UNIVERSITY OF CALIFORNIA, SAN DIEGO

Towards More Biologically-Plausible Computational Models for Cognition, Texture Classification, and Network Replication

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Rupert Charles James Minnett

Committee in charge:

Professor Robert Hecht-Nielsen, Chair
Professor Clark Guest, Co-Chair
Professor Gert Cauwenberghs
Professor William Hodgkiss
Professor Virginia de Sa

2012

The dissertation of Rupert Charles James Minnett is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2012

DEDICATION

To my wonderfully patient and supportive wife, Tricia,
and our delightfully determined son, Owen.

EPIGRAPH

I think that it is a relatively good approximation to truth - which is much too complicated to allow anything but approximations - that mathematical ideas originate in empirics. But, once they are conceived, the subject begins to live a peculiar life of its own and is ... governed by almost entirely aesthetical motivations. In other words, at a great distance from its empirical source, or after much "abstract" inbreeding, a mathematical subject is in danger of degeneration. Whenever this stage is reached the only remedy seems to me to be the rejuvenating return to the source: the reinjection of more or less directly empirical ideas.

John von Neumann, *The Mathematician*, 1947

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I express my sincere gratitude to all of you who have helped me throughout my educational career. First and foremost, I thank my graduate advisor and mentor, Professor Robert Hecht-Nielsen, without whom this dissertation would not have been written. You inspired me, with your enthusiasm for the future of computational neuroscience and your passionate lectures in your Neurocomputing class, to pursue research in our exciting interdisciplinary field. I will always be grateful for the many diverse and insightful conversations we have had. You have made many outstanding contributions to academia and industry and your breadth and depth of knowledge continues to astound me. I look forward to, hopefully, many more productive years of collaboration. I also appreciatively thank my committee co-chair, Professor Clark Guest, and committee members, Professors Gert Cauwenberghs, William Hodgkiss, and Virginia de Sa, for your guidance, inspiration, and unwavering support of my academic career.

I am extremely grateful to my colleagues at the Scripps Institution of Oceanography (SIO) and the College of Earth, Ocean, and Atmospheric Sciences (CEOAS) at Oregon State University, all of whom have been exceptionally supportive and understanding of my, perhaps overly ambitious, choice to simultaneously pursue graduate degrees and continue full-time employment at SIO. Anthony, you have been, and continue to be, consistently encouraging and accommodating and I am eternally grateful for your guidance, enthusiasm, friendship, and demand for perfection. Cathy, Lisa, and Hubert, thank you ever so much for your mentorship in my feeble attempts to understand paleomagnetism, and for your patience and friendship, especially during the last year. You are all exceptional academics and have made the last six years extremely enjoyable and rewarding.

Soren, your vision, determination, and focus are inspiring and contagious. I truly appreciate the many coffee breaks, extensive white-board discussions, and impressive neuroanatomy conversations you have shared with me. You led the way for my graduate school career and I can only hope to continue to follow in your footsteps.

Andrew, my graduate school kin, you have greatly enriched the hundreds of meetings we have attended together with your wit, endless knowledge and wisdom. Life in the lab would not have been the same without you and I look forward to many more years of post-graduate friendship.

Mehrdad, Geoff, Gavin and Bill, thank you for invigorating the lab with your new ideas and momentum, for the many colorful discussions and meetings, and for taking over some of the administrivia of running the lab; your humor will serve you well. Hopefully the gameless game nights will continue.

Arthur and Alyssa, you are wonderful friends. Arthur, I consider myself very lucky to have gone our separate ways after meeting at Cornell and still to have shared graduate school together. I look back very fondly on the dozens, if not hundreds, of squash matches we have played, luckily without any black eyes and only a few scrapes.

Wayne, my best man, thank you for keeping life light-hearted and reminding me to live. You will have to try living on the west coast, eventually, but not until your new company goes public.

Chris and Michelle, your friendship has meant a great deal to me and Tricia and we always enjoy your company. Thank you both for being so supportive of us and, Chris, for enriching my life for well over a decade.

Thank you to all of my graduate and undergraduate schoolmates who made the doctorate program pass by a little faster. Omer, Bharath, Ramsin and Matt, thank you for all of the stimulating conversations and your company over lunches and through many classes at UCSD. Omer and Bharath, I enjoyed the many squash matches we played, even the many I lost. Derek, you are a fantastic friend and made undergraduate school and your visits during graduate school very memorable.

Dr. Steven Browdy, thank you ever so much for your friendship and guidance, which led me to graduate school and to a career at SIO. Professor Stephen Murrell, your mentorship for two years over a decade ago taught me invaluable lessons and good practices in programming and academia.

I am very thankful for the love and support of my family through many of the challenges these last six years have introduced. Mum and Dad, I am extremely grateful to you both for the wonderful parents you have been. Dad, you have been an inspiration to me throughout my life and I hope that I can have as positive of an influence on Owen as you have had on me. Mum, you have always cared for me and you instilled in me a determination to succeed academically at a very young age. Edward, my brother and my oldest friend, you have always been interested and encouraging and very supportive of my life. Beth, Raul, and Joe, you adopted me into your family with open arms and raised the most important person in my life.

Most importantly, I must thank my wife and my best friend, Tricia. My life truly began in Italy when I first met you over a decade ago. My happiest moments have been shared with you, and my worst have been without you. My education, my career, and most of all, our son, Owen, are a direct reflection of your unconditional love and support, for which I am eternally grateful. This dissertation could not have been written without you.

Chapter 2, in full, is a modified reprint with minor revisions of the material as it appears in Physics of Life Reviews. Solari, Soren; Smith, Andrew T.; Minnett, Rupert C.J.; Hecht-Nielsen, Robert, Elsevier, 2008. The dissertation author was an investigator and author of this material.

Chapter 3, in part, has been submitted for publication of the material. Minnett, Rupert C.J.; Hecht-Nielsen, Robert. The dissertation author was the primary investigator and author of this material.

Chapter 4, in part, has been submitted for publication of the material. Minnett, Rupert C.J.; Hecht-Nielsen, Robert. The dissertation author was the primary investigator and author of this material.

Chapter 5, in full, is a modified reprint with minor revisions of the material as it appears in Neural Networks. Minnett, Rupert C.J.; Smith, Andrew T.; Lennon Jr., William C.; Hecht-Nielsen, Robert, Elsevier, 2011. The dissertation author was the primary investigator and author of this material.

VITA

2000	Who's Who in American High School Students, Miami, Florida
2004	Most Outstanding Computer Engineering Student, University of Florida
2005	Bachelor of Science (Summa Cum Laude) in Electrical Engineering, University of Florida
2005	Bachelor of Science in Computer Engineering, University of Florida
2000-2005	Programmer Consultant, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Florida
2008	Master of Science in Electrical Engineering(Signal and Image Processing), University of California, San Diego
2009	Candidate of Philosophy in Electrical Engineering (Signal and Image Processing), University of California, San Diego
2005-2012	Programmer/Analyst II, Scripps Institution of Oceanography, University of California, San Diego
2012	Doctor of Philosophy in Electrical Engineering (Signal and Image Processing), University of California, San Diego

PUBLICATIONS

Rupert C.J. Minnett, Robert Hecht-Nielsen, "Human and Computational Texture Classification Performance", *Optical Engineering*, In Review, 2012.

Rupert C.J. Minnett, Andrew T. Smith, William C. Lennon Jr., Robert Hecht-Nielsen, "Neural network tomography: Network replication from output surface geometry", *Neural Networks*, Volume 24, Issue 5, June 2011, Pages 484-492, ISSN 0893-6080, DOI: 10.1016/j.neunet., 11.01.006.

William C. Lennon Jr., **Rupert C.J. Minnett**, Andrew T. Smith, Robert Hecht-Nielsen, "Direct positioning of a neural network's hidden units", *Society for Neuroscience*, Volume 40, November 2010.

Anthony A.P. Koppers, **Rupert C.J. Minnett**, Lisa Tauxe, Cathy Constable,"MagIC database: Comprehensive archiving and visualization of rock- and paleomagnetic data using web 2.0 technology", *Geochimica Cosmochimica Acta*, Volume 74, Issue 12, Supplement 1, June 2010, Page A531.

Anthony A.P. Koppers, Hubert Staudigel, **Rupert C.J. Minnett**, "Seamount Catalog: Seamount Morphology, Maps, and Data Files", *Oceanography*, Volume 23, Issue 1, March 2010, Page 37.

Soren Solari, Andrew T. Smith, **Rupert C.J. Minnett**, Robert Hecht-Nielsen, “Confabulation Theory”, *Physics of Life Reviews*, Volume 5, Issue 2, June 2008, Pages 106-120, ISSN 1571-0645, DOI: 10.1016/j.plrev.2008.03.003.

TALKS, REPORTS, AND ABSTRACTS

Rupert C.J. Minnett, Anthony A.P. Koppers, Lisa Tauxe, Cathy Constable, Nicholas A. Jarboe, “Solving the challenges of data preprocessing, uploading, archiving, retrieval, analysis and visualization for large heterogeneous paleo- and rock magnetic datasets”, *Eos, Transactions, American Geophysical Union*, Volume 92, Issue 52, December 2011, Abstract IN13B-1330.

Lisa Tauxe, Cathy Constable, Anthony A.P.iKoppers, **Rupert C.J. Minnett**, Nicholas A. Jarboe, Fabio Donadini, Andy Biggin, “Paleointensity: Where we stand and where the MagIC database can take us”, *Eos, Transactions, American Geophysical Union*, Volume 92, Issue 52, December 2011, Abstract GP14A-03 (invited talk).

Cathy Constable, **Rupert C.J. Minnett**, Anthony A.P. Koppers, Lisa Tauxe, Nicholas A. Jarboe, "The MagIC Online Database: Improving the Archive Quality via a New Review System”, *Eos, Transactions, American Geophysical Union*, Volume 92, Issue 52, December 2011, Abstract GP11A-1008.

Nicholas A. Jarboe, Lisa Tauxe, Cathy Constable, Anthony A.P. Koppers, **Rupert C.J. Minnett**, “Transitional Field Studies Using a Large Scale Database (MagIC)”, *Eos, Transactions, American Geophysical Union*, Volume 92, Issue 52, December 2011, Abstract GP23A-1035.

Rupert C.J. Minnett, Anthony A.P. Koppers, Lisa Tauxe, Cathy Constable, “Recent Advances in the MagIC Online Database: Rock- and Paleomagnetic Data Archiving, Analysis, and Visualization”, *Eos, Transactions, American Geophysical Union*, Volume 91, Issue 52, December 2010, Abstract GP11A-0750.

Rupert C.J. Minnett, Anthony A.P. Koppers, Lisa Tauxe, Cathy Constable, “The MagIC Database: A Rock- and Paleomagnetic Online Comprehensive Archive, Scientific Analysis Environment, and Rich Visualization Platform”, *Eos, Transactions, American Geophysical Union*, Volume 90, Issue 52, December 2009, Abstract GP11A-0738.

Anthony A.P. Koppers, **Rupert C.J. Minnett**, Lisa Tauxe, Cathy Constable, Fabio Donadini, “Managing Rock and Paleomagnetic Data Flow with the MagIC Database: from Measurement and Analysis to Comprehensive Archive and Visualization”, *Eos, Transactions, American Geophysical Union*, Volume 89, Issue 53, December 2008, Abstract GP11A-0708.

Rupert C.J. Minnett, Anthony A.P. Koppers, Daniel T. Staudigel, Hubert Staudigel, “Efficiently Communicating Rich Heterogeneous Geospatial Data from the FeMO2008 Dive Cruise with FlashMap on EarthRef.org”, *Eos, Transactions, American Geophysical Union*, Volume 89, Issue 53, December 2008, Abstract IN41A-1121.

Daniel T. Staudigel, **Rupert C.J. Minnett**, Anthony A.P. Koppers, Hubert Staudigel, “FlashMap: A Versatile and Intuitive Web-based User Interface for Rich Heterogeneous Geospatial Data”, *Eos, Transactions, American Geophysical Union*, Volume 88, Issue 52, December 2007, Abstract IN32A-04 (talk).

Anthony A.P. Koppers, **Rupert C.J. Minnett**, Lisa Tauxe, Cathy Constable, Fabio Donadini, “Uploading, Searching and Visualizing of Paleomagnetic and Rock Magnetic Data in the Online MagIC Database”, *Eos, Transactions, American Geophysical Union*, Volume 88, Issue 52, December 2007, Abstract GP21A-0114.

Cathy Constable, Anthony A.P. Koppers, Lisa Tauxe, **Rupert C.J. Minnett**, “Cyberinfrastructure for the Magnetics Information Consortium (MagIC)”, *Geological Society of America Abstracts with Programs*, Volume 39, Issue 6, October 2007, Page 211, Paper Number 78-10 (talk).

Cathy Constable, Anthony A.P. Koppers, Lisa Tauxe, **Rupert C.e. Minnett**, “The Five Dimensions of MagIC”, *Eos, Transactions, American Geophysical Union*, Volume 87, Issue 52, December 2006, Abstract IN13C-1172 (invited).

Rupert C.J. Minnett, Anthony A.P. Koppers, Cathy Constable, Lisa Tauxe, Sergei A. Pisarevsky, Michael J. Jackson, Peter Solheid, Sudipto Banerjee, Catherine Johnson, “The Magnetics Decormation Consortium (MagIC) Online Database: Uploading, Searching and Visualizing Paleomagnetic and Rock Magnetic Data”, *Eos, Transactions, American Geophysical Union*, Volume 87, Issue 52, December 2006, Abstract GP11A-0067.

Daniel T. Staudigel, **Rupert C.J. Minnett**, Anthony A.P. Koppers, Hubert Staudigel, Jasper Konter, “Google for Seamounts”, *Eos, Transactions, American Geophysical Union*, Volume 87, Issue 52, December 2006, Abstract V13A-0636.

Hubert Staudigel, Anthony A.P. Koppers, **Rupert C.J. Minnett**, Daniel T. Staudigel, Jasper Konter, Paul Martin, “Google for Seamounts”, *Eos, Transactions, American Geophysical Union*, Volume 87, Issue 52, December 2006, Abstract V13A-0635.

Edward J. Kearns, Steven Browdy, **Rupert C.J. Minnett**, Christine Caruso-Magee, Geoffrey K. Morrison, Rod G. Zika, “Marine Observations from the International SeaKeepers Society’s Autonomous VOS Fleet”, *JCOMM Technical Report Number 16, WMO-TD/No. 1118*, Goa, India, February 2002, Pages 90-97.

ABSTRACT OF THE DISSERTATION

Towards More Biologically-Plausible Computational Models for Cognition, Texture Classification, and Network Replication

by

Rupert Charles James Minnett

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2012

Professor Robert Hecht-Nielsen, Chair
Professor Clark Guest, Co-Chair

Neuroscience and machine learning often operate at two ends of a spectrum. The former sometimes finds itself entrenched in the details of experimentation, and the latter sometimes finds itself drifting into the expanse of theory. Both fields can mutually coexist, and when they do, have produced invaluable results in computational neuroscience towards more plausible models of biological solutions. This dissertation presents two detailed investigations into the benefits of this interdisciplinary field: a model for cognition and a model for vision. Experiments during these investigations led us to a third result: a new learning approach called neural network tomography.

We introduce our universal theory of cognition, Confabulation Theory, and discuss its biological plausibility. Confabulation Theory posits that the cerebral cortex, in conjunction with the thalamus, is implementing a repeated functional architecture of thalamocortical modules, each encoding one attribute which an object in the individual's mental universe may possess. These modules are interconnected with concurrence statistics called knowledge links, are capable of confabulating a state, and are carefully controlled with action commands. We use Confabulation Theory to build a model for natural language processing and present striking results in sentence generation with context.

Subsequently, we focus on the task of texture classification, which we argue is a more primitive operation than object recognition, and therefore, appropriate for investigation with the goal of elucidating biology's solution for processing visual stimuli. We develop a hierarchical model for texture classification, carefully informed by neuroscience results, and demonstrate state-of-the-art performance on a challenging texture classification dataset in the context of our human psychophysical experiment.

Finally, we survey existing methods in neural network learning and propose a new approach with several valuable theoretical advantages. By rephrasing the task of function approximation as replicating the topology and weights of an existing universal approximator network, we show that several of the drawbacks of classical backpropagation learning can be avoided. We define a new objective function, mean squared curvature (MSC), and demonstrate that minimizing the MSC of the difference between the networks during the replication process produces favorable results and allows networks to be reverse-engineered iteratively.

Chapter 1

Introduction

Researching the mammalian anatomy alone, even in great detail, does not elucidate the intricate mechanisms that interact to allow us to process information, otherwise Constantin von Economo would have understood how the cerebral cortex works in 1925 with his cytoarchitectonic study of the adult human cerebral cortex [von Economo & Koskinas, 1925; von Economo, 1929; von Economo & Koskinas, 2008]. Researching the physiology alone (e.g. mammalian vision, see Chapter 3) only provides glimpses at the outputs of the functional contributions made by different stages of the visual processing stream black box, otherwise the emergence of orientation selectivity in primary visual cortex would no longer be a highly controversial topic today, fifty years after Hubel and Wiesel proposed a solution from their physiological studies of receptive fields [Martinez, 2011; Hubel & Wiesel, 1962]. A thorough understanding of the physiology from carefully controlled experiments [Crick, 1979; Siegel & Callaway, 2004], constrained by monumental cytoarchitectonic studies of detailed neuroanatomy, is necessary to truly uncover the mysteries of mammalian cortical information processing. Concurrently modeling these integrated systems can often inform where the experimental neuroscientists should concentrate to elucidate the complexities of the physiology (e.g. the primary visual cortex simple cell parameterization by Daugman [1985] informing the physiological experiments by Jones et al. [1987]).

1.1 Comparative Neuroanatomy

One of the most informative sources for neurophysiology is comparative neuroanatomy. Two of the most comprehensive comparative neuroanatomy studies were published over a century ago and, surprisingly, have yet to be reproduced with more modern neuroscience techniques. Brodmann studied the cytoarchitectonics across 64 different mammalian species with Nissl stains and produced his now widely accepted magnum opus, a cortical map labeled with 52 discrete areas [Brodmann, 1909], which was subsequently, but less popularly, extended to 200 regions [Vogt & Vogt, 1919] and followed by revised nomenclature [von Economo & Koskinas, 1925]. Ramón y Cajal published his cytoarchitectonic study comparing human and vertebrate neuroanatomy using Golgi staining at the end of the 19th century [Ramón Y Cajal, 1899]. Careful observations of the subtle, and sometimes stark, differences in the neuroanatomy between species in the context of those species' behaviors can provide extremely valuable insights into the function and importance of neurological structures. For example, only large-brained mammals have unusually large neurons in layer 5 of primary visual cortex with thick myelinated axons which are thought to be implicated in rapid motion detection of the magnocellular pathway [Wang et al., 2008]. Without these specialized neurons, conduction delays over the longer distances in large-brained mammals would likely interfere with the time-sensitivity of visual motion perception. This comparative neuroanatomy result and many others provide valuable evidence for understanding the neurophysiology of the mammalian brain.

Comparative neuroanatomy can also provide clues about which underlying neurophysiological results are less critical to the integrated system neuroscience function. Hubel and Wiesel reported on the surprisingly organized structure of orientation columns and ocular dominance columns in the cat striate cortex [Hubel & Wiesel, 1974] and for decades this property of columnar organization was thought to be a universal characteristic of mammalian cortical architecture and necessary for high visual acuity. However, rodents and lagomorphs do not share this property of semiregular, smoothly-varying orientation maps in primary visual cortex and appear not to suffer from any related visual deficits [Van Hooser et al., 2005]. Perhaps, orientation column organization is not a necessary component of the visual processing stream mechanism and may only be a solution, when necessary in most mammals, to the retinotopically limited arborization of isotropic lateral connectivity in primary visual cortex.

1.2 Lesion Studies

Another compelling set of sources for clues about visual neurophysiology are lesion studies. Humans, due to unfortunate circumstances, sometimes suffer from accidental (e.g. head trauma, infection) or necessary (e.g. treatment of epilepsy or tumors) cortical lesions or severe atrophy. These circumstances, though, have repeatedly provided crucial insights into the physiology of the affected cortical areas from the observed behavioral deficits the subjects experience. For example, a very recent functional MRI (fMRI) study [Cavina-Pratesi et al., 2010] involving two humans suffering from severe visual agnosia convincingly identified, through behavioral double dissociation and neuroimaging dissociation, two areas that are independently implicated in shape recognition (not object recognition) and surface texture recognition. One subject had damage to his collateral sulcus (CoS) from infectious encephalitis and could not perform surface texture recognition (healthy subjects showed activity in the posterior CoS during this task), but could perform shape recognition. The other subject suffered bilateral lesions to her lateral occipital cortex (LOC) from a hypoxic episode and could not perform shape recognition (healthy subjects showed activity in the LOC during this task), but could perform surface texture recognition.

1.3 Experimental and Model Complexity

In an effort to perfectly replicate mammalian vision in a computational model, one might be tempted to model every neuron in the biological model, or every synapse or even every ligand [Markram, 2006]. However, this rapidly becomes computationally intractable, overparameterized, and largely unnecessary. In classical mechanics, many problems are drastically simplified with a minimal compromise in uncertainty (e.g. modeling the trajectory of a ball as a particle traveling through a vacuum and neglecting aerodynamics). Similarly, there exists a minimal degree of complexity at which mammalian vision or cognition can be modeled without any substantial loss in explained variance. Determining that minimum level of complexity, though, is not trivial.

This dissertation presents three related investigations, carefully informed by modern neuroscience evidence, towards determining this minimum level of complexity in different modalities. First, we present a framework for computationally modeling cognition and apply it to the task of sentence generation with context. Then, we develop a model for computational vision, apply it to the task of texture classification, and compare it with our measured human performance. Finally, we introduce a new neural network learning method and demonstrate its advantages, even in trial cases, over classical backpropagation.

1.4 Chapter References

Brodmann, K. (1909). *Brodmann's "Localisation in the Cerebral Cortex"*. Smith-Gordon, London.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010). Separate processing of texture and form in the ventral stream: evidence from fmri and visual agnosia. *Cerebral Cortex*, 20(2), 433–446.

Crick, F. H. (1979). Thinking about the brain. *Scientific American*, 241(3), 219–232.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A Optics and image science*, 2(7), 1160–1169.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154.2.

Hubel, D. H. & Wiesel, T. N. (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *Journal of Comparative Neurology*, 158(3), 267–293.

Jones, J. P., Stepnoski, A., & Palmer, L. A. (1987). The two-dimensional spectral structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6), 1233–1258.

Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, 7(2), 153–60.

Martinez, L. M. (2011). A new angle on the role of feedfoward inputs in the generation of orientation selectivity in primary visual cortex. *The Journal of Physiology*, 589(Pt 12), 2921–2922.

Ramón Y Cajal, S. (1899). Estudios sobre la corteza cerebral humana i. corteza visual. *Revista Trimestral Micrográfica*, 4.

Siegel, R. M. & Callaway, E. M. (2004). Francis Crick's legacy for neuroscience: Between the α and the Ω. *PLoS Biology*, 2(12), e419.

Van Hooser, S. D., Heimel, J. A., Chung, S., Nelson, S. B., & Toth, L. J. (2005). Orientation selectivity without orientation maps in visual cortex of a highly visual mammal. *Journal of Neuroscience*, 25(1), 19–28.

Vogt, C. & Vogt, O. (1919). Allgemeinere ergebnisse unserer hirnforschung. vierte mitteilung. die physiologischebedeutung der architektonischen rindenfelderung auf grund neuer rindenreizungen. *Journal für Psychologie und Neurologie*, 25, 399–462.

von Economo, C. (1929). The cytoarchitectonics of the human cerebral cortex. *Journal of Anatomy*, 63(Pt 3), 389.

von Economo, C. & Koskinas, G. N. (1925). *Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen*. Springer Verlag, Vienna.

von Economo, C. & Koskinas, G. N. (2008). *Atlas of Cytoarchitectonics of the Adult Human Cerebral Cortex* (translated, revised and edited by L.C. Triarhou). Karger, Basel.

Wang, S. S.-H., Shultz, J. R., Burish, M. J., Harrison, K. H., Hof, P. R., Towns, L. C., Wagers, M. W., & Wyatt, K. D. (2008). Functional trade-offs in white matter axonal scaling. *Journal of Neuroscience*, 28(15), 4047–4056.

Chapter 2

Cognition

Here, we briefly describe our Confabulation Theory and discuss experimental results in its support. Simply put, Confabulation Theory proposes that thinking is like moving. In humans, the theory postulates that there are roughly 4000 *thalamocortical modules*, the "muscles of thought". Each module performs an internal competition (*confabulation*) between its *symbols*, influenced by inputs delivered via learned axonal associations with symbols in other modules. In each module, this competition is controlled, as in an individual muscle, by a single graded (i.e., analog) *thought control signal.* The final result of this confabulation process is a single active symbol, the expression of which also results in launching of *action commands* that trigger and control subsequent movements and/or thought processes. Modules are manipulated in groups under coordinated, event-contingent control, in a similar manner to our 700 muscles. Confabulation Theory hypothesizes that the control of thinking is a direct evolutionary outgrowth of the control of movement.

2.1 Introduction

The formal academic study of human and animal cognition has been underway for over 2360 years [Finger, 2001] (e.g., Aristotle's pioneering studies of logical thought [Barnes, 1984] were published c. 350 BC). Yet, even today, roughly all that can be stated with certainty is that there is strong evidence suggesting that the storage and processing of information involved in all aspects of human *cognition* (seeing, hearing, planning, language, reasoning, control of movement and thought, etc.) is carried out by the cerebral cortex and its related subcortical nuclei. Beyond general statements of this sort (primarily based on deficits after cortical lesions [Catani & Ffytche, 2005; Penfield & Rasmussen, 1968]), little is known about how cognition (also referred to here as *thinking* Theory.

Section 2 provides a conceptual framework for the key elements of Confabulation Theory. Sections 2.3.1, 2.3.2, 2.3.3 and 2.3.4 detail the four key elements of Confabulation Theory. Finally, Section 2.4 briefly surveys some natural language processing experiments using computer implementations of the theory. These illustrate the enormous impact Confabulation Theory is likely to have on practical information processing. The way forward to building artificial intelligence systems is now clear, and the mysteries of the human brain are primed to be finally unlocked.

2.2 A Conceptual Framework for Confabulation Theory

This section presents a general conceptual overview of the four key elements of Confabulation Theory, the specific details of which are described in Section 2.3.

We hypothesize that cognitive information processing is a direct evolutionary re-application of the neural circuits controlling movement, and thus functions just like movement. Brains seem to have developed to utilize sensory inputs to coordinate muscle contractions [Lieber, 2002; Squire, 2004], which thereby increased the evolutionary fitness of these animals. Since neural circuitry already existed to contract muscles, the specific thalamus [Jones, 2007] and six-layered cerebral cortex in mammals [Northcutt & Kaas, 1995] developed to perform cognition utilizing the same control mechanisms as for muscles (i.e., basal ganglia, brainstem, cerebellum, etc.).

Conceptually, brains are composed of many "muscles of thought" (termed *thalamocortical modules* in mammals), each of which describes a single *attribute* which an object in the individual's mental universe may possess. A module contains *symbols*, each of which is a sparse collection of neurons which functions as a *descriptor* of the attribute of that module. For example, if the attribute of a module is the *visual form* of faces [Tsao et al., 2006], then a single symbol represents a particular face's visual form. If the attribute of the module is words (in some specific human language) describing the name of an object, then a single symbol represents a particular word.

Each thalamocortical module is connected to many other modules through the cortical *white matte*r in the brain. When two symbols are active simultaneously, each in different modules, they are said to *co-occur*, which creates the opportunity to associate the two symbols. For instance, after seeing a face and hearing a name together, the symbols representing each may become associated. These learned associations are implemented by unidirectional axonal synaptic connections between the neurons representing each of the pair of symbols. Each strengthened unidirectional association between two symbols is termed a *knowledge link*. Collectively knowledge links comprise all *cognitive knowledge*.

Each thalamocortical module performs the same single information processing operation, which can be thought of as a "contraction of a list of symbols", termed a *confabulation*. Throughout a confabulation, *input excitation* is delivered to the module through knowledge links from active symbols in other modules' lists of candidate conclusion symbols, driving the activation of these knowledge link's *target* symbols in the module performing the confabulation. When a thalamocortical module's list of candidate conclusion symbols contracts, there is no physical movement in the brain, rather symbols currently on the list compete (based upon their relative excitation levels) for eventual exclusive activation (a so-called "winner-take-all" competition) within that module and, as a result, the number of active symbols is gradually reduced. Crucially, this contraction of the candidate conclusion symbol list in each thalamocortical module is externally controlled by a *thought control signal* delivered to the module (in exact analogy with the motorneuron input delivered from an external source to a muscle).

Physical muscle contractions are controlled by graded analog inputs provided by alpha motor neurons [Lieber, 2002]. Similarly, a confabulation in a thalamocortical module is controlled by a graded analog control input, the thought control signal, which determines how much overall symbol activity there can be in the module. The thought control signal determines how many symbols remain in the competition, but has no effect on selecting which symbols are in the competition. Which symbols are in the competition is determined by the excitation level of a symbol as it dynamically reacts to knowledge link input from active symbols in other modules (which cause its excitation level to increase) or to a reduction or cessation of such input (which causes its excitation level to fall). Ultimately, the thought control signal is used to dynamically contract the number of active symbols in a module from an initial many less-active symbols to, at the end of the confabulation, a single maximally-active symbol. The resulting single active symbol is termed the confabulation *conclusion*.

Each time a module reaches a conclusion, the module immediately launches *action* *commands* (by activating a separate collection of specialized neurons within the module which have been associated from the conclusion symbol). Some action commands (which proceed from the module to subcortical nuclei) directly cause a specific movement process or thought process (each a type of behavioral *action*) to be launched. Others modify ongoing actions. The learned association between each symbol of a module and its set of action commands is termed *skill knowledge*. Skill knowledge is stored in the module, but the learning of these associations is controlled by subcortical brain nuclei.

In summary, the brain is composed of many thalamocortical modules ("muscles of thought"), which, through controlled input, expand and contract the list of active symbols in the module. The list of active symbols is determined by input from active symbols in other modules via knowledge links, thus all the modules interact dynamically, "comparing notes", while a thought control input contracts the number of active symbols in each module to a single active symbol conclusion. When a conclusion is reached in a module, those action commands which have a learned association from that conclusion symbol are instantly launched. These issued action commands are proposed as the source of all non-reflexive and non-autonomic behaviors.

Thalamocortical modules performing confabulations, delivering excitation through knowledge links, and applying skill knowledge through the issuance of action commands constitute the complete foundation of all mammalian cognition.

2.3 The Key Elements of Confabulation Theory

Confabulation Theory primarily consists of four *key elements* that form the fundamental underpinnings of all cognition. Although Confabulation Theory in its most general form likely applies to cognitive information processing in all animal nervous systems, here we focus on describing Confabulation Theory from the perspective of mammalian neuroanatomy (with specific emphasis on the human case). The dominant neuronal structures (gray matter) and gross anatomical projections (white matter) of the cerebral cortex in all mammals have a virtually identical organization [Striedter, 2005], therefore the four key elements presented here apply equivalently to all mammals, including humans. Although all of the important functions of cognition are covered by the four key elements, many ancillary details are still waiting to be elucidated.

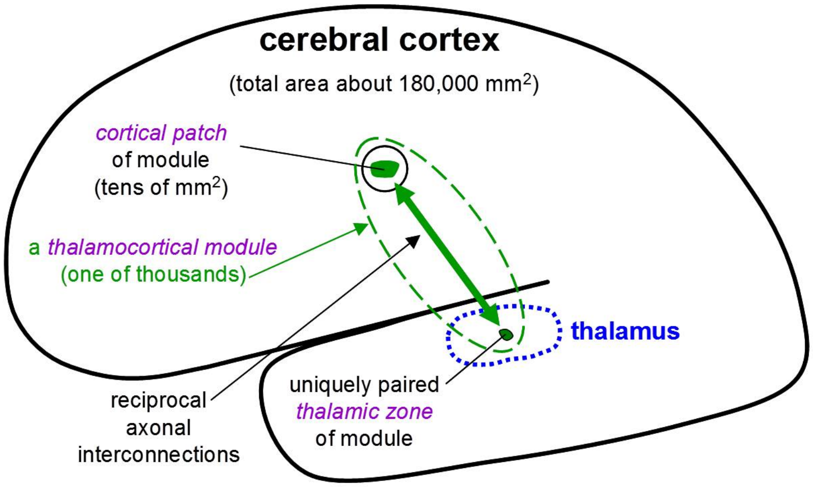

cerebral cortex
(total area about 180,000 mm^2)
cortical patch of module (tens of mm^2)
a thalamocortical module (one of thousands)
thalamus
reciprocal axonal interconnections
uniquely paired thalamic zone of module

Figure 2.1: A human *thalamocortical module* (one of thousands in human cerebral cortex).
Each thalamocortical module is composed of a localized *patch* (having an area of a few tens of square millimeters) of the six-layer cortical sheet along with a uniquely paired, reciprocal, small *zone* of specific thalamus. The cortical patch of each module is reciprocally axonally connected with the thalamic zone of the module. Although cortical patches (and thalamic zones) of different modules are largely disjoint, partial overlaps do likely occur.

2.3.1 Thalamocortical modules and symbols: describing *attributes* of *objects* in the individual's mental universe

For over a hundred and fifty years, the cerebral cortex has been known to have localized functionality, for example, vision, language, and movement are each processed in separate cortical areas [Finger, 2001; Penfield & Rasmussen, 1968]. Even though each area of the cerebral cortex carries out seemingly different types of information processing, every area of the cortex has the same 6-layered structure and equivalent reciprocal axonal connections with some part of the thalamus [Brodmann, 1909; Jones, 2007]. The similarity across all regions of cortex and thalamus strongly suggest that how information is stored and processed in each cortical area is the same even though what is stored and processed is different. Surprisingly, given the detailed knowledge of cortical organization, very little is known about exactly how or exactly what is stored and processed in any part of cortex. Confabulation Theory proposes that human cerebral cortex is divided into thousands of *thalamocortical modules*, each including a localized patch of the cortical sheet with an area on the order of tens of mm^2 (see Figures 2.1 and 2.2). Each module also includes a small zone of thalamus that is reciprocally axonally connected with its cortical patch.

Each thalamocortical module is responsible for describing one *attribute* which an *object* (e.g., a sensory object, a language object, a movement or thought process object, a plan object, etc.) of the individual's *mental universe* may possess. To carry out this job, each module develops and permanently stores a finite set of discrete *symbols*, each representing a single *descriptor* of the module's attribute. For example, if a particular module is used (along with other modules) to describe bodies of English text, it might contain over 200,000 symbols, each describing a single English word, word phrase, or punctuation ("tree", "Tree", ";", "jet fighter", "Winston Churchill", etc.). Similarly, symbols in a visual module might be used to describe the localized visual form of a portion of a visual object [Tanaka, 2003].

In humans, each module typically possesses thousands to hundreds of thousands of symbols. Each human cortical module contains hundreds of thousands to millions of neurons. However, each symbol is represented by a small collection of tens to hundreds of neurons within each of multiple populations of symbol-representing neurons within the module. Although neuron collections representing two symbols often overlap by a small number of neurons, the mathematics of neuronal processing are such that any significant interference between symbols is exceedingly unlikely [Hecht-Nielsen, 2007] (much like the probability that all of the air molecules in a room would congregate into one corner).

When a module is being used to describe a particular mental world object, the neurons representing one specific symbol within one special neuronal population of the module are all firing at a high level of activity (the exact definition of what action potential rate and what level of action potential synchrony this involves is not yet known) and all other such symbol-representing neurons are largely inactive. In this situation, we say that the module is *expressing* that one symbol. Often (e.g., during *multiconfabulations*, see Section 2.3.3 below) a module will have multiple symbols partially activated.

Only when a symbol is active is it sending excitation through knowledge link axons to other associated symbols. Because of the sparse coding of symbols, one or several symbols can be fully or partially active at the same time in a module (or none may be active). Consider that the human cerebral cortex has a module whose attribute is the visual form of faces [Tsao et al., 2006], and another whose attribute is words. Suppose a symbol representing the visual form of "Bob's" face is active in one module and a symbol representing Bob's name is active in another module; and that a knowledge link from

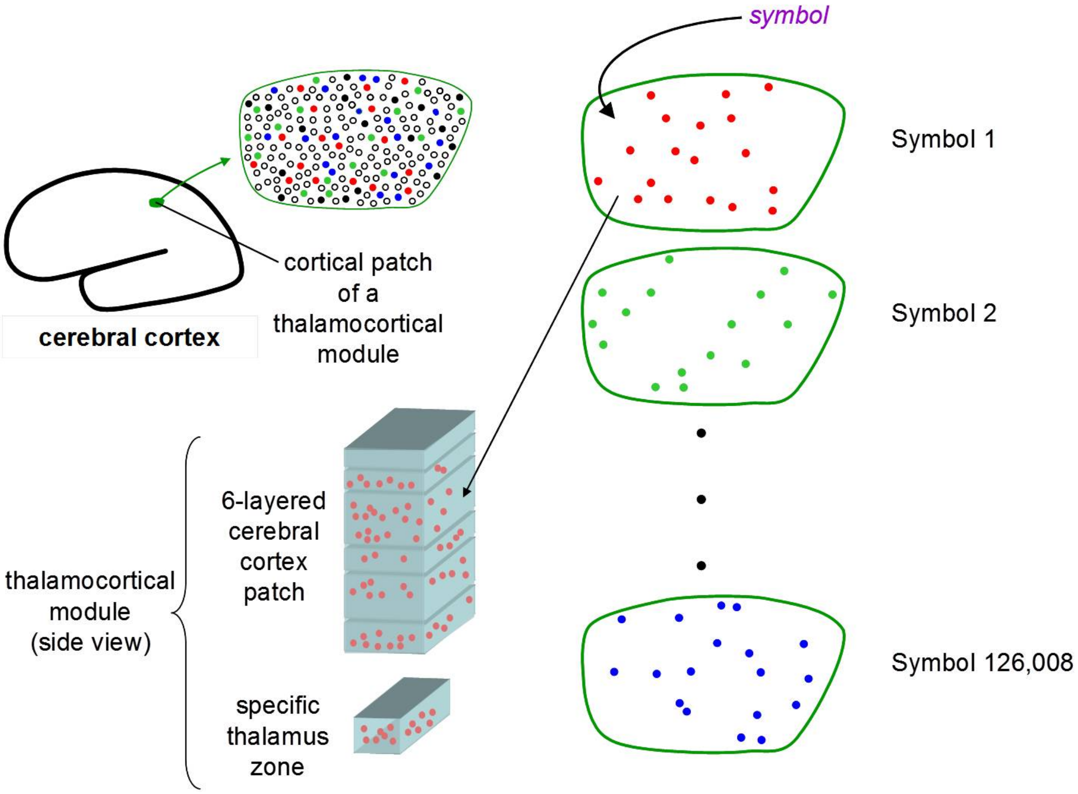

symbol
cerebral cortex
cortical patch of a thalamocortical module
Symbol 1
Symbol 2
Symbol 126,008
thalamocortical module (side view)
6-layered cerebral cortex patch
specific thalamus zone

Figure 2.2: A thalamocortical module stores and processes symbols.
A primary function of each thalamocortical module is to store and process exactly one *attribute* which an *object* (e.g., a sensory object, a visual object, a language object, a movement or thought process object, a plan object, etc.) of the individual's mental universe may possess. To carry out this function, each module develops and stores a large collection of *symbols*, which are each a different *descriptor* of the attribute of the module (e.g., a symbol may represent a particular word in a module that describes words, or a particular face in a module that describes faces). Each symbol representation is composed of many tens to hundreds of neurons (shown as colored dots within the enlarged depiction of the module's cortical patch). When a module is describing an object, typically a single symbol is active in the module. As shown, a module is composed of a 6-layered piece of cerebral cortex and small zone of specific thalamus. Within each layer different populations of neurons have different anatomical projections and, as a result, different functions. As an example, the conclusion symbol representations (shown here) likely reside in the lower part of cortical layer 3. These neurons store the symbol conclusions that are ultimately mapped to action commands in cortical layer 5. Here, a module with 126,008 symbols is depicted.

the first of these symbols to the second has already been formed. Then, if a symbol representing the visual form of Bob's face is active in the first module, excitation will be delivered to the symbol representing his name in the second module. That such symbol pair "co-activation-based" knowledge links would be sufficient to explain all aspects of cognition seems preposterous. Yet, that is exactly the claim of Confabulation Theory. And, how this works is illustrated in Section 2.4.

Anatomically, thalamocortical modules have the exact same structure in all mammals [Brodmann, 1909; Northcutt & Kaas, 1995], providing strong evidence that cognition functions identically in all mammals. Understanding the structure of a thalamocortical module, therefore, is essential to understanding how cognition functions. Within every thalamocortical module, the same distinct populations of neurons exist (roughly aligned with the six layers of the cortex and the two layers of thalamus) each having homotypical white matter projections to different regions of the brain [Barbas & Hilgetag, 2002; Braitenberg & Schüz, 1998; de No, 1943]. We call each of these separate neuron populations a *neural field.* The afferent and efferent connectivity of each neural field defines its function. Each symbol has a separate sparse neuronal representation within each neural field. Our computational models suggest the central function of this detailed anatomical organization is performing the controlled winner-take-all competition of confabulation using neurons and synapses (see Section 2.3.3).

2.3.2 Knowledge Links: the basis of all cognitive knowledge

In 1949, Donald Hebb [Hebb, 2002] postulated that *learning* in brains was the strengthening of synapses linking two groups of neurons (which he called "cell assemblies") with axonal connections between them. He postulated that this occurred whenever the first cell assembly helped cause the second cell assembly to become active (the involved synapses going in this direction between the two cell assemblies are then strengthened). Ample neurological evidence supports Hebb's postulate; however, no comprehensive examination of the role of cell assemblies in learning has yet occurred (see Figure 2.3 for an illustration of the role of Hebb's idea in Confabulation Theory). When two symbols are co-active they may become associated by strengthening the synapses linking them. The unidirectional association between two symbols is termed a *knowledge link*; a reciprocal pair of knowledge links may exist between two symbols. Each knowledge link is considered a single *item of knowledge.* An active *source* symbol delivers *input excitation* to all

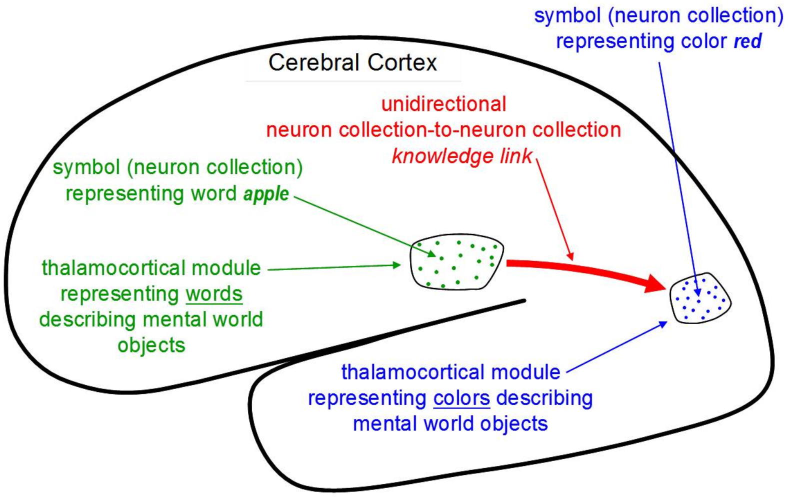


Figure 2.3: A cognitive *knowledge link*.
Here, a human subject is viewing and considering a red apple. A visual thalamocortical module contains an active symbol for the *color* of the apple. At the same time, a language thalamocortical module contains an active symbol for the English *name* of the apple. Pairs of symbols which *meaningfully co-occur* in this manner have unidirectional axonal links, termed *knowