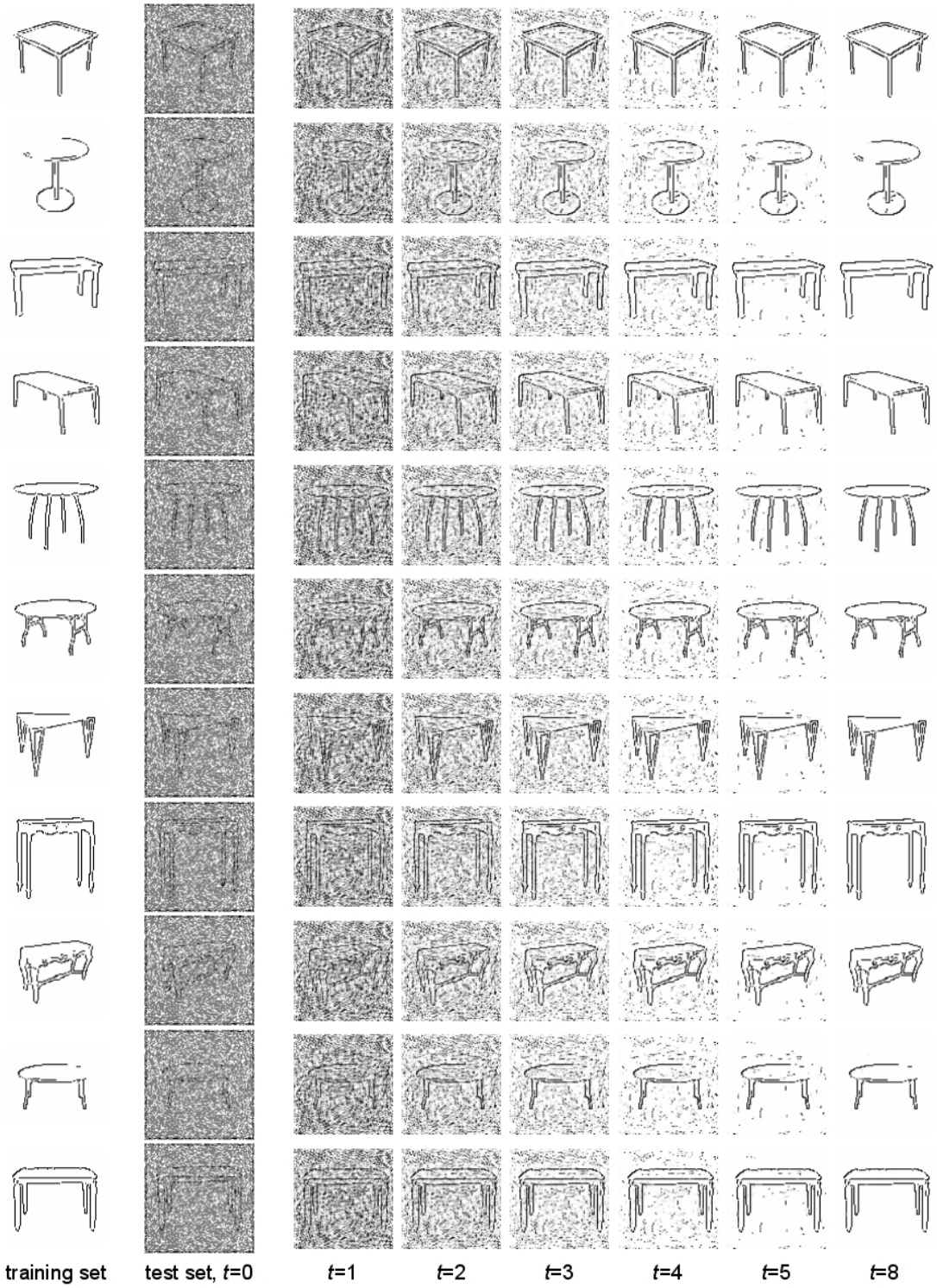
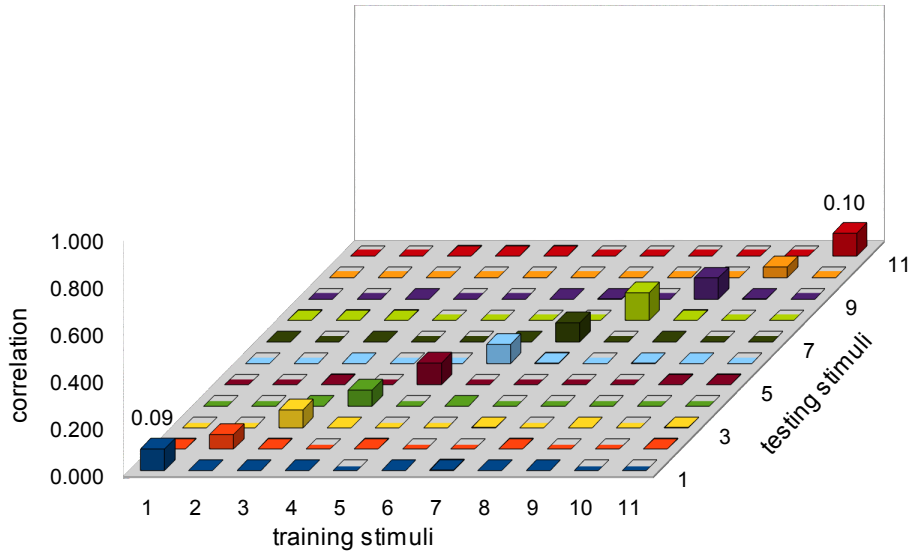
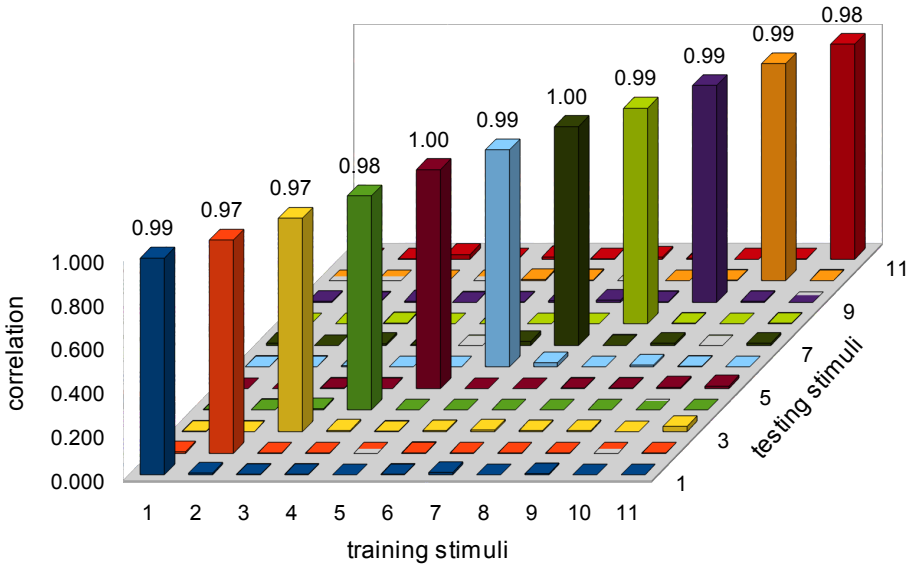


Figure 9: Experiment 1. Testing stimuli with 70% random ablations

Pixels that are depicted as white or black visualize network states that have reached a corner of the state hypercube and there is a high probability that this corner belongs to an attractor. Pixels depicted in gray haven't yet converged to a corner and their intensity visualizes their proximity to white and black.



Graph 1: Initial pairwise correlations



Graph 2: Final pairwise correlations

From Figure 9 it's obvious that the system can clear the noise and converge in just a few steps to a global state that resembles remarkably one of the learned stimuli. In quantitative terms, Graph 1 and Graph 2 present the correlations between all possible pairs of testing and training stimuli before and after the dynamics respectively. No testing stimulus has an initial correlation of more than 0.1 with any of the training stimuli. After the dynamics all outputs converge to the correct training stimulus with a correlation of more than 0.97, whereas the correlations with the non corresponding training stimuli do not exceed the value of 0.02 for any of them.

Although the level of noise (70%) is very high, one could claim that the testing stimuli retain adequate information to get recognized. Therefore, we make a second test with the level of noise increased to 90%. In this experimental setup nine out of ten pixels are deleted. The test set in Figure 10 is extremely hard, if possible at all, to recognize and the initial correlations (Graph 3) are all lower than 0.02. Even in this case, though, the dynamics converges to a state where the outputs for all testing stimuli have a high similarity to the corresponding training ones. Figure 10 depicts iterations 1 through 5, and 50. Of course, with this level of noise it takes longer for the system to reconstruct the stimuli but it manages to do that with the exact same parameters and the same number of hetero-associative connections as before. The final pairwise correlations (Graph 4) are not as high as in Graph 2 but they're still significant (μ=0.79, σ=0.09) while none of the correlations with the non corresponding stimuli exceeds a value of 0.02.

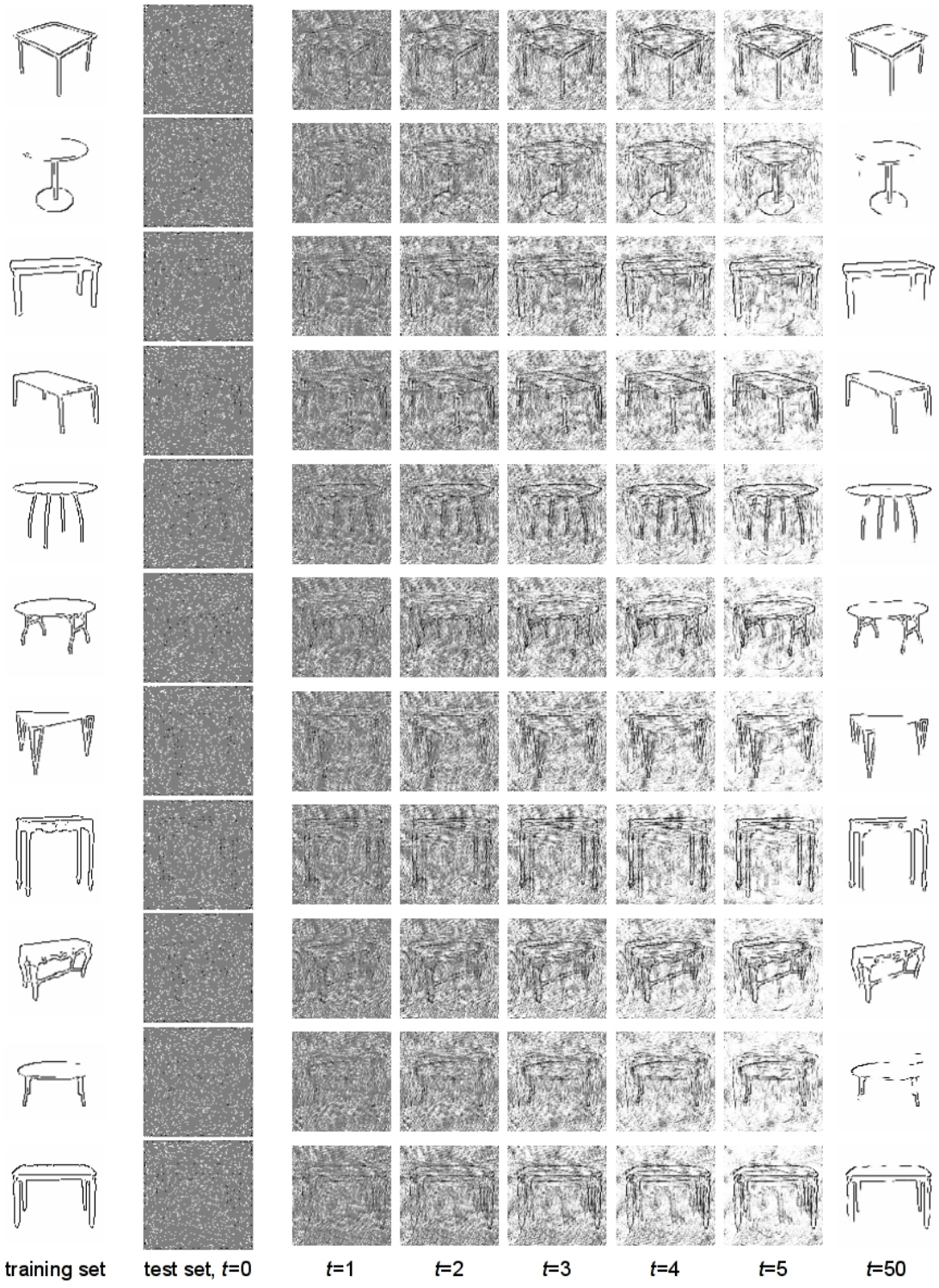
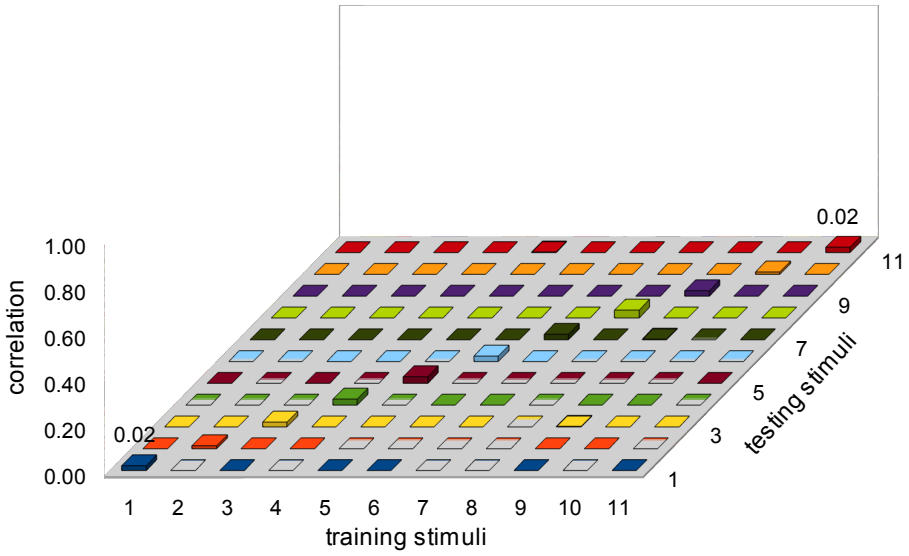
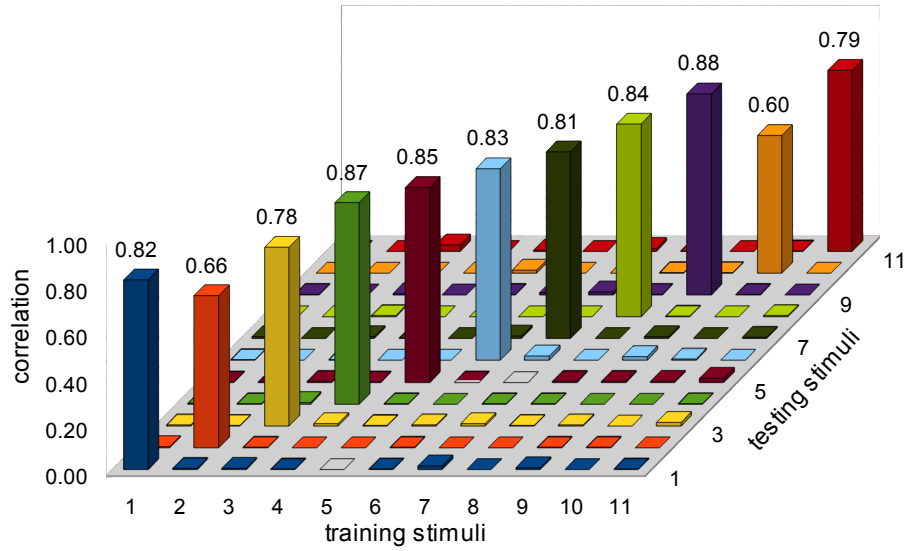


Figure 10: Experiment 1: Testing stimuli with 90% random ablations



Graph 3: Initial pairwise correlations



Graph 4: Final pairwise correlations

5.4.2. Experiment 2

In the second experiment we test a variation of the previous scenario. Instead of random ablations we now delete whole image blocks – i.e. contiguous image regions. But why is this important if, from what we've seen in the first experiment,

the system is tolerant to levels of noise even as high as 90%? In the case of randomly distributed noise the information that the signal retains will also be randomly distributed. So, despite the level of noise, there is some remnant *islands* of information that, given a sufficient degree of connectivity and time, the system can exploit in order to reconstruct the noisy parts. However, in the case of ablated image blocks, there will be no such *islands* of information to drive the system's activity. In this experiment, the only available information will lie outside the borders of the ablated regions. So, the question is whether the system will manage to overcome these relatively huge image gaps (30x30 pixels for an image of 100x100 pixels) and rebuild their contents from scratch by relying exclusively on associations with distant image regions.

A feasible solution to the problem would be the use of a large number of distant connections, but it's not a preferable one because its cost increases with the size of the ablated region – the larger the gap the larger the surroundings that we need to cover. Therefore, we test the system with a small number of hetero-associations, twenty, that have a Gaussian distribution. Practically, this means that each network is fully connected to the ring of the eight adjacent networks, plus 8 out of the 16 networks of the second ring, plus 4 out of 24 of the third ring. So, the longest distance an association can reach is two image points apart whereas the ablated region is 30 points wide. Apparently, in order for the system to reconstruct the ablated region it would have to let the networks propagate their information and gradually fill in the blanks with the information that the associations dictate. Visually, we would expect the ablated part to start shrinking and the reconstruction to occur inwards. And this is exactly what happens.

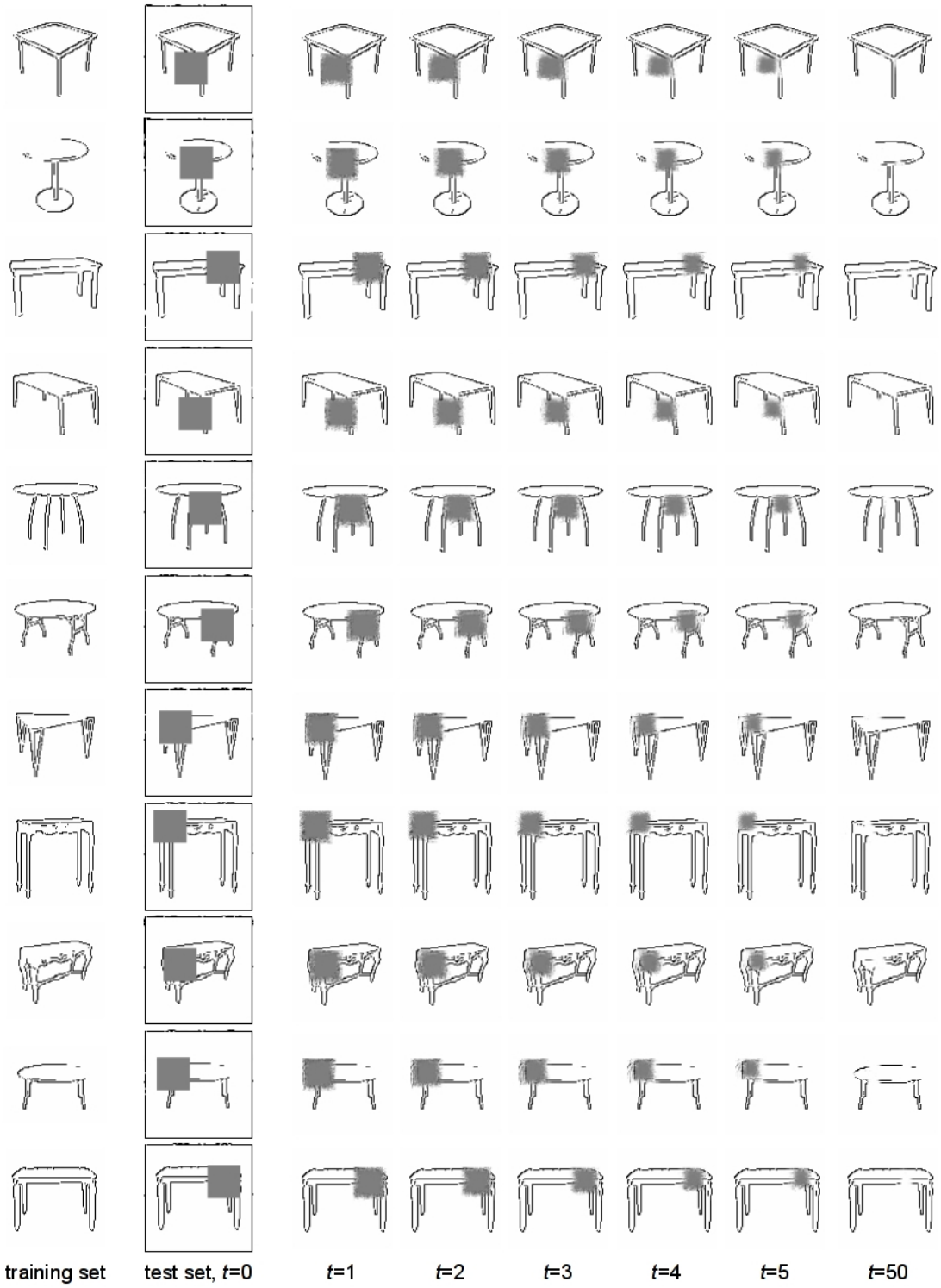


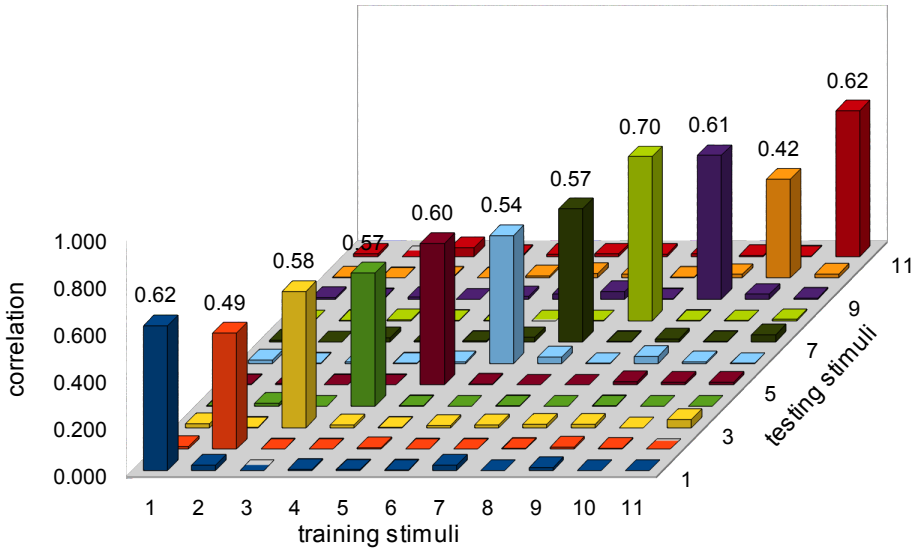
Figure 11: Experiment 2: Testing stimuli with ablated blocks

Figure 11 depicts the testing stimuli along with the output of the system in iterations 1 through 5, and 50. For the training process we used the Widrow-Hoff learning algorithm with a stopping criterion a vector similarity (cosine) of 0.9999 or 200 iterations, whichever comes first. The parameters of the attractor networks are all fixed to α=0.9, γ=0.2, and δ=1 and we let the whole system iterate for 50 time steps. The processing time for 64 CPUs was 7min. As predicted, the ablated regions are gradually shrinking and filled in with information driven by the learned associations. In this particular configuration the framework manages to extinguish the noise for most testing stimuli in about 10-30 time steps. Although the time of convergence varies among the testing stimuli all outputs at time step 50 have a high resemblance with the corresponding training stimuli (compare the first and last column in the figure).

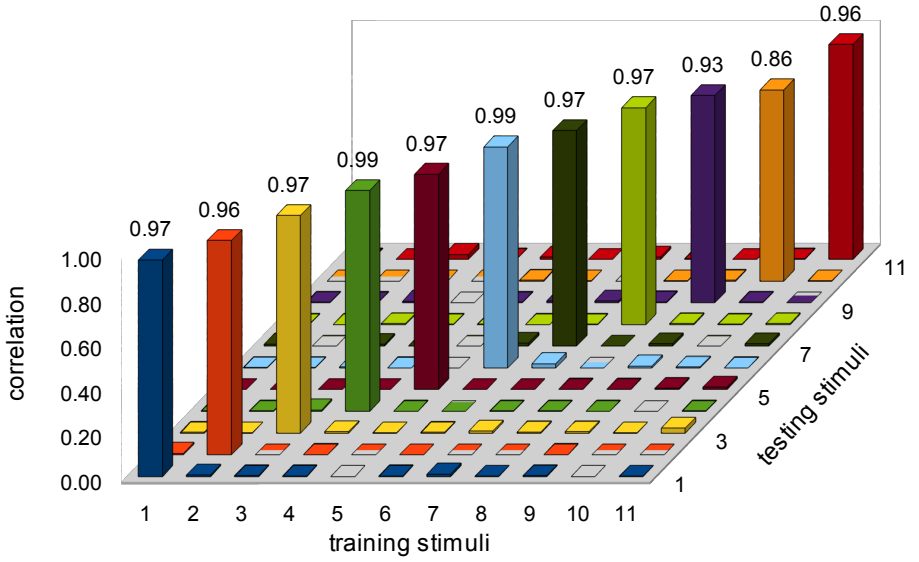
Quantitatively, Graph 5 and Graph 6 present the initial and the final pairwise correlations, respectively, between the testing stimuli and the corresponding training ones. The initial correlations (average 0.57) might look pretty high but considering the fact that the ablated regions consist only 9% of the images (i.e. 91% of the image information remains the same) the initial correlations are very low. On the other hand, the final correlations average at a value of 0.96 which is very high and consists a great recognition performance over the testing stimuli. As for the correlations with the non-corresponding stimuli they all remain at very low levels, less than 0.02, which shows that the system retains its specificity.

An interesting aspect of this experiment is the trajectories that the states of the system describe in their course. The first thing to notice is that the time of

convergence varies significantly among the testing stimuli and it seems to depend on the location of the ablation in conjunction with the complexity of the stimulus at that position. The richer the stimulus the easier for the neural assembly to fill in the missing points.



Graph 5: Initial pairwise correlations



Graph 6: Final pairwise correlations

However, what is more interesting in this experiment is the metastability of the dynamical systems. Figure 12 depicts two instances of this phenomenon that occurs in the dynamics of stimuli #1 and #6.

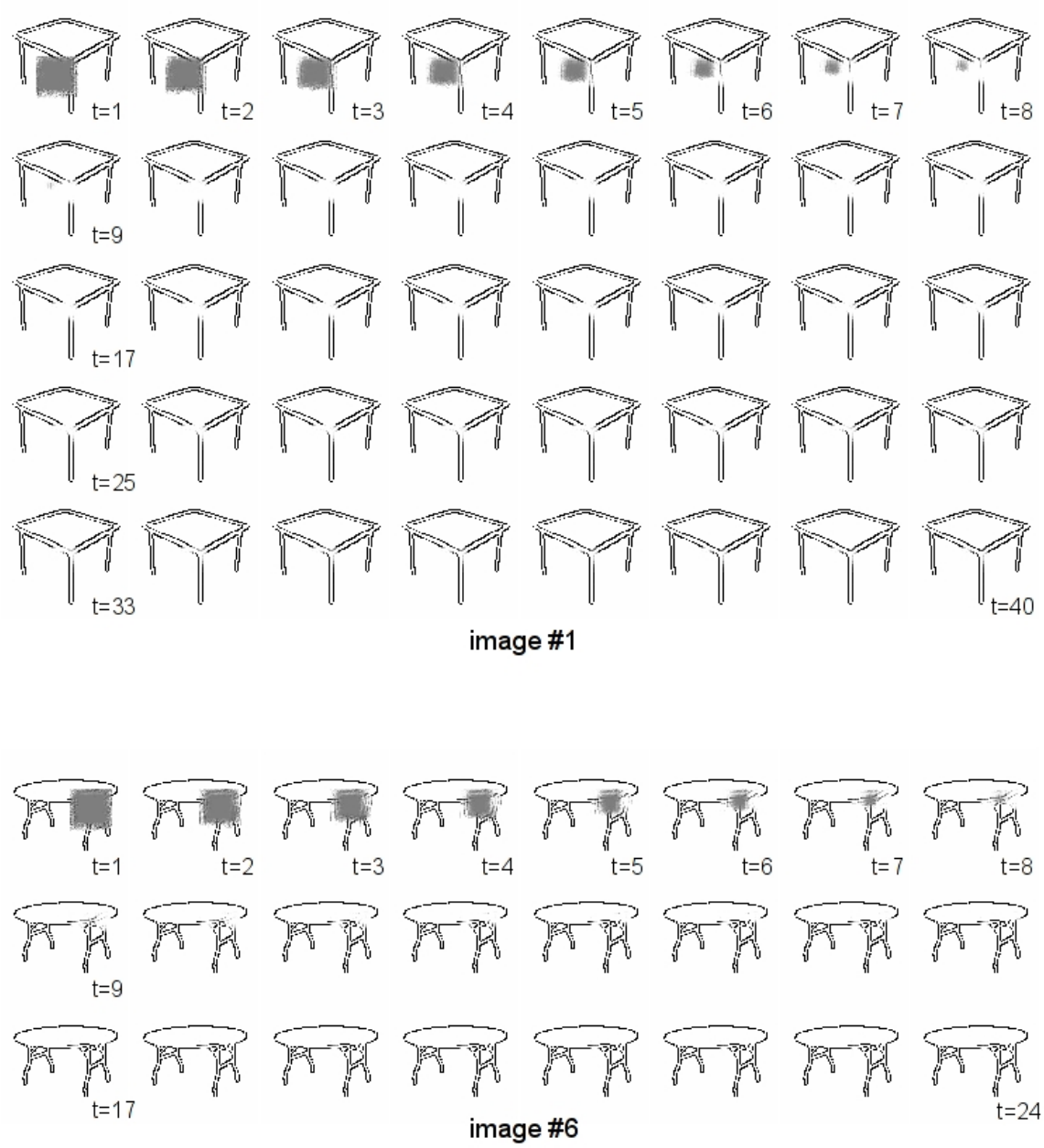


Figure 12: Experiment 2: Metastability in visual recognition

As explained earlier, pixels depicted in gray are visualizing noisy (unstable) states that gradually converge to stable attractors (i.e. pixels depicted with white or black color). However, we see in the figure above that there are many attractor networks that initially converge to one attractor (in this case *white*) and later make a transition to another one (in this case *black*). As we can see, by time step 8, all noisy states in *gray*, for both testing stimuli #1 and #6, have reached a *black* or *white* state. But, starting in time step 9, some *white* states start converting to *black*.

Metastability is a well known phenomenon not only in physical dynamical systems but in the brain as well. According to Kelso and Tognoli 2007 "metastability is a result of a symmetry-breaking caused by the subtle interplay of two forces: the tendency of the components to couple together and the tendency of the components to express their intrinsic independent behavior. The metastable regime reconciles the well known tendencies of specialized brain regions to express their autonomy (segregation) and the tendencies for those regions to work together as a synergy (integration)". So, the *white* states in the figure are transient metastable states that last for a few time steps before they converge to a more stable final state. This metastability of the neuronal interactions is also capable of causing synchronous oscillatory activity (Mioche and Singer 1989). However, in this experiment we only observed oscillations in single neurons that were not able to cause any significant oscillations at the level of the attractor networks.

5.4.3. Experiment 3

In the previous experiments we tested the sensitivity of the system to input variability and noise. In the third experiment we test the opposite, its specificity to unseen stimuli. We train the system with a set of 10 out of 11 stimuli and then we test the recognition performance on the stimulus that was excluded. The experimental process is repeated for all 11 stimuli of Figure 8. A good specificity performance would mean that the system can differentiate the unseen stimuli from the learned ones. Provided with a testing stimulus that was not part of the training set the system should converge to a state that differs substantially[4] from all the training stimuli. Needless to say, given that the stimuli are clear of noise, the system has a perfect recognition performance on the learned stimuli. Therefore, we will only discuss the tests on the previously unseen stimuli.

Figure 13 depicts in each row the results of the 11 repetitions of the specificity experiment – one for each testing stimulus that is excluded from the training set. For the training process we used the Widrow-Hoff learning algorithm with a stopping criterion a vector similarity (cosine) of 0.9999 or 200 iterations, whichever comes first. The parameters of the attractor networks are all fixed to α=0.9, γ=0.2, and δ=0 and we let the whole system iterate for 50 time steps. The figure presents the visualizations of the system's state in iterations 1 through 5, and 50. The reason that in this experiment we use a parameter of δ=0 instead of δ=1 is because we want the system to evolve devoid of the initial input. Otherwise it will be hard to make transitions to radically new states.

4 We have to keep in mind that the proposed framework is not a classifier that makes decisions or gives outputs with discrete labels. It's only through the visualization of the states that we acquire a sense of similarity or difference between stimuli.

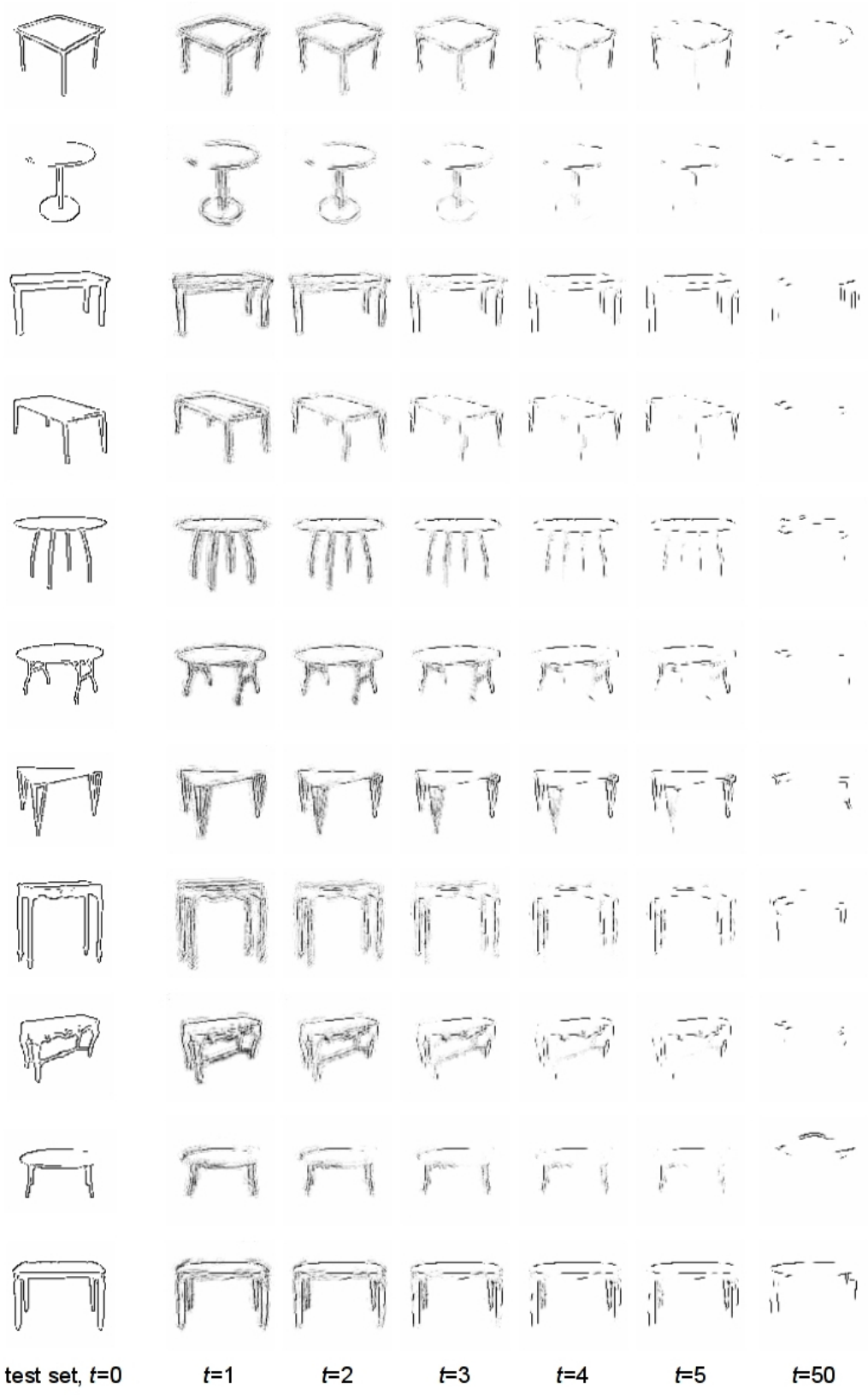
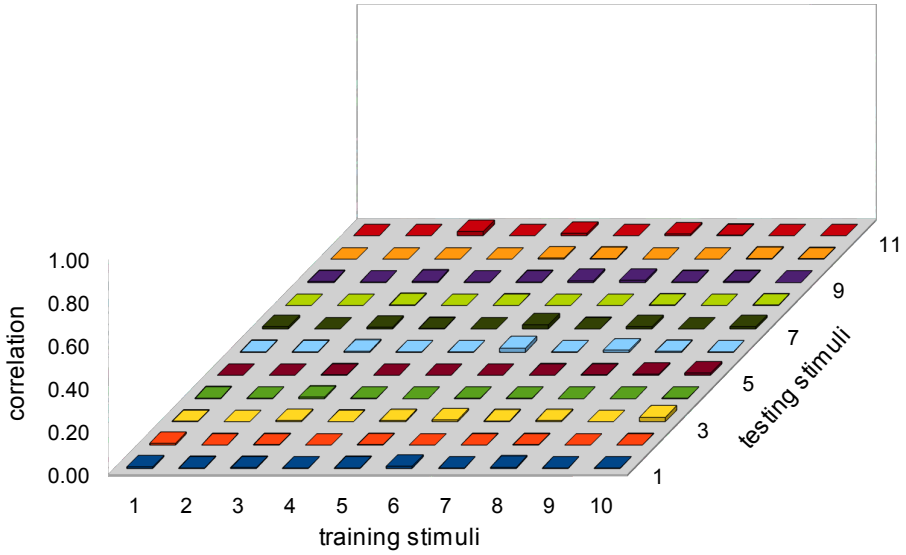
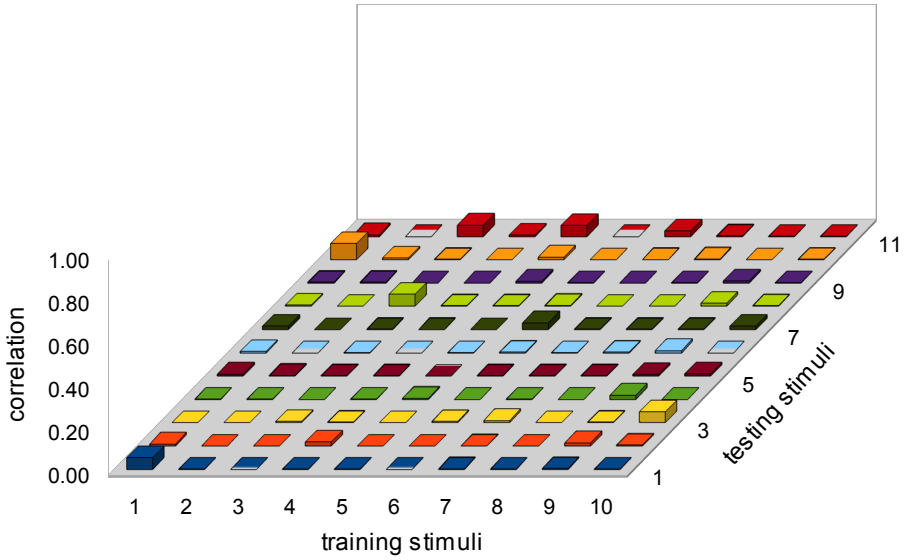


Figure 13: Experiment 3: Stimulus specificity

The results show that by time step 50 the output of the system is unlike any of the training stimuli. In fact, this happens much earlier, around time step 10, for most of the testing stimuli.



Graph 7: Initial pairwise correlations



Graph 8: Final pairwise correlations

In quantitative terms, Graph 7 and Graph 8 depict the correlations between the testing and the training stimuli before and after the dynamics takes place. The maximum correlation at the beginning has a value of 0.02 and it doesn't exceed a value of 0.08 at the end of the process. It's clear that the system exhibits a good specificity and manages to distinguish the unseen testing stimuli from the training ones.

5.4.4. Discussion

Stimulus sensitivity and specificity in humans depends on a variety of factors that pertain to different stages of cognitive processing (e.g. the perceptual context, the nature of the noisy/ablated part, the distinctiveness an image part provides to the object it belongs to, the visual/verbal priming, etc.). Although our experimentation in this study can only account for a certain stage of processing and the results we acquire do not generalize easily, there are still some useful conclusions and predictions we can make.

The tolerance of the system to random noise up to some level (e.g. 70%) was not that surprising. But the minimal decline in the performance when noise reached a level of 90% was kind of unexpected. This is mainly due to the fact that we tend to judge the easiness of a task based on our own performance to similar tasks. Experimental findings (Biederman 1987) show that humans can preserve their recognition performance with degraded objects at various levels of information loss. But these experiments do not show what stages of processing are involved in this behavior. The proposed framework, in its current form, is agnostic of the higher level associations and the top-down effects of cognition. It doesn't have a

holistic view of the images it perceives. Humans, to the contrary, behave very differently. The way they eliminate noise in the image is not by simply applying a local or semi-local (e.g. 8 hetero-associations) process. The fact that our relatively naïve approach is able to eliminate up to 90% of noise while humans can't go that far, suggests that there are holistic processes in the brain that work at a global level and prevent us form doing so. These processes may not have a great influence at lower levels of noise (e.g. up to 70%) but when noise reaches a higher level they hinder visual recognition. We could test this prediction by having humans recognize small patches of our testing stimuli instead of the whole images. We hypothesize that in this case, where higher associations are not present, humans would have a difficulty with recognition even at lower levels of noise (e.g. 70%). If this is verified it shows that the benefits of higher-level associativity come at the expense of side effects. The more complex a system gets the less unbiased it can remain from certain influences. And this is something we empirically observe in cognition in general.

In the second experiment we observed a lot of metastability during recognition. This metastability was always a one-time switch from the wrong attractor to the correct one without leading to any oscillations. But what is more interesting is the location it occurred. If you had the chance to observe the visualizations in a higher resolution you would see that the metastable attractor networks lie in the middle of the ablated area, i.e. the deepest inside the noisy image part. This is not a coincidence. These networks have no direct connectivity with the networks that process the clean image parts, so it takes time for the associative information to reach their location. Therefore, it is to be expected that in the

absence of the right information they might reach the wrong attractor before they switch to the correct one. If we didn't provide enough time for the system to converge, no metastability would occur and some networks would have concluded in the wrong attractor. This is probably the strongest evidence for the significance of dynamics. Conventional approaches to pattern recognition do not provide for this flexibility. From an application point of view this might not be that important, but from a cognitive modeling point of view this is quite crucial. If we want to be able to account for human behavior then we can't neglect the temporal aspect and the various phenomena of metastability it causes all the time.

5.5. Study II – Affine transformations

The primate visual system possesses a remarkable ability to recognize various types of affine transformations of learned stimuli. Evidence from psychophysics, neurophysiology and behavioral experiments shows that visual recognition can to some extent be shift, size, and rotation invariant (Biederman and Cooper 1991, Biederman and Cooper 1992, Tovee et al. 1994, Ito et al. 1995). From a mathematical point of view all affine transformations are grouped in one class of transformations, but this is not necessarily true from a cognitive perspective. There is no evidence so far supporting a unified mechanism capable of all affine transformations. To the contrary, the fact that the various transformations require unequal effort (in terms of time) during visual recognition implies that there are different mechanisms that take place. For instance, behavioral performance on scaled stimuli seems to be quite effortless while performance for rotated stimuli is correlated with the angle of rotation (Shepard and Metzler 1971). In another case, neurophysiological data shows that view-tuned IT units (VTUs) are quite tolerant to scaling and translation in the image plane but not to rotation in depth (Logothetis et al. 1995). It's quite clear from this data that we shouldn't be looking for a unified model that is so powerful to account for all possible affine transformations. Moreover, with regard to rotation of mental imagery in particular, there is experimental evidence showing that it requires a considerable post-perceptual effort and involves multiple neural systems, at least the visual and the motor one (Parsons 1987).

Therefore, in this study, we will not consider all affine transformations but will focus our experimentation on two-dimensional translation and scaling. Shift and size invariance is well documented in literature and there exist quite a few computational models that are trying to account for it (Wiskott 2004). Most of these models adhere to one of two approaches: *normalization* or *invariant features*. In the first case, visual recognition is preceded by an internal transformation to a standard position and size. In the second case, visual recognition is performed on a set of invariant features that are extracted from the image. Both of these approaches have pros and cons and neither one can provide a convincing explanation to the issue. Moreover, it seems from the research findings that shift and size invariances are only particular instantiations of a more generic brain ability to process invariant representations. If this is indeed the case then it would be much more interesting, instead of building a customized system that targets a certain invariance, to examine the potential and the limitations of a more developmental approach. This is the approach we will take in this study. We will simulate an ideal observer and the learning that occurs in her perceptual system during cognitive development and then test what kind of invariances, if any, the system is able to develop.

We will run two experimental scenarios in order to decompose the two components of optical flow, translation and scaling. In the first one we will examine the kind of shift/size invariance that can be developed by learning translated stimuli, and in the second one we will repeat the process for scaled stimuli.

5.5.1. Experiment 4

In this experiment we will presume an observer that is able to move her gaze parallel to the scene while preserving her distance from it. We will also assume that in the course of development the observer will have the chance to observe, and theoretically learn, all possible displacements of a certain stimulus. For the sake of simplicity we will experiment with a single artificial stimulus, a rectangular wireframe with a size of 20x20 points. We will further assume that the observer is able to scan the visual field for a range of 40x40 points. This process will cause the projection of 1600 different images onto the visual system. Figure 14 presents a random subset of these stimuli.

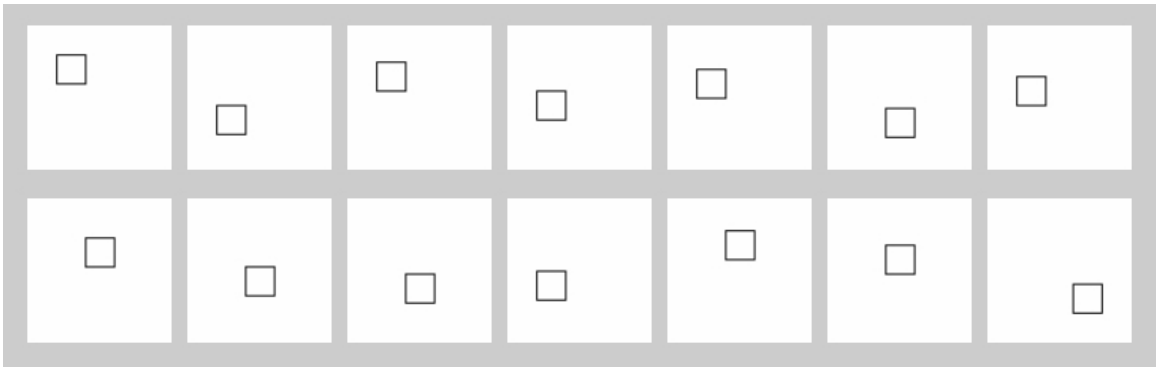


Figure 14: Experiment 4: A random subset of the 1600 training stimuli

If we allow for a prolonged exposure of the system to these images, we would expect that the associations that are formed will uniformly pick the principal statistics of the stimulus. Of course, the attractor networks of the proposed system can only learn a small number of these associations, so learning will never complete. However, after a large number of input presentations, some primitive form of module assemblies should have been learned (e.g. corners with edges, edges with edges, etc.). Most important, though, is the fact that none of

these associations will hold a global view of the stimulus they are collectively tuned to respond to. There will be no representation of size or position of this stimulus whatsoever. The only thing encoded in these small attractor networks will be the local associations with neighboring networks. Nevertheless, the system should still be able to respond to stimuli that share the same local properties with the rectangular wireframe regardless of their position and size. Not every case will be that simple, of course, but a certain degree of scale and position invariance should be able to emerge.

Figure 15 depicts the first set of results for this experiment. In the left column we see the testing stimuli that consist of three types of shapes: a *rectangular*, a *diamond*, and a *circle*. All testing stimuli have an equal size with the training stimuli in Figure 14 but are positioned in random locations of the visual field in order to test the system's response to stimulus translation. If the system has developed some form of position invariance it should be sensitive to the *rectangles* in different positions but insensitive to the *diamonds* and the *circles*. For the training process with the 1600 stimuli we used the Widrow-Hoff learning algorithm. In this case there was no stopping criterion as in the previous study because it's impossible for the networks to learn all 1600 configurations. Therefore, we trained the system for a fixed number of 100 epochs (repetitions of the full training set). Each network had 4 hetero-associations with the adjacent networks. The parameters of the dynamics were all fixed to α=0.9, γ=0.2, and δ=0 and we let the whole system iterate for 80 time steps. The processing time for the whole experiment, using 16 CPUs, was 5h 40min. Most of the processing was allocated to the high demands of learning.

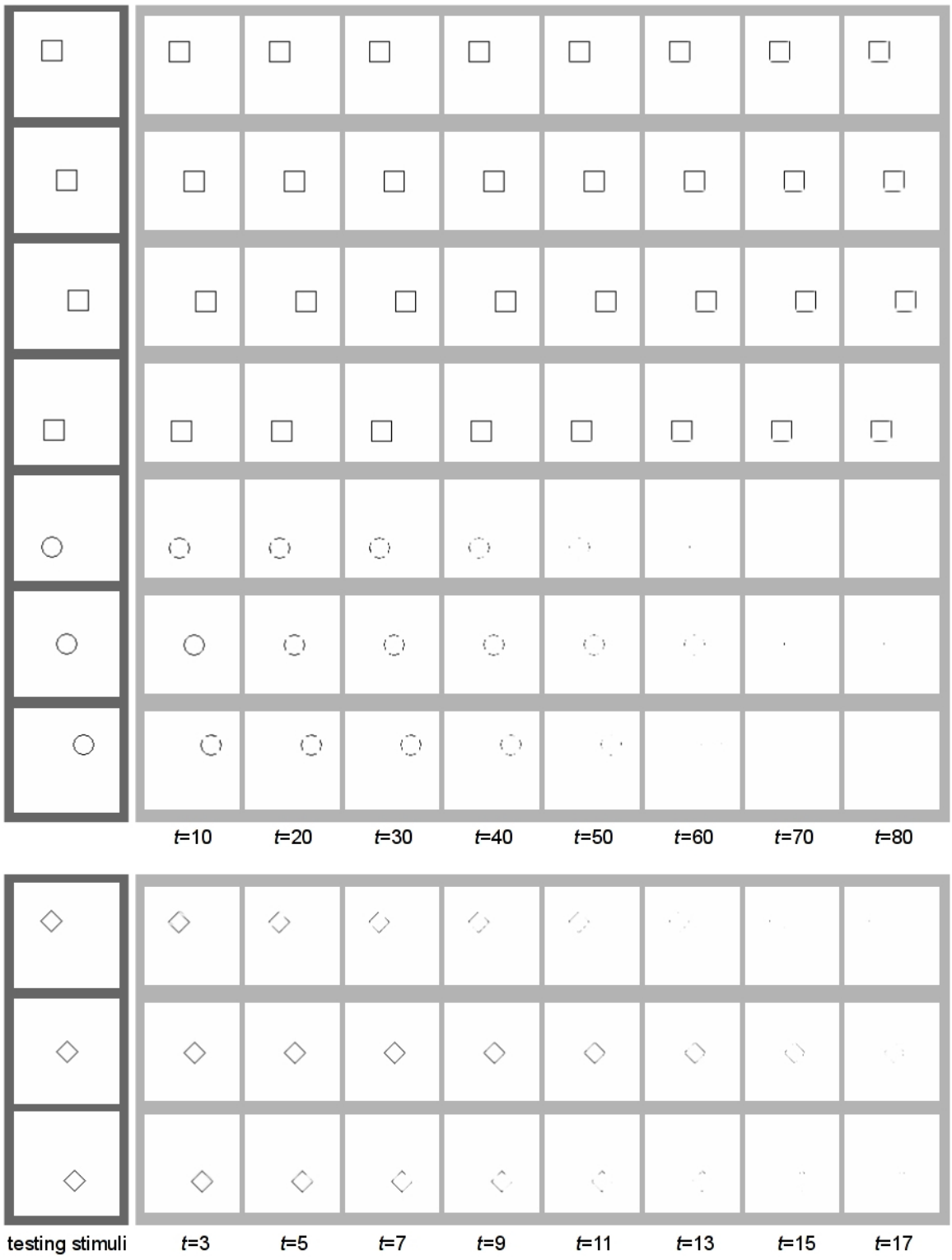


Figure 15: Experiment 4: Position invariance with translated stimuli

As depicted in Figure 15 it takes about 15 steps for the system to reject the *diamonds* and around 50 steps to reject the *circles*. Recognition of *rectangles* in any location is pretty robust for much longer up to the last iteration. We have to stress once again that there is no binary decision for recognition or rejection of a stimulus. Stimuli that are not recognized (i.e. rejected) are fading in time while the ones that are recognized can simply retain an active module assembly for much longer a duration. The reason why *circles* are harder to reject than *diamonds* is because the pixelation of *circles* in this particular experiment is causing a lot of overlap with the *rectangles*. This is more obvious in the next test-set (Figure 16) where the shapes are larger and, therefore, *circles* have a larger overlap with *rectangles*. The results of this experiment may not seem that surprising since all the recognized *rectangles* belonged to the initial training set. However, considering the fact that the training process was by far unable to complete the learning of all those stimuli, it is quite surprising that even a minor tuning to the target stimulus was able to generate this clear recognition performance. This shows that dynamical systems are not only robust to conditions of incomplete training but, most important, that perfect learning is not necessarily a prerequisite for a good visual recognition.

Figure 16 depicts the results of a second test-set where, contrary to the previous one, the three types of shapes (*rectangular*, *diamond*, *circle)* have varying sizes (see the left column in the figure). With this test-set we want to examine the extend to which size-invariant recognition can emerge from a system that is trained exclusively on translations of a fixed-size stimulus.

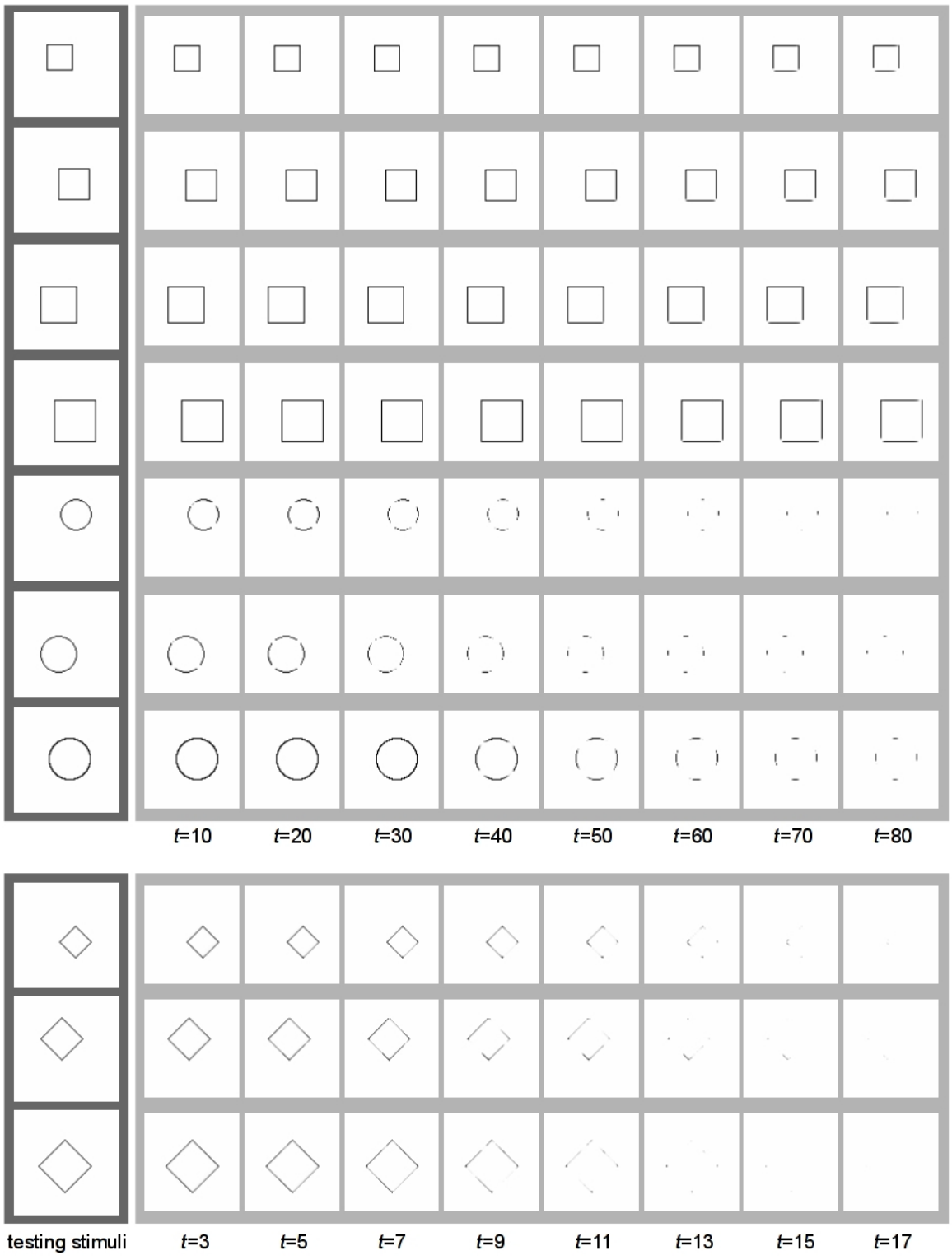


Figure 16: Experiment 4: Shift/size invariance with translated stimuli

The results are on the same line with the outcome of the previous test-set. The system is able to recognize *rectangles* in various sizes and positions and rejects *diamonds* and *circles* regardless of their size and position. The time it takes to reject a *diamond* seems to be independent of its size. In the case of *circles*, the larger the shape the larger the overlap with a *rectangular* and the longer the time to reject it. In either case, though, there is a clear rejection whereas recognition of rectangles lasts for much longer.

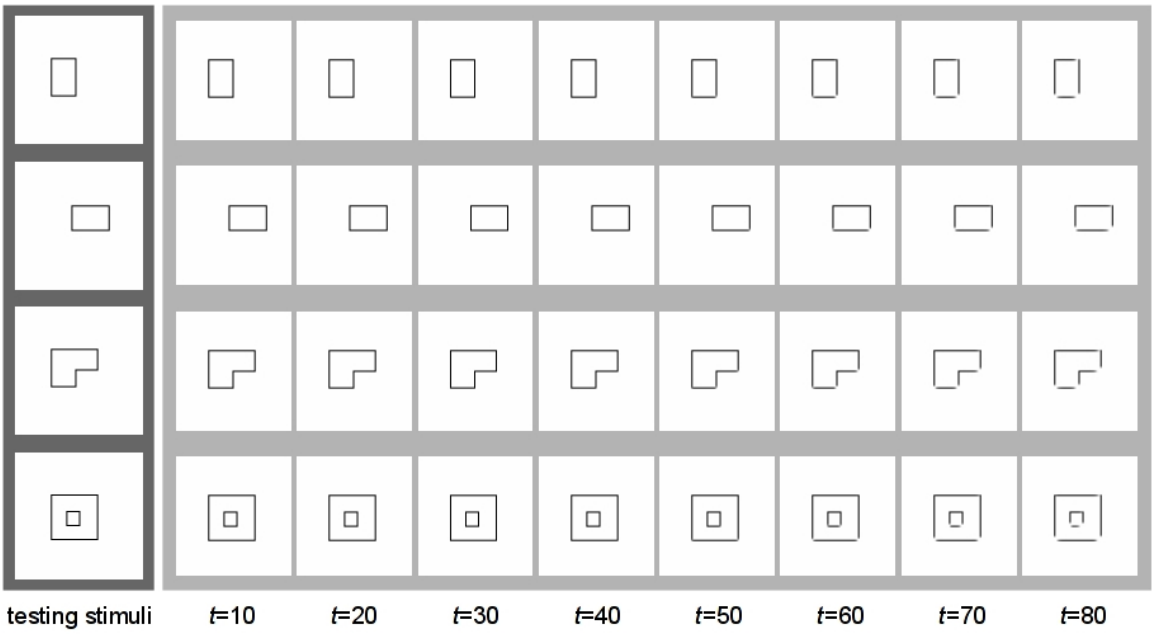


Figure 17: Experiment 4: False recognition of non-learned stimuli

These results give a nice illustration of how a seemingly trivial developmental process of visual learning can elicit a neural representation that despite being agnostic of the whole can still make distinctions between learned and non-learned stimuli in a variety of scales and positions. This doesn't mean that the proposed approach is always successful. Figure 17 shows examples of false recognitions that are due to the very fact that training in this experiment was

based on a single and simple *rectangle* only. In a more naturalistic scenario, though, there will be a richer set of associations that can presumably overcome this kind of shortcomings. In either case, the emergence of some form of shift and size invariance without employing thorough training or elaborate processing is definitely compelling.

One last note. We can also think of the *diamonds* as a rotated version of the *rectangles*. If this is the case then it shows that the framework, in its current configuration at least, has no capability for rotation invariant recognition. This is not surprising – in the beginning of this study we stressed the fact that not all affine transformations use the same mechanisms, and that mental rotation requires post-perceptual effort that recruits multiple neural systems. However, it's very likely that the rotation invariance could simply depend on the perceptual encoding and the relational representation of local image parts. This is supported by behavioral studies on the discrimination of inverted faces (Rossion and Gauthier 2002). Therefore, it's not unlikely that an alternative configuration of the same framework could exhibit some rotation invariant properties.

5.5.2. Experiment 5

In this experiment we test what kind of invariances the system might develop if the learning is based on scaled, instead of translated, stimuli. We will presume an observer that fixates her gaze to the center of the scene and moves back and forth in order to observe different scales of the image. For the sake of simplicity we will experiment again with a single artificial stimulus, a rectangular wireframe. In the current scenario the observer will have the chance to observe, and

theoretically learn during development, 45 scales of the stimulus. Figure 18 presents an ordered subset of the projected images. Although the size of the training set is much smaller than in the previous experiment, learning is still quite difficult to complete in this case.

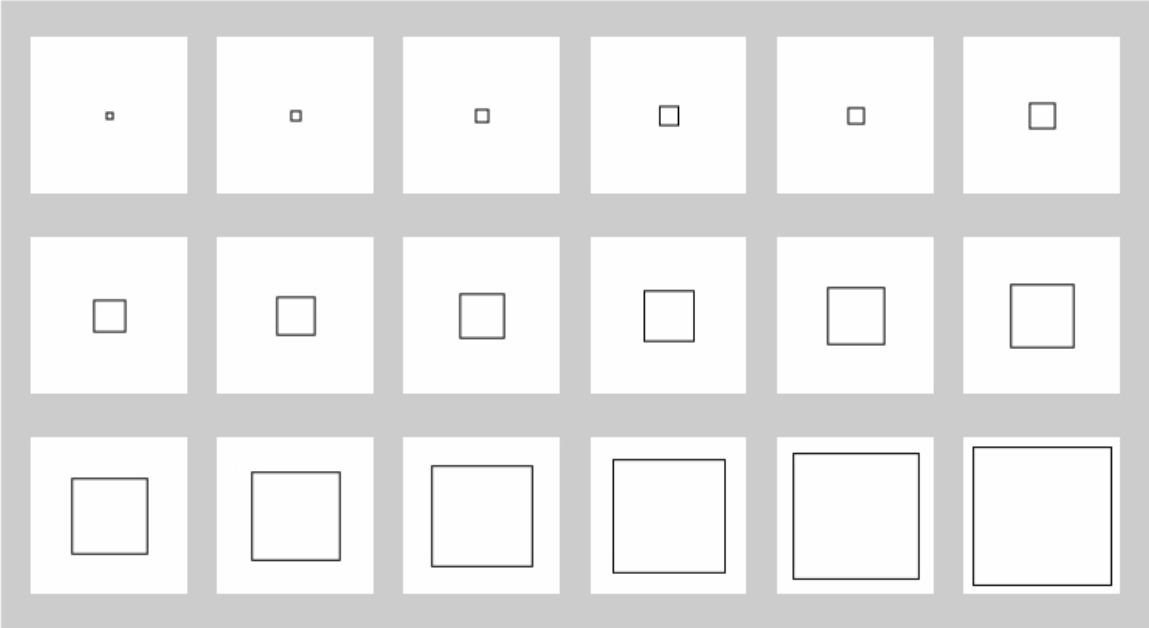


Figure 18: Experiment 5: A subset of the 45 training stimuli

It's obvious that this kind of developmental setup doesn't provide for a learning experience as rich as in the previous setup. Most of the attractor networks will still be able to experience a rich set of *edges* but they won't learn as many associations with *corners* as before. Therefore, we should expect a mixed performance where recognition will mostly depend on the position of the testing stimulus (i.e. whether the underlying attractor networks had the chance to learn the needed associations). On the other hand, the size of the testing stimulus shouldn't be a major issue.

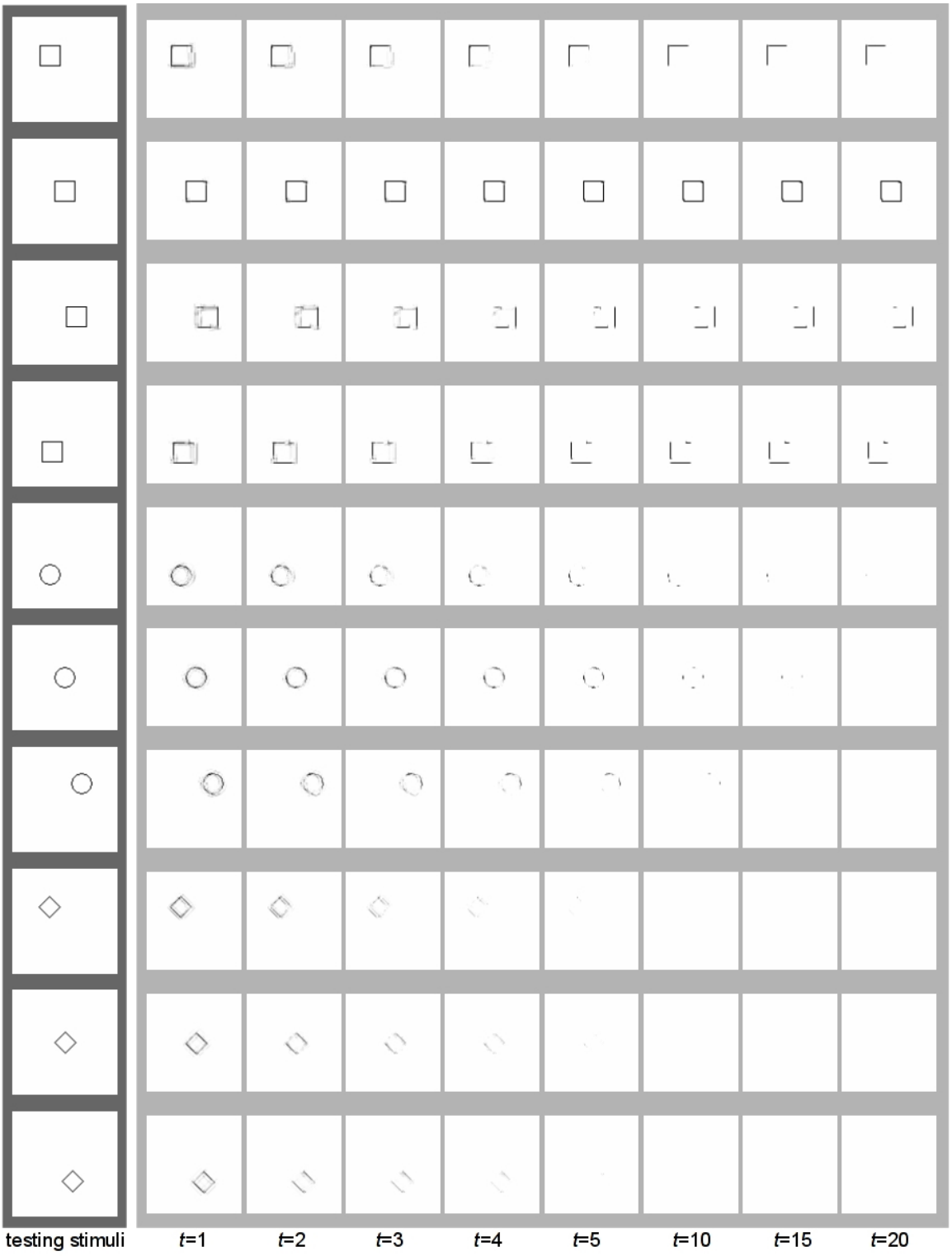


Figure 19: Experiment 5: Position invariance with scaled stimuli

Figure 19 depicts the first set of results for this experiment. In the left column we see the same set of testing stimuli we've seen in the previous experiment. All testing stimuli have an equal size but are positioned in random locations of the visual field in order to test the system's response to stimulus translation. For the training process we used the Widrow-Hoff learning algorithm for a fixed number of 100 epochs. Each network had 4 hetero-associations with the adjacent networks. The parameters of the dynamics were all fixed to α=0.9, γ=0.2, and δ=0 and we let the whole system iterate for 80 time steps. The processing time for the whole experiment, using 16 CPUs, was 18 min.

As expected, the system has a mixed performance. It takes about 10 steps to reject the *diamonds* and the *circles* (faster than before) but the recognition of *rectangles* is not very consistent. It clearly depends on the location of the testing stimulus. When the position of the testing stimulus happens to coincide with the position of a training stimulus (e.g. in second row) the recognition is excellent. In all other cases, though, the system achieves a partial recognition that depends on the overlap between training and testing stimuli (e.g. first, third, and fourth row). Performance on the second test-set is roughly the same. As depicted in Figure 20 rejection of *circles* takes a bit longer (because of their larger size and larger overlap with *rectangles*) but other than that there is no major difference in terms of recognition. As for the non-learned stimuli (Figure 21), their rejection seems to be a mere consequence of their position in the visual field and not a result of a correct recognition. For instance, a better (i.e. centered) positioning of the fourth stimulus could have very well produced the side effect of a good recognition.

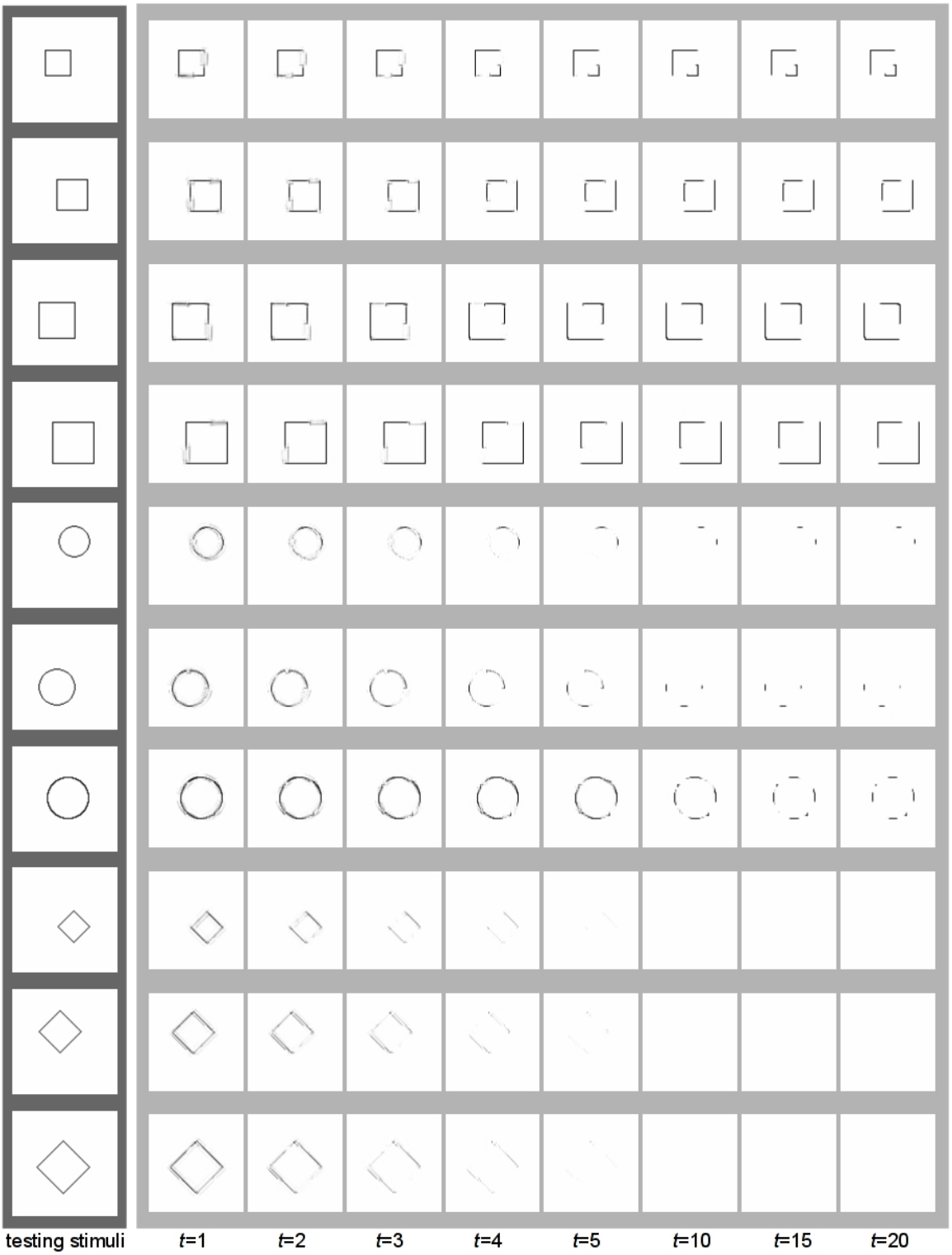


Figure 20: Experiment 5: Shift/size invariance with scaled stimuli

Figure 21: Experiment 5: Performance on non-learned stimuli

5.5.3. Discussion

At first glance, the developmental approach we followed may appear very different from what most models in literature are suggesting. This is not necessarily true though. In an abstract level of description, *normalization* and *invariant features* suggest a whole different approach for recognition. The first one claims that images have to be transformed in scale and position, whereas the second one suggests that there exists some set of invariant features we can extract from every image. How different are they in a lower level though? How differently can these two be implemented in a biologically plausible computational level? We believe that the proposed framework can provide some hints for these questions. By scanning the visual field during cognitive development we feed the system with normalized versions of visual percepts at various positions and scales. By training the attractor networks on the immense number of local

patches in their receptive fields we make them learn nothing but the most universal features. So the proposed framework shows that normalization and invariant features may be two sides of the same coin, simply in a higher level of description. After all, the real question is not whether the visual system employs *normalization* or *invariant features* but how exactly does it do it. The literature suggests that although the *invariant features* approach seems more prominent, both classes of models have pros and cons and it's very likely that the truth lies somewhere in between (Wiskott 2004). The proposed framework gives a hint for how this *between* might look like.

On another note, the fist study raised the question whether the recognition performance was good because of the small training set or because of the underlying properties of the proposed framework. This second study is giving some answers to this question. In experiments 4 and 5 the training sets consisted of 1600 and 45 stimuli respectively. One would expect the recognition performance in experiment 4 to be a lot worse than in experiment 5. This was not the case, though. Although in experiment 4 the system was never able to fully learn all the training stimuli and the learning phase had to stop prematurely in every run, the end result was much better. Because it wasn't the size of the training set that mattered most but the setup of the training. Depending on the goal of learning, it might sometimes be better to have a superficial but extensive training than a complete but limited one. The minor tuning of the system from an exhaustive but incomplete learning proved to be more adequate in this case. This study showed clearly that by increasing the size of the training set we don't necessarily have a deterioration in the performance of the proposed framework.

Of course, the goal of learning plays a crucial role too. This couldn't have happened with any kind of cognitive task.

Last, we have to note that this study exemplifies the need for a hierarchical architecture. The failure of the system to retain its specificity in the case of the non-learned irregular stimuli shows clearly that a single layer architecture has certain limitations. When it gets to more complex visual configurations it's impossible to form the variety of associations that are needed for recognition without constructing some kind of hierarchical associativity. And this is regardless of the number and the distribution of the hetero-associations that are formed within each layer.

5.6. Study III – Reaction time

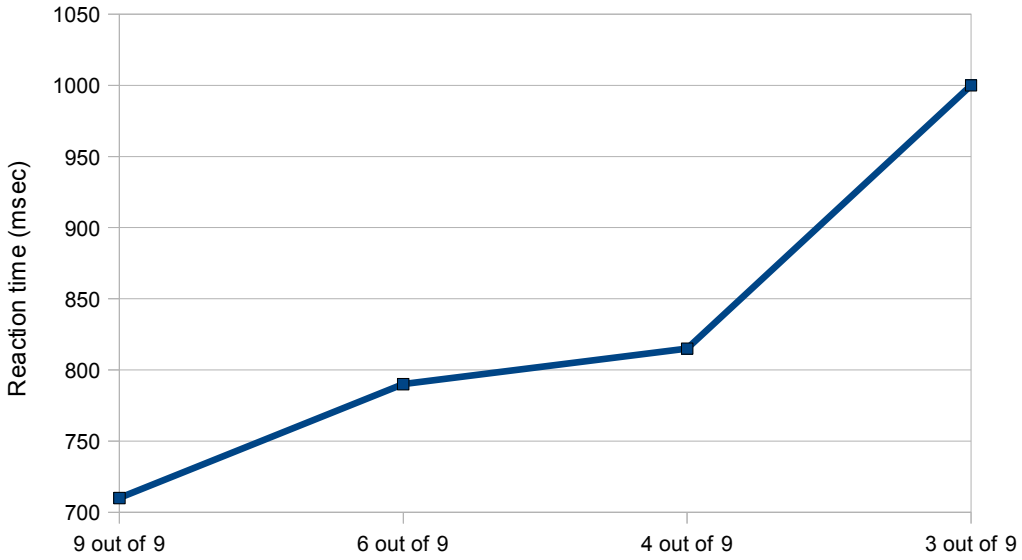
Reaction time is a critical aspect of cognitive processing because it indicates a variety of things like the amount of effort required, the difficulty of the task, the complexity of the stimulus, etc. The temporal dimension of processing, for biological and artificial dynamical systems alike, is expressed inherently through feedback loops and recurrent connections. This isomorphism gives a great advantage to dynamical systems approach when modeling of human behavior pertains to issues related to reaction time. Although it's hard to make quantitative comparisons between simulated and actual response times (Thorpe et al. 1996) a qualitative comparison should still be possible.

Behavioral studies in visual recognition (Biederman 1987) have shown that human performance is dependent on the degree and the type of completeness of the visual percepts. Absence of object parts, deletion of contours, and partial occlusion deteriorate the recognition performance and increase the reaction time. In this study we are experimenting with these conditions. In the first experiment we are testing how the size of an ablated image part affects the reaction time. In the second experiment we make a comparative study between ablations of mid-segments and vertices. In the third experiment we run a parametric analysis of the two of them combined. All results are compared quantitatively to the actual reaction times of behavioral studies.

5.6.1. Experiment 6

The more incomplete an object is the longer it takes for humans to recognize it and the more the mistakes they make. Biederman 1987 has studied this behavior

by training human subjects on 9 objects which consisted of up to nine components. Graph 9 presents the mean reaction time (ms) of recognition for testing stimuli that had a varying number of parts missing from the image. Although a thorough examination of the results shows that not all parts have an equal discriminative value, the general pattern shows clearly that an increase in the number of parts being absent from the image causes an increased reaction time for recognition.



Graph 9: Mean reaction time of incomplete objects for humans

In this experiment we will test how the proposed framework can account for this behavior. For the training stimuli we will use the same set of 11 images we used before (Figure 8). For the testing set, Biederman's stimuli were tailored to examine his theory that objects are formed by a number of components. In our case, though, there is no such claim so the testing stimuli are designed with simply an increasing degree of incompleteness without a preference for certain image locations or components.

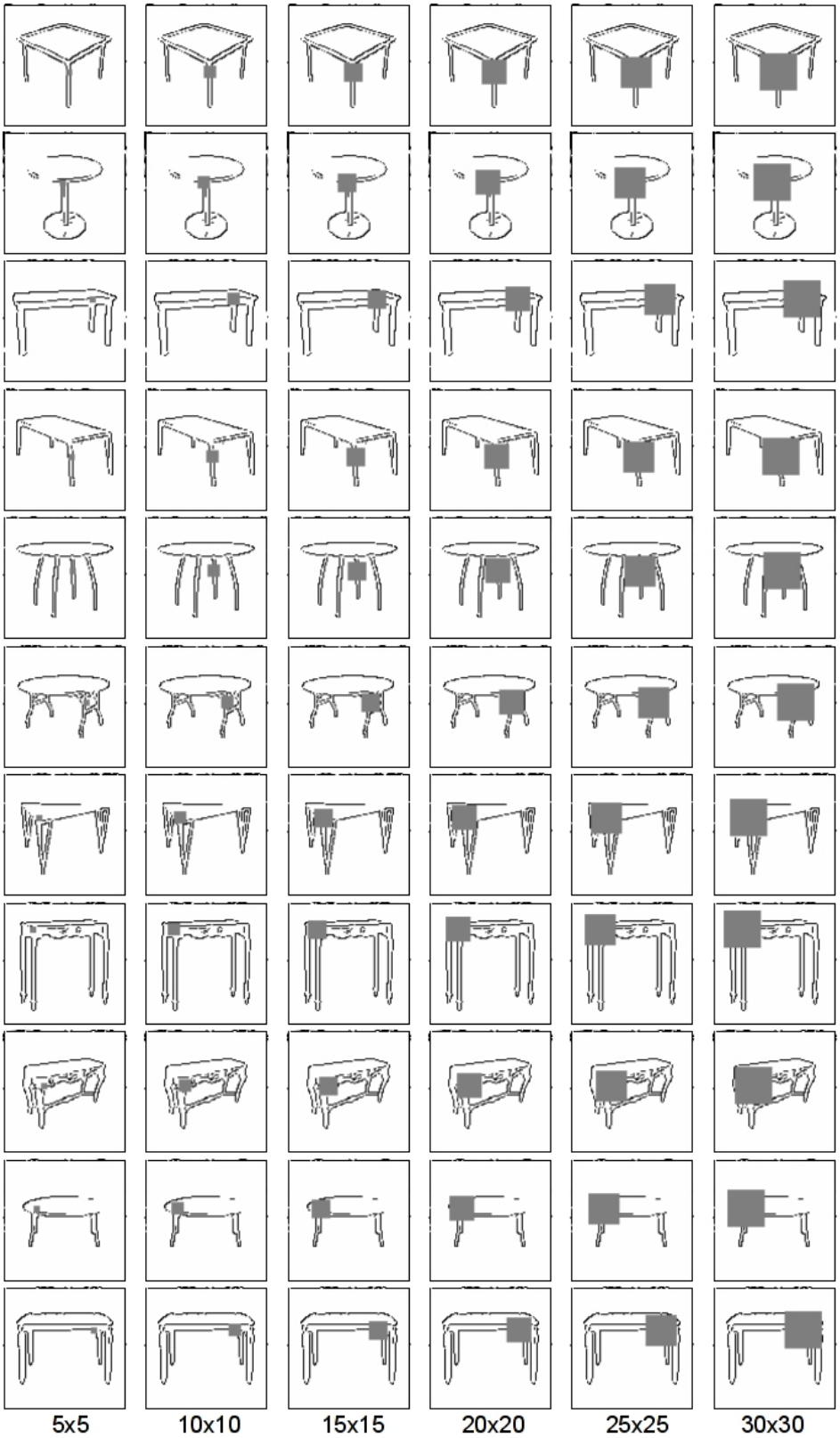


Figure 22: Experiment 6: Testing stimuli with 6 degrees of ablation

Figure 22 depicts the testing stimuli that consist of six degrees of ablation of the original training stimuli (5x5, 10x10, 15x15, 20x20, 25x25, 30x30 - in pixels). The ablations occur at random locations but all degrees of ablation for a certain stimulus have the same origin so that results are comparable. We would expect that as the size of the ablation increases the recognition performance would deteriorate and the reaction time would increase. A critical parameter in this setup, though, is what are we going to consider as the moment of recognition. As we stressed many times earlier, in a dynamical system like the one proposed there is no binary decision to be made, there is only a flow of activity. In the previous experiments time was not an issue so we could just stop the dynamics at a certain time step and compare the results. We can't do this in this experiment though. Time is what we are testing. What we do instead is to let the system iterate for a long enough number of iterations (e.g. 100 or 200) and then search through the flow of activity and pinpoint the moment that recognition stabilizes. Although moments of absolute stability are hard to find there exist quasi-stable states after which there is practically no change in the overall system's state. Figure 23 depicts these moments that we consider recognition as completed (one to one mapping with the stimuli in Figure 22).

Another thing we have to note here is that Graph 9 shows only the results of the successful recognitions and doesn't count the cases where recognition was not possible for the subjects. We won't do this kind of correction. We assume that recognition is always successful no matter how good the output is. Besides, from what we've seen in experiment 2, and is verified here as well, recognition performance for this type of stimuli is pretty high so it shouldn't matter anyway.

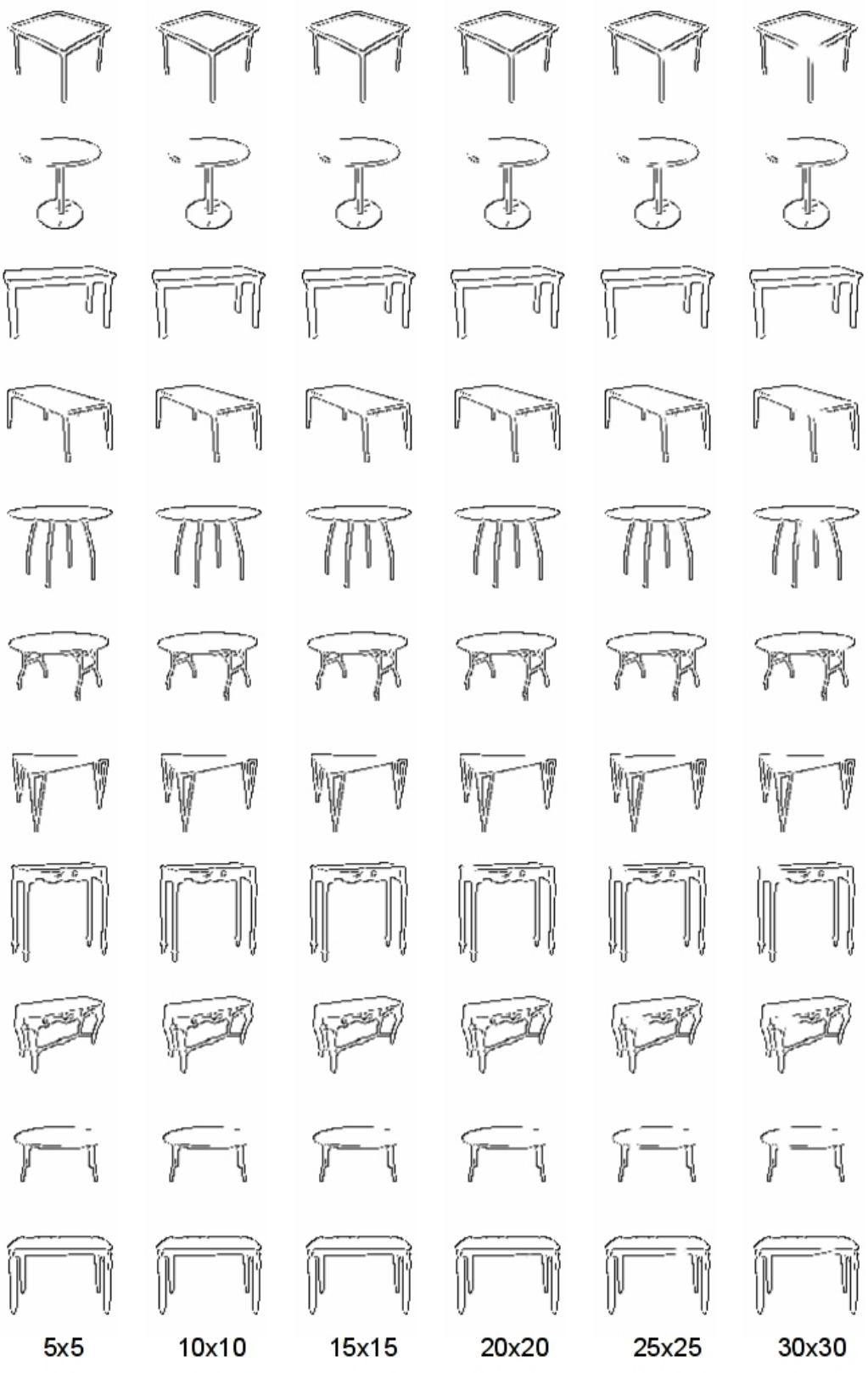
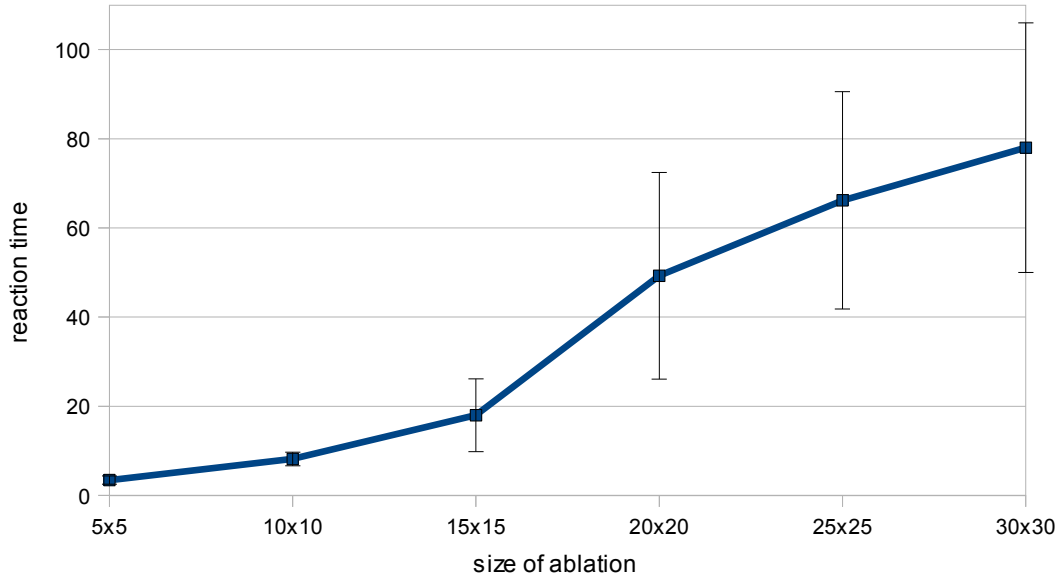


Figure 23: Experiment 6: Quasi-stable recognition (8 associations)

For the training process we used the Widrow-Hoff learning algorithm with a stopping criterion being a vector similarity (cosine) of 0.9999 or 100 iterations, whichever comes first. Each network has 8 hetero-associations with the adjacent networks. The parameters of the dynamics are all fixed to α=0.9, γ=0.2, and δ=1 and we let the whole system iterate for 100 time steps. The processing time for the whole experiment, using 16 CPUs, was 40 min.



Graph 10: Mean reaction time for 11 ablated stimuli (8 hetero-associations)

Graph 10 depicts the mean reaction time of the 11 testing stimuli for the 6 degrees of ablation. The reaction times are expressed with the number of iterations that occurred in the course of dynamics before the system reached the states depicted in Figure 23. Although there is no direct relationship with the reaction time of humans that is expressed in ms, it is obvious that the general trend of the graph is similar to the one in Graph 9. The larger the ablation the longer it takes to recognize the stimulus.

One can also notice in Figure 23 that the larger the ablation the less accurate the recognition is. The same trend exists in subjects as well. However, in the case of human behavior this results in errors of recognition, which are discrete and thus measurable, whereas in our case it results in lower correlation between output and training stimuli (as we've seen in experiment 2) which is hard and even unrealistic to convert to some form of recognition error. This is why in this study we're only focusing on the reaction times. Another thing to notice in Graph 10 is that the standard deviation of the reaction time increases disproportionally to the reaction time itself. The reason is that in the case of large ablations there is a high metastability in the system which causes greater fluctuations in the reaction time. This is not surprising considering the fact that for some stimuli the ablation can occupy an area as high as 50% of the image size and there is no visual context or other information to complement the loss.

One might wonder how this outcome relates to the connectivity of the networks. Would the same trend hold in case we changed the number of the hetero-associative connections among the attractor networks? In order to check this we ran the same experiment with an increased number. This time we doubled the connections from 8 to 16. We used a Gaussian-like distribution which means that each network connects to the 8 adjacent networks plus 8 out of 16 networks that lie right next to them. All other parameters remained the same. Figure 24 presents the results of recognition which are slightly improved but overall quite similar to those in Figure 23. This makes absolute sense since the increased connectivity can facilitate a higher associativity and thus a better recognition performance.

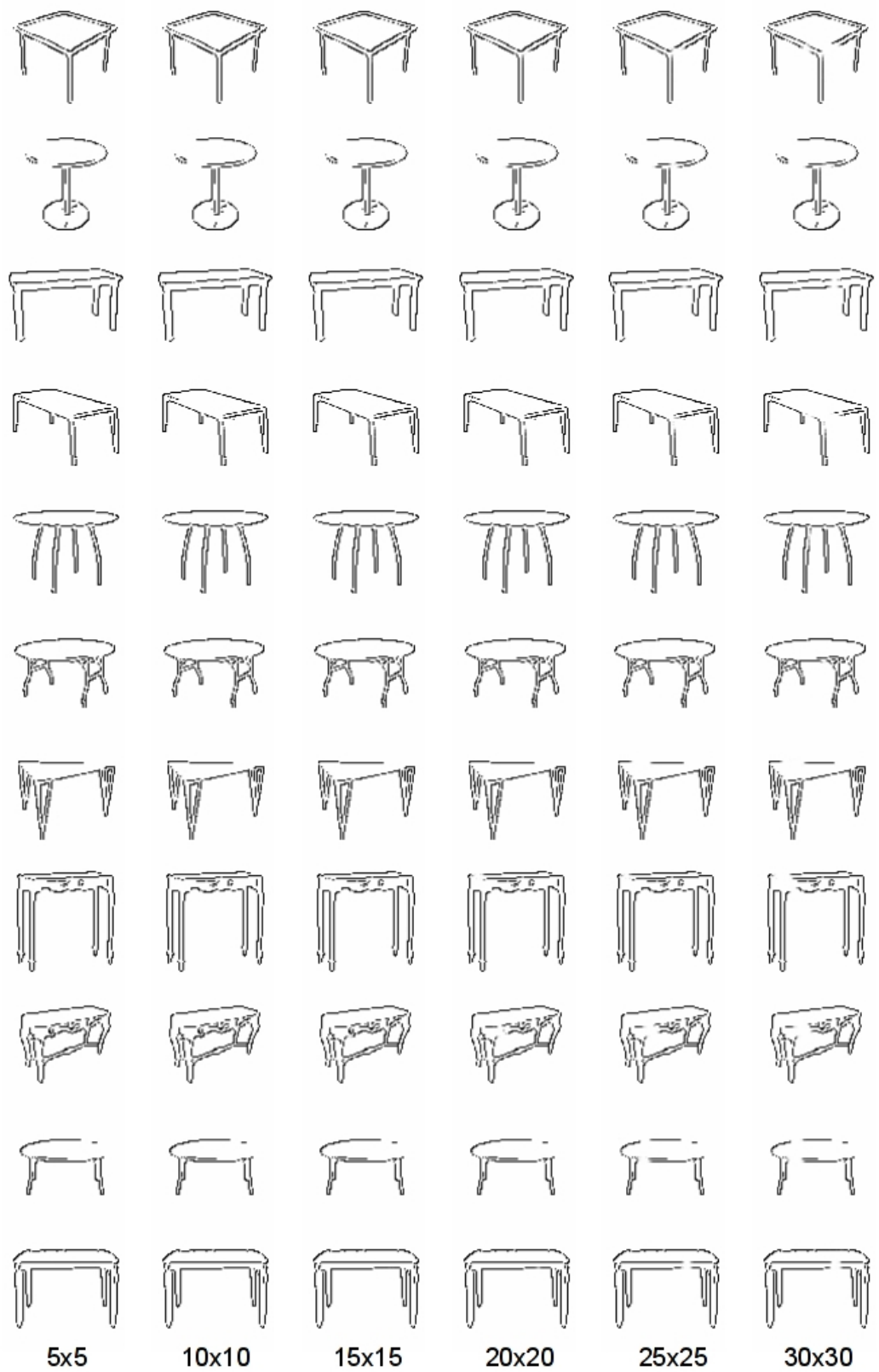
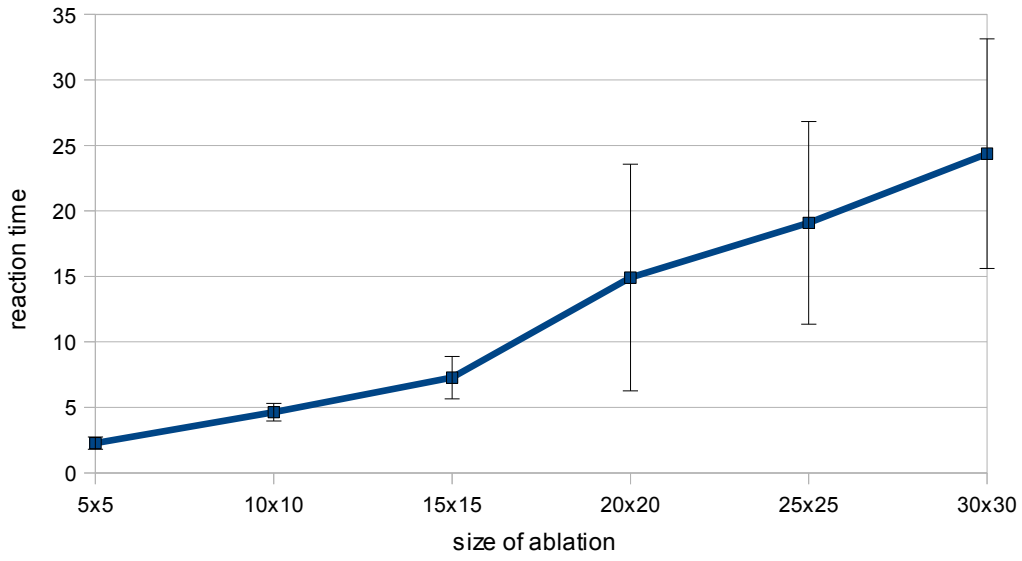


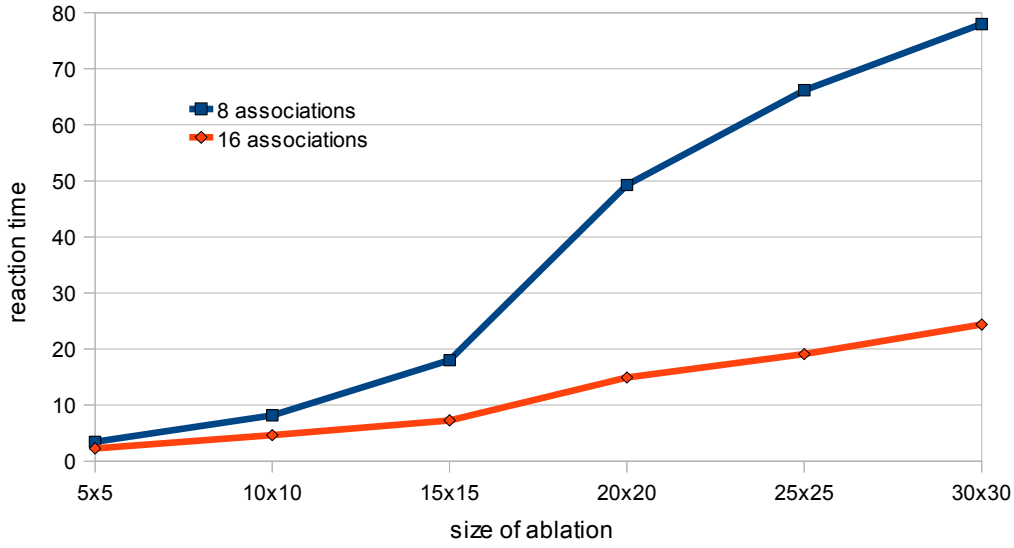
Figure 24: Experiment 6: Quasi-stable recognition (16 associations)

Graph 11 depicts the mean reaction times of the 11 stimuli for the 6 ablation conditions. The overall trend is the same as before and this is also true for the standard deviations as well.



Graph 11: Mean reaction time for ablated stimuli (16 hetero-associations)

In absolute values, though, both reaction times and standard deviations for the case of 16 associations are much smaller than for the case of 8. Graph 12 depicts the effect of this connectivity in the response of the system. The reaction time with 8 associations for small ablations was as short as 3 iterations but for large ablations it reached the 100 iterations in many cases. In the case of 16 associations the reaction time for small ablations was as short as 2 iterations (not that different) but for small ablations it never exceeded the 45 iterations (quite different). Moreover, it seems that with a higher number of associations the system becomes more stable while the trend of the reaction time acquires a more linear relation with the size of the ablation.

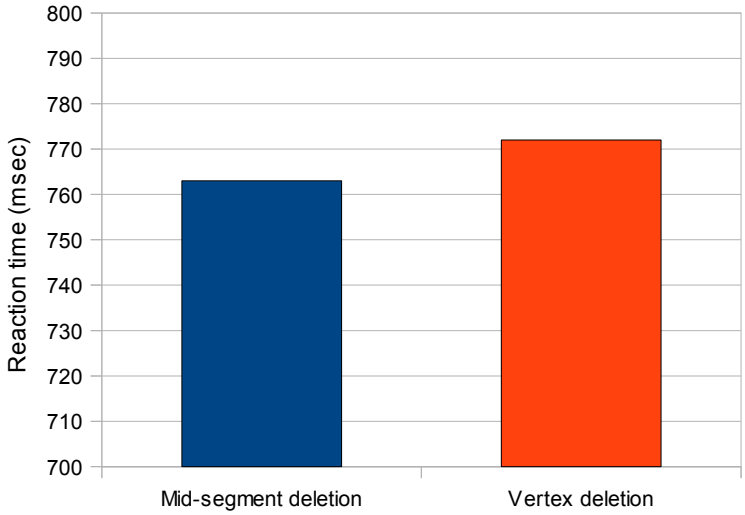


Graph 12: Effect of the degree of connectivity on reaction time

5.6.2. Experiment 7

In the previous experiment the ablation was agnostic of its location in the image. The result verified the findings of behavioral studies regarding the trend of the reaction time but it couldn't tell anything about how the complexity of the stimulus itself can affect the reaction time of recognition. The common intuition says that not all parts of an image have the same contribution to its recognition. This is true but it mostly pertains to recognition errors and not to the reaction time per se. Behavioral experiments with degraded objects and deleted contours have shown that reaction time in case of correct recognition is quite independent of the locus of deletion (Biederman 1987). In one of these experiments object contours were deleted in two types of locations, on mid-segments and on vertices, and presented to subjects for different durations. The results showed that there was no significant difference in the reaction time for the two conditions. Graph 13

depicts the mean reaction time for 18 objects with 25% of their contours deleted (averaged for exposure durations of 100ms, 200 ms and 750 ms).



Graph 13: Mean reaction time of humans

The mean reaction time for mid-segment deletions was 763 ms while for vertices it was 772 ms. A minor difference for sure. The goal of our experiment is to see how these two conditions relate in the proposed framework. For the training stimuli we will use the same set of 11 images that we used before (Figure 25a). For testing stimuli we designed two sets of stimuli one for each condition (Figure 25b for mid-segment and Figure 25d for vertices). The location of the ablations was selected manually in order to match the requirements of the experiment in the best possible way. We also tried to cover a representative number of mid-segments (or vertices) in each image while retaining the overall size of the ablation roughly similar for most stimuli. What we believe matters most, though, is the size of each ablated spot, not the sum of them, therefore we fixed this to 13x13 pixels for all the ablations and for every stimulus.

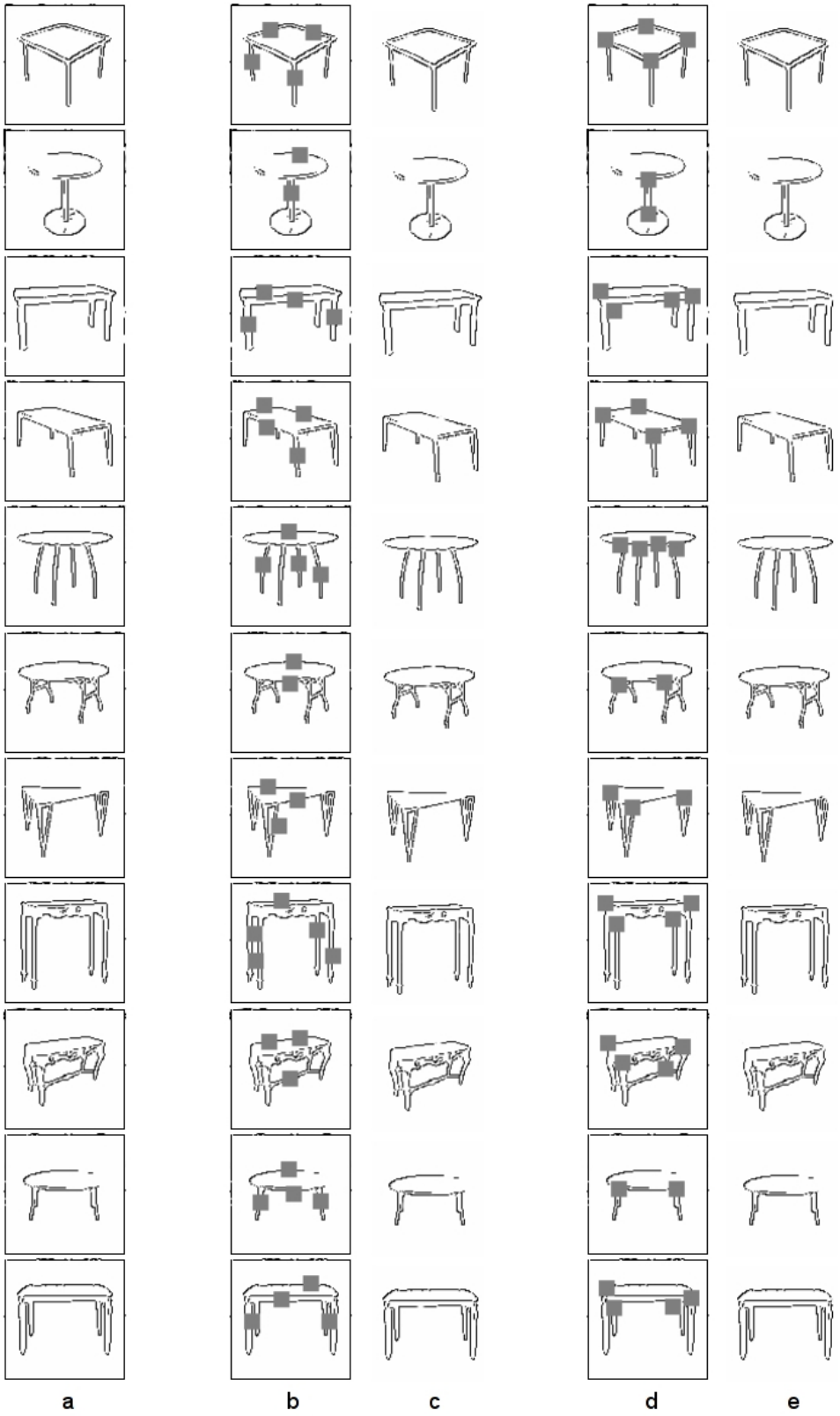
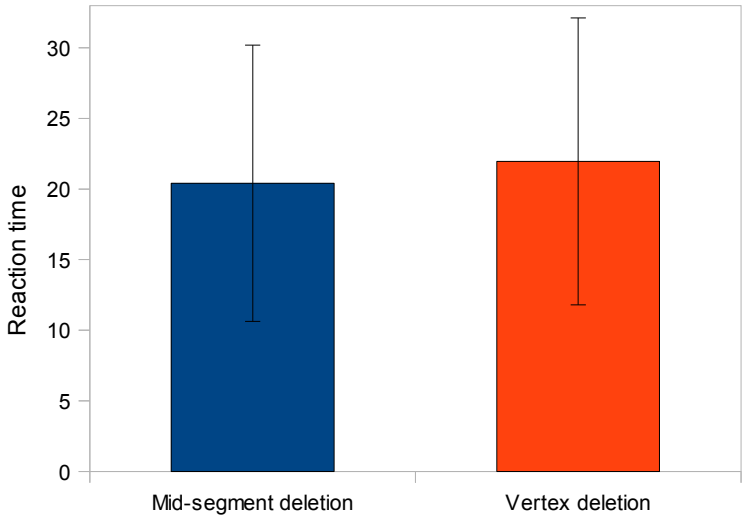


Figure 25: Experiment 7: Results for 8 hetero-associations

Columns *c* and *e* in Figure 25 present the quasi-stable states that we identified as the moment of recognition according to the procedure we described in a previous experiment. Recognition performance seems to be pretty good (compare with column *a*). For the training process we used the Widrow-Hoff learning algorithm with a stopping criterion being a vector similarity (cosine) of 0.9999 or 100 iterations, whichever comes first. Each network has 8 hetero-associations with the adjacent networks. The parameters of the dynamics are all fixed to α=0.9, γ=0.2, and δ=1 and we let the whole system iterate for 100 time steps. The processing time for the whole experiment, using 16 CPUs, was 18 min. Graph 14 presents the results of the reaction time for the two conditions (mid-segment: μ=20.41, σ=9.79, vertices: μ=21.95, σ=10.16). Although the values are not directly comparable to the ones in Graph 13 they seem to be on the same line. Deletion of vertices appears to incur a bit longer of a reaction time similar to what happens with human subjects.



Graph 14: Mean reaction time (8 associations)

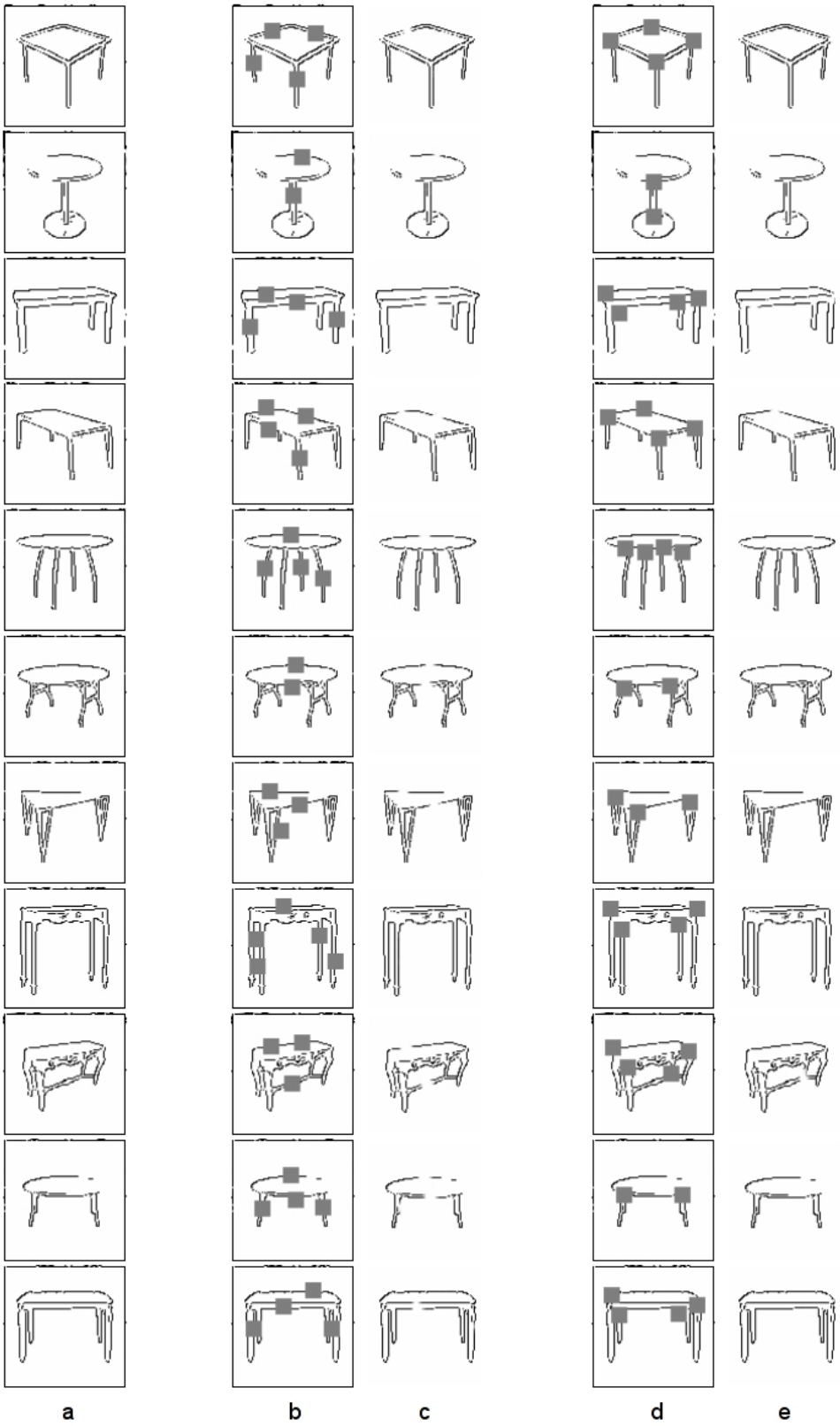
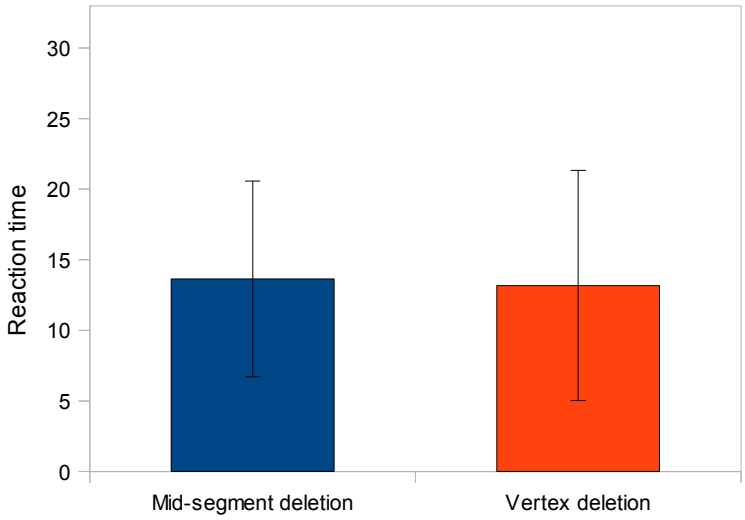


Figure 26: Experiment 7: Results for 16 hetero-associations

However, it's not certain whether this is due to the stimulus itself or the specifics of the associations. For instance, in some testing stimuli (e.g. 6th, 8th, 10th, 11th) reaction time is faster for vertices instead of mid-segments. In order to examine the effect of connectivity we will have to run a second test with a different number of hetero-associations. Like we did before, we're doubling the associations from 8 to 16 and we keep everything else the same. Figure 26 depicts the results in the same arrangement as before. The recognition performance seems overall pretty similar. The reaction time, however, has shifted a bit to the reverse (Graph 15).

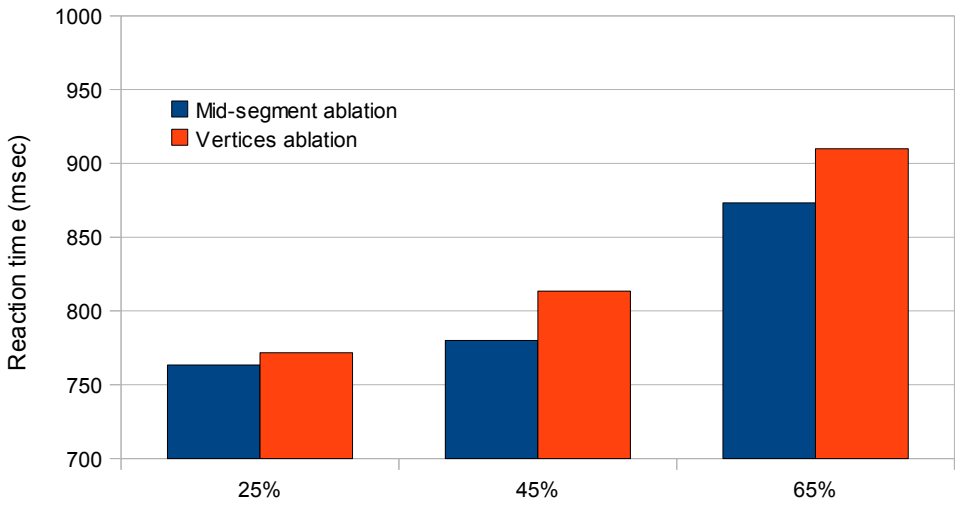


Graph 15: Mean reaction time (16 associations)

The mean reaction time for mid-segment ablations is now 13.64 (σ=6.93) and the respective for vertices is 13.18 (σ=8.15). The general trend is still on the same line with Graph 13 but the reversal indicates that the number of associations has a greater role to play. It seems that the increase of the hetero-associations provided for a richer set of connections that areas with a high degree of complexity (e.g. vertices) could take advantage of.

5.6.3. Experiment 8

In the last experiment of this study we want to investigate the interplay of the variables we tested in the previous two. We will run a parametric analysis on the size of the ablation with respect to the two conditions, mid-segment and vertices deletion. The goal is to examine whether, and to what extend, the trends we observed in the previous experiments hold for larger ablations as well. Experimental evidence with human subjects showed that reaction time for mid-segment versus vertices deletion is not significantly affected by the size of the ablation when recognition is successful (Biederman 1987).



Graph 16: Mean reaction time of humans

Graph 16 depicts the mean reaction times (msec) for three degrees of contour deletion (25%, 45%, 65%) for the two conditions we discuss. The RTs range from 760 to 910 ms and two things are clear. First, the larger the deletion the longer it takes to recognize an object, and second, recognition of objects with deleted mid-segments is slightly faster regardless of the size of the deletion. So, we will

test a similar situation with the proposed framework and see how the results are related to these.

For the training stimuli we will use the same set of 11 images that we used before (Figure 8). As for the testing stimuli we would have to design six sets (for the three degrees of ablation and the two conditions). From a cognitive point of view the percentages of deletion in Biederman's study are not very realistic since they only count the contour of the image and neglect the rest of the visual information (e.g. background). In our case, as we've seen before, we ablate whole image blocks so we would have to choose smaller percentages of ablation (e.g. 15%, 30%, 50%) in order to have a plausible comparison. For the first degree of ablation (15%) we will reuse the testing stimuli and the results in Figure 26 (columns *b/c* and *d/e*). The testing stimuli in this case have an ablation of roughly 15% of the image size. As for the other two degrees of ablation (30% and 50%) Figure 27 depicts the testing stimuli and the respective results of recognition.

For the training process we used the Widrow-Hoff learning algorithm with a stopping criterion being a vector similarity (cosine) of 0.9999 or 100 iterations, whichever comes first. Each network has 16 hetero-associations with a Gaussian-like distribution. The parameters of the dynamics are all fixed to α=0.9, γ=0.2, and δ=1 and we let the whole system iterate for 100 time steps. The processing time for this subset of the experiment was 15 min using 64 CPUs. Regarding the recognition performance, the system exhibits a better accuracy in

the case of ablated vertices. In a pairwise comparison this seems to be the case for most stimuli. This is also verified by the reaction times.

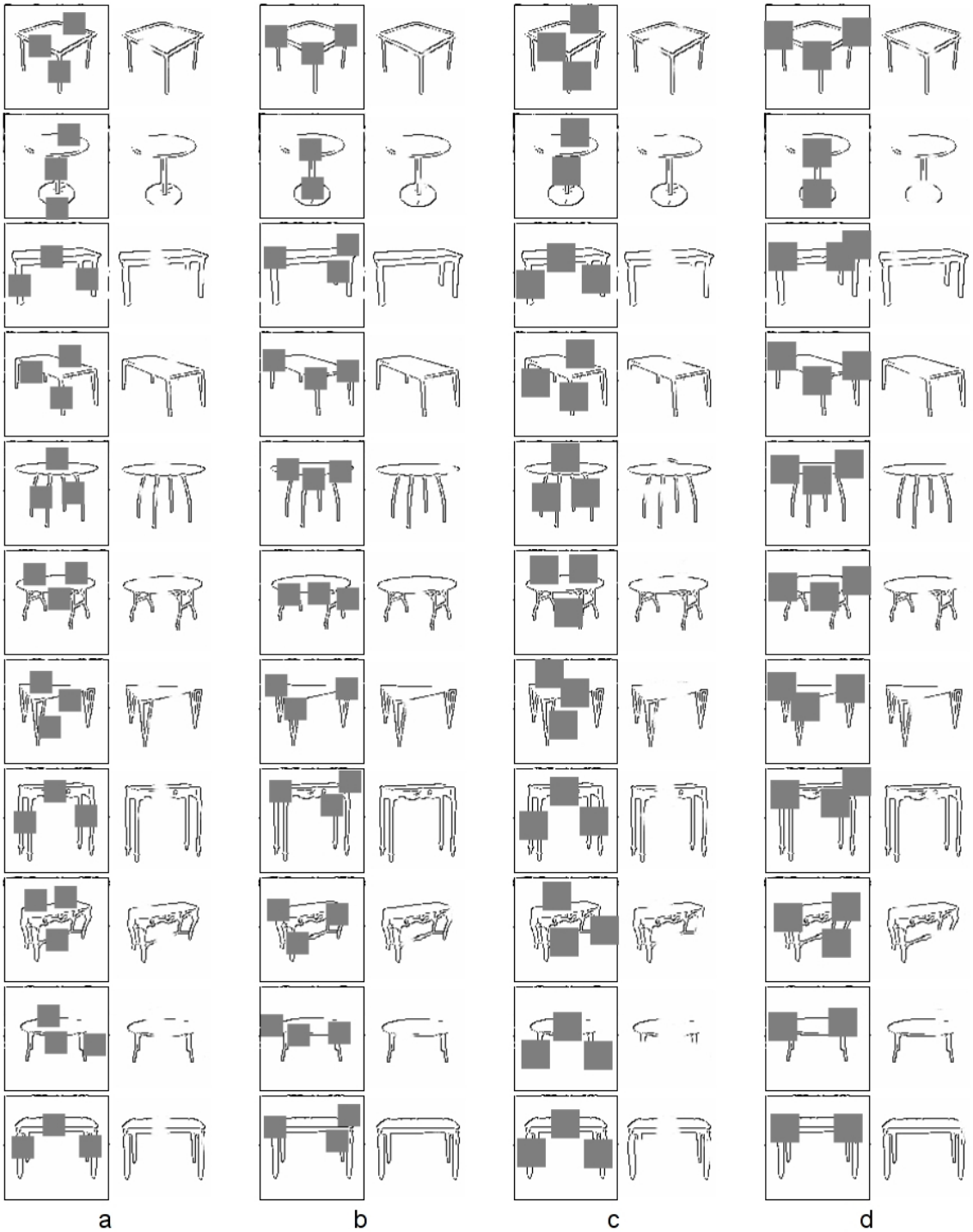
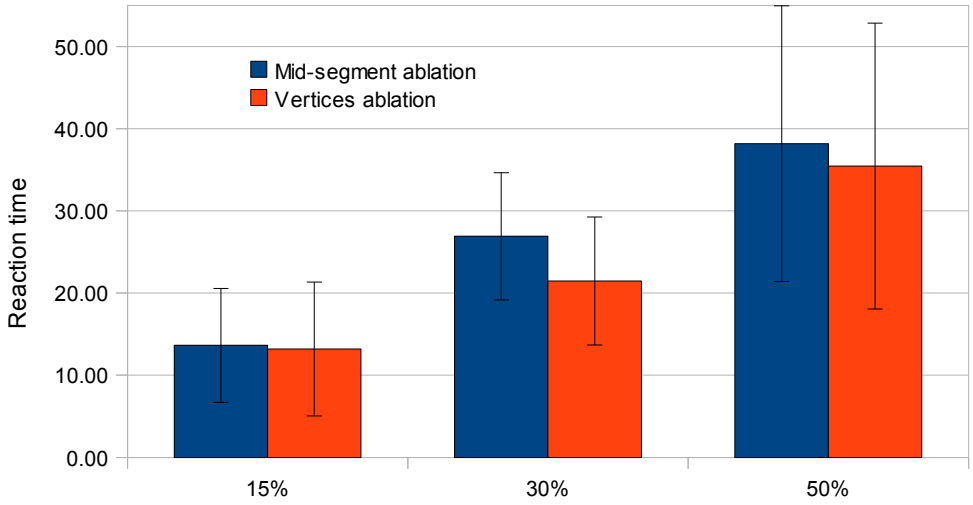


Figure 27: Experiment 8: Testing stimuli versus the resulted recognition

Graph 17 depicts the mean reaction time for the 11 stimuli of each testing set. Although, for the degree of ablation, it follows the general trend of Graph 16, for the relation of the two conditions, we see a reversal. It's the ablation of vertices that seems to be favored over the ablation of mid-segments as we increase the size of the ablation in the stimuli.



Graph 17: Parametric analysis of the degree and locus of ablation

This is more evident in the case of 30%. In the latter case (50%) there is quite a lot of metastability in the system and the preference for ablated vertices is minimal.

5.6.4. Discussion

The results of this study verified the behavioral findings that visual complexity facilitates recognition. The more complex or complete an object is the less time it takes to recognize it. These findings contradict the common computational intuition, and some cognitive proposals too (e.g. Ullman 1984), that support the idea that a complex or richer stimulus needs more time to be processed. This

may be surprising from a serial processing point of view, but it makes perfect sense if we think of visual recognition as a parallel distributed process. To illustrate this better let us take the argument to the limit. Suppose the testing image is identical to one of the training stimuli. In this case the proposed framework would take a single step of dynamics to converge while a serial approach would have to examine the whole image in a number of cycles commensurate with the size of the stimulus.

Another conclusion that came out of this study is that the degree of connectivity can have an impact on the reaction time. Experiments 7 and 8 showed that if we increase the number of hetero-associations beyond a point, the recognition of stimuli with ablated vertices becomes a little faster than the recognition of stimuli with ablated mid-segments. This may sound a bit counterintuitive but it seems as if the increase in the connectivity is wasted when there is no complexity in the underlying stimulus (e.g. mid-segments). By contrast, the regions with higher complexity (e.g. vertices) take more advantage of the extra connections and this leads to a shorter reaction time during visual recognition. Apparently, the distribution of these connections has a crucial role to play as well. We didn't have the time to experiment with it, but we would expect that an agnostic distribution (e.g. uniform or Gaussian) would be less effective in this case. Unfortunately, all these questions regarding the reversal between the two conditions and the role of the degree and the distribution of the connectivity, are not easy to test with behavioral experiments. Perhaps, future research in neuroscience might be able to shed more light and provide some answers.

On a different note, the results of this study showed that recognition performance, and reaction time in particular, do not depend on the total percentage of the ablation in the image but on the size of the largest ablated part only. For example, if an image has two ablations and the largest one is 30x30 points, the reaction time of recognition will be the same regardless of the size of the smaller ablations. The number of the ablations in the image doesn't play a role either, to some reasonable extend of course. This finding supports the idea of a parallel distributed processing and makes the following prediction. Recognition progresses as a parallel wave that fills in the ablated part from the surroundings. Hence, the reaction time is a function of the distance that the neural waves have to travel before they meet in the center of the ablation. To test this prediction we can modify and repeat Biederman's experiments as follows. Instead of examining the reaction time as a function of the total ablation in the image, we can test it as a function of the shape of the ablation. We suspect that the same size of ablation will cause quite different reaction times if we change the shape considerably. For instance, a long narrow strip will be much easier to reconstruct than a square area of the same size. If this prediction is verified it has quite interesting implications for the modularity and the connectivity of the visual architecture.

6. Conclusions

If we look back to the history of cognitive science we will see that most approaches to modeling have emanated from ideas and theories that originated in different fields (e.g. AI, logic, statistics, etc). Despite our constant claim that the discoveries in cognitive science will be the ones to drive the evolution of computational intelligence, it seems that so far we mostly observe the exact opposite trend. It is usually some computational tools that are driving cognitive models (e.g. formal logic for language, AI for decision making and problem solving, statistics for learning, etc.). Although these tools are very sound and convenient there is no doubt that brain is far more complex and versatile to be modeled with these tools only. We believe it's time for cognitive science to stop relying on mere tools and start developing a more complete theory of the mind. It doesn't matter what approach and set of tools we will employ in this endeavor, as long as they are powerful enough to account for all aspects of the complex behavior that the mind exhibits. In this thesis we took a tiny step towards this direction and tried to articulate a theory that works at the neuronal level of description. We proposed a modular computational framework that introduces a coherent integration of the spatial and temporal domains and provides explanations of how learning and dynamics may interact during processing. We believe it is versatile enough to account for the multifarious processing that brain performs. The results are promising but there is still a long way to go.

The most fundamental question for the approach we take is probably what are the limitations of a purely associative system. Many studies, including the current

one, demonstrate that this kind of systems are quite powerful for modeling perception. Extending such a system to account for perception-action interactions is not hard to imagine either. From a computational point of view there is not much of a difference between afferent and efferent nerve signals. So, an associative system should in principle be able to learn any type of coupling either perceptual, motor, or sensory-motor. When it comes to cognition and language, though, things get trickier. What happens with things we've never experienced before? Is it possible for novel concepts to emerge out of a network of associations? If novelty is simply the result of some form of integration then the answer is probably yes. But if novelty demands something more procedural, then the question becomes how feasible is for associativity to give rise to complicated mental processes. The proposed framework in its current form cannot provide such a function. However, if we turn the spatial interactions into temporal ones we can convert the network of dynamical systems into a causal network. It is our belief that with a multiregional multilayer architecture spatial and temporal interactions can coexist to produce all kinds of mental processes.

On another note, perceptual encoding is a crucial but often overlooked determinant of visual recognition. Most models put an emphasis on the processing of the information and pay less or no attention at all to the encoding of the stimulus itself. The reason why this is so important is because the properties of the encoding can change completely the limitations of the recognition. For instance, a set of stimuli encoded in a certain way may be extremely hard to classify but an alternative encoding may make the recognition much easier. A typical example of this situation is the XOR problem. If we think of

it as a problem in two dimensions then we need a nonlinear system to classify the patterns. But if we think of it as a problem in 3 dimensions then all we need is a simple linear classifier. This makes much more sense in visual recognition where most of the systems take the two dimensions for granted. If we were able to discover the visual properties (e.g. depth, motion, etc.) that optimize the perceptual encoding we might have been able to simplify considerably the recognition models we build. And this optimization is not only about the richness of the stimulus. It has also to do with the statistical properties and the distribution of the perceptual encoding. In the proposed system, for example, the capacity of the learning depends on the orthogonality of the learned vector patterns. The more orthogonal the encoded stimuli are the larger the number that the system can learn. But it also affects the dynamics. The less orthogonal the perceptual encoding is the closer the attractors are to each other, and the higher the chances for the system to converge to the wrong basin. Apparently, an evenly distributed encoding would facilitate considerably the recognition performance. In general, visual recognition could benefit a lot from a perceptual encoding that is not as blind as we tend to assume but it takes advantage of the features and the statistics of the visual world. Considering the superior recognition performance of the primate visual system it is very likely that the brain applies some sort of optimization in the perceptual encoding.

As a future direction of this project, the first thing that needs to be implemented is the expansion of the framework to the proposed multilayer hierarchical architecture. Although the experimentation with a single layer was able to give us a very good overview of the framework's potential, the full version with the

vertical integration will be able to answer many more interesting questions and achieve an even better recognition performance. One thing, for instance, that became apparent during the experiments is that there are certain limitations in the horizontal integration within a single layer. The larger the number of the hetero-associative connections the more susceptible the nodes are to neighboring activity. But this is not always good, especially when performance is not optimal and there are many fluctuations of metastability in the nodes. So, if we want the recognition to have a more *global view* we face the following trade-off: the more connections we add the more unstable the system becomes. From a cortical organization point of view the solution to this dilemma comes from vertical integration. Cortical columns have dense local connectivity and sparse distant connections. Yet they still achieve the desired *global view* by forming larger and larger assemblies as they ascend the cortical hierarchy. By implementing the proposed vertical integration we will avoid instabilities while being able to maintain the local scope within layers (small number of hetero-associations per node) and a more global scope through higher layers. This, for instance, would improve considerably the position and scale invariance in the second study.

In this thesis we proposed a novel and unconventional approach to visual recognition and to cognitive modeling in general. We hope it provided an interesting source of brainstorming.

7. Bibliography

Ackley D., Hinton G. and Sejnowski T., "A Learning Algorithm for Boltzmann Machines", Cognitive Science, 9 (1), 147-169, 1985

Amit D.J., "Modeling brain function: The world of attractor neural networks", Cambridge University Press, 1989

Yali Amit and Donald Geman, "A Computational Model for Visual Selection", Neural Computation, 11 (7), 1691-1715, 1999

James Anderson, "The BSB Model: A simple nonlinear autoassociative neural network", In Associative Neural Memories, Hassoun M. (eds.), Oxford University Press, 1993

James Anderson and Jeffrey Sutton, "A network of networks: Computation and neurobiology", World Congress of Neural Networks, 1, 561-568, 1995

Anderson C. and Van Essen D., "Shifter circuits: a computational strategy for dynamic aspects of visual processing", Proceedings of the National Academy of Sciences, 84 (17), 6297-6301, 1987

James A. Anderson, Paul Allopenna, Gerald S. Guralnik, David Sheinberg, John A. Santini, Jr., Socrates Dimitriadis, Benjamin B. Machta, and Brian T. Merritt, "Programming a Parallel Computer: The Ersatz Brain Project", In Challenges to Computational Intelligence, Wlodzislaw Duch and Jacek Mandziuk (eds.), Springer: Berlin, 2007

Moshe Bar, "Viewpoint Dependency in Visual Object Recognition Does Not Necessarily Imply Viewer-Centered Representation", Journal of Cognitive Neuroscience, 13 (6), 793-799, 2001

Mark F. Bear, Barry W. Connors, and Michael A. Paradiso, "Neuroscience: Exploring the brain", Lippincott Williams & Wilkins, 2001

Irving Biederman, "Recognition-by-Components: A Theory of Human Image Understanding", Psychological Review, 94 (2), 115-147, 1987

Irving Biederman and Eric E. Cooper, "Evidence for complete translational and reflectional invariance invisual object priming", Perception, 20 (5), 585-593, 1991

Irving Biederman and Eric E. Cooper, "Size invariance in visual object priming", Journal of ExperimentalPsychology: Human Perception and Performance, 18 (1), 1992

Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training", COLT: Proceedings of the Workshop on Computational Learning Theory, 1998

Valentino Braitenberg and Almut Schuz, "Cortex: statistics and geometry of neuronal connectivity", Springer, 1998

Bulthoff H. and Edelman S., "Psychophysical support for a two-dimensional view interpolation theory of object recognition", Proceedings of the National Academy of Sciences, 89, 60-64, 1992

Edward M. Callaway, "Feedforward, feedback and inhibitory connections in primate visual cortex", Neural Networks, 17 (5-6), 625-632, 2004

John Canny, "A Computational Approach To Edge Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, 8 (6), 679-698, 1986

John G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles", Vision Research, 20 (10), 847-856, 1980

Gustavo Deco and Edmund T. Rolls, "A Neurodynamical cortical model of visual attention and invariant object recognition", Vision Research, 44 (6), 621-642, 2004

Socrates Dimitriadis, "Cortically-inspired parallel processing", 13th SIAM conference on Parallel Processing for Scientific Computing, 2008

Socrates Dimitriadis, "Visual processing as a large scale integration of associative neural networks", 4th Computational Cognitive Neuroscience Conference, 2009

Socrates Dimitriadis and James Anderson, "A network of networks model of pop-out phenomena in visual attention", 2nd Annual Computational Cognitive Neuroscience Conference, 2006

Rodney J. Douglas and Kevan A.C. Martin, "Neuronal circuits of the neocortex", Annual Review of Neuroscience, 27, 419-451, 2004

Shimon Edelman, "Representing 3D objects by sets of activities of receptive fields", Biological Cybernetics, 70, 37-45, 1993

Shimon Edelman, "Receptive fields for Vision: from Hyperacuity to Object Recognition", In Vision, R. J. Watt (eds.), MIT Press, 1996

Daniel J. Felleman and David C. Van Essen, "Distributed Hierarchical Processing in the Primate Cerebral Cortex ", Cerebral Cortex, 1 (1), 1-47, 1991

David J. Field, "What is the goal of sensory coding?", Neural Computation, 6 (4), 559-601, 1994

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

Kunihiko Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition", Neural Networks, 1 (2), 119-130, 1988

Kunihiko Fukushima, "Restoring partly occluded patterns: a neural network model", Neural Networks, 18 (1), 33-43, 2005

J. M. Fuster, "Inferotemporal units in selective visual attention and short-term memory", Journal of Neurophysiology, 64 (3), 681-697, 1990

Stuart Geman and Elie Bienenstock, "Neural Networks and the Bias/Variance Dilemma", Neural Computation, 4, 1-58, 1992

Donald O. Hebb, "The organization of behavior", Wiley, New York, 1949

Hinton, G. E. and Sejnowski, T. J., "Optimal Perceptual Inference", Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 1983

Hinton G., Dayan P., Frey B., and Neal R., "The wake-sleep algorithm for unsupervised neural networks", Science, 268 (5214), 1158-1161, 1995

Hinton, G. E, Osindero, S., and Teh, Y. W., "A fast learning algorithm for deep belief nets", Neural Computation, 18 (7), 1527-1554, 2006

John Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proceedings of the National Academy of Sciences, 79, 2554-2558, 1982

Hubel D. and Wiesel T. N., "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex", Journal of Physiology, 160 (1), 106–154, 1962

Hubel D. and Wiesel T. N., "Sequence regularity and geometry of orientation columns in the monkey striate cortex", Journal of Comparative Neurology, 158, 267-294, 1974

Hummel J. and Biederman I., "Dynamic binding in a neural network for shape recognition", Psychological Review, 99, 480-517, 1992

Ito, M., H. Tamura, I. Fujita, and K. Tanaka, "Size and position invariance of neural responses inmonkey inferotemporal cortex", Journal of Neurophysiology, 73 (1), 218-226, 1995

Xiong Jiang, Ezra Rosen, Thomas Zeffiro, John VanMeter, Volker Blanz, and Maximilian Riesenhuber, "Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques", Neuron, 50, 159-172, 2006

J. A. Scott Kelso and Emmanuelle Tognoli, "Toward a Complementary Neuroscience: Metastable Coordination Dynamics of the Brain", In Neurodynamics of Cognition and Consciousness, (eds.), Springer Berlin / Heidelberg, 2007

Kohonen Teuvo, "Self-organized formation of topologically correct feature maps", Biological Cybernetics, 43, 59-69, 1982

Tai Sing Lee, "Image Representation Using 2D Gabor Wavelets", IEEE Transactions on Pattern Analysis and Machine Intelligence, 18 (10), 959-971, 1996

Nikos K. Logothetis and David L. Sheinberg, "Visual object recognition", Annual Review of Neuroscience, 19, 577-621, 1996

Logothetis N., Pauls J., Bulthoff H. and Poggio T., "View-dependent object recognition by monkeys", Current Biology, 4, 401-414, 1994

Nikos K. Logothetis, Jon Pauls and Tomaso Poggio, "Shape representation in the inferior temporal cortex of monkeys", Current Biology, 5 (5), 552-563, 1995

S. Marcelja, "Mathematical description of the responses of simple cortical cells", Journal of the Optical Society of America, 70 (11), 1297-1300, 1980

Henry Markram, "The Blue Brain project", Nature reviews Neuroscience, 7 (2), 153-160, 2006

Henry Markram, Joachim Lubke, Michael Frotscher and Bert Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs", Science, 275 (5297), 213-215, 1997

David Marr, "Vision. A Computational Investigation into the Human Representation and Processing of Visual Information", W.H. Freeman and Company, 1982

D. Marr and H. K. Nishihara , "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes", Proceedings of the Royal Society of London. Series B, Biological Sciences, 200 (1140), 269-294, 1978

Bartlett W. Mel, "SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition", Neural Computation, 9 (4), 777-804, 1997

L. Mioche and W. Singer , "Chronic recordings from single sites of kitten striate cortex during experience-dependent modifications of receptive-field properties", Journal of Neurophysiology, 62 (1), 185-197, 1989

Vernon B. Mountcastle, "Introduction to special issue on cortical columns", Cerebral Cortex, 13 (1), 2-4, 2003

Mountcastle VB, Berman AL, Davies PW, "Topographic organization and modality representation in first somatic area of cat’s cerebral cortex by method of single unit analysis", American Journal of Physiology, 183 (646), 1955

David Mumford, "On the computational architecture of the neocortex. I. The role of the thalamo-cortical loop", Biological Cybernetics, 65 (2), 135-145, 1991

David Mumford, "On the computational architecture of the neocortex. II. The role of cortico-cortical loops", Biological Cybernetics, 66 (3), 241-251, 1992

Randal C. Nelson and Andrea Selinger, "A Cubist Approach to Object Recognition", Sixth International Conference on Computer Vision, 1998

Jorn Niessing and Rainer W. Friedrich, "Olfactory pattern classification by discreteneuronal network states", Nature, 465, 47-52, 2010

Bruno A. Olshausen and David J. Field, "How Close Are We to Understanding V1?", Neural Computation, 17, 1665-1699, 2005

Olshausen, B., Anderson, C., and Van Essen, D., "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information", Journal of Neuroscience, 13, 4700-4719, 1993

Lawrence M. Parsons, "Imagined spatial transformations of one's hands and feet", Cognitive Psychology, 19 (2), 178-241, 1987

D. I. Perrett and M. W. Oram, "Neurophysiology of shape processing", Image and Vision Computing, 11 (6), 317-333, 1993

Poggio T. and Edelman S., "A network that learns to recognize 3D objects", Nature, 343, 263-266, 1990

Uri Polat and Dov Sagi, "The architecture of perceptual spatial interactions", Vision Research, 34 (1), 73-78, 1994

Marc'Aurelio Ranzato, Fu-Jie Huang, Y-Lan Boureau, Yann LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition", CVPR 2007, 2007

Rajesh P. N. Rao and Dana H. Ballard, "Dynamic Model of Visual Recognition Predicts Neural Response Properties in the Visual Cortex", Neural Computation, 9 (4), 721-763, 1997

Maximilian Riesenhuber and Tomaso Poggio, "Hierarchical models of object recognition in cortex", Nature Neuroscience, 2 (11), 1019-1025, 1999

Maximilian Riesenhuber and Tomaso Poggio, "Models of object recognition", Nature Neuroscience, 3 (sup), 1199-1204, 2000

Riesenhuber M. and Poggio T., "Computational Models of Object Recognition in Cortex: A Review", AI Memo 1695/CBCL Paper 190, 2000

Edmund T. Rolls and Martin J. Tovee, "Processing Speed in the Cerebral Cortex and the Neurophysiology of Visual Masking", Proceedings: Biological Sciences, 257 (1348), 9-15, 1994

Rosenblatt Frank, "The perceptron: A probabilistic model for information storage and organization in the brain", Psychological Review, 65 (6), 386-408, 1958

Bruno Rossion and Isabel Gauthier, "How does the brain process upright and inverted faces?", Behavioral and Cognitive Neuroscience Reviews, 1 (1), 63-75, 2002

Serre T., Kreiman G., Kouh M., Cadieu C., Knoblich U. and Poggio T., "A quantitative theory of immediate visual recognition", Progress in Brain Research, 165, 33-56, 2007

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio, "Robust Object Recognition with Cortex-Like Mechanisms", IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (3), 411-426, 2007

R. M. Shapley and D. J. Tolhurst, "Edge detectors in human vision", Journal of Physiology, 229 (1), 165–183, 1973

Roger N. Shepard and Jacqueline Metzler, "Mental Rotation of Three-Dimensional Objects", Science, 171 (3972), 701-703, 1971

Sperry, R. W., "Chemoaffinity in the orderly growth of nerve fiber patterns and connections", Proceedings of the National Academy of Sciences, USA, 50, 703-710, 1963

Michael Spivey, "The continuity of mind", Oxford university press, 2007

Michael J. Spivey and Rick Dale, "Continuous Dynamics in Real-Time Cognition", Current Directions in Psychological Science, 15 (5), 207 - 211, 2006

Keiji Tanaka, "Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities", Cerebral Cortex, 13, 90-99, 2003

Tarr Michael, "Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects", Psychonomics Bulletin and Review, 2, 55-82, 1995

Tarr Michael, "News on views: pandemonium revisited", Nature Neuroscience, 2, 932-935, 1999

Tarr M. and Bulthoff H., "Image-based object recognition in man, monkey and machine", Cognition, 67, 1-20, 1998

Tarr M., Williams P., Hayward W. and Gauthier I., "Three-dimensional object recognition is viewpoint-dependent", Nature Neuroscience, 1, 275-277, 1998

Alex M. Thomson and A. Peter Bannister, "Interlaminar Connections in the Neocortex", Cerebral Cortex, 13 (1), 5-14, 2003

Thorpe S., Fize D. and Marlot C., "Speed of processing in the human visual system", Nature, 381, 520-522, 1996

M. J. Tovee, E. T. Rolls, A. Treves and R. P. Bellis, "Information encoding and the responses of single neurons in the primate temporal visual cortex", Journal of Neurophysiology, 70 (2), 640-654, 1993

Tovee M. J., E. T. Rolls, and P. Azzopardi, "Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque", Journal of Neurophysiology, 72 (3), 1049-1060, 1994

Treisman, A. M. and Gelade, G., "A feature integration theory of attention", Cognitive Psychology, 12, 97-136, 1980

Alessandro Treves, "Mean-field analysis of neuronal spike dynamics", Network: Computation in Neural Systems, 4 (3), 259-284, 1993

John K. Tsotsos, "Analyzing vision at the complexity level", Behavioral and Brain Sciences, 13, 423-469, 1990

Matthew Turk and Alex Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, 3 (1), 71-86, 1991

Shimon Ullman, "Visual routines", Cognition, 18 (1-3), 97-159, 1984

Shimon Ullman, "Object recognition and segmentation by a fragment-based hierarchy", Trends in Cognitive Sciences, 11 (2), 58-64, 2006

Shimon Ullman and Erez Sali, "Object classification using a fragment-based representation", Proc. Lecture Notes in Computer Science, 2000

Shimon Ullman, Michel Vidal-Naquet and Erez Sali, "Visual features of intermediate complexity and their use in classification", Nature Neuroscience, 5, 682-687, 2002

Tim Valentine, "Upside-down faces: A review of the effect of inversion upon face", British Journal of Psychology, 79 (4), 471-491, 1988

Guy Wallis and Heinrich Bulthoff, "Learning to recognize objects", Trends in Cognitive Sciences, 3 (1), 22-31, 1999

Guy Wallis and Edmund T. Rolls, "Invariant face and object recognition in the visual system", Progress in Neurobiology, 51 (2), 167-194, 1997

Ruye Wang, "A hybrid learning network for shift-invariant recognition", Neural Networks, 14 (8), 1061-1073, 2001

Widrow Bernard and Hoff Marcian E., "Adaptive switching circuits", IRE WESCON Convention Record, 4, 96-104, 1960

Tom J. Wills, Colin Lever, Francesca Cacucci, Neil Burgess, John O'Keefe, "Attractor Dynamics in the Hippocampal Representation of the Local Environment", Science, 308 (5723), 873-876, 2005

Laurenz Wiskott, "How Does Our Visual System AchieveShift and Size Invariance?", In Problems in Systems Neuroscience, J. L. van Hemmen and T. J. Sejnowski (eds.), Oxford University Press, 2004

Jeremy M. Wolfe and Todd S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?", Nature Reviews Neuroscience, 5, 495-501, 2004

Li I. Zhang, Huizhong W. Tao, Christine E. Holt, William A. Harris and Mu-ming Poo, "A critical window for cooperation and competition among developing retinotectal synapses", Nature, 395 (6697), 37-44, 1998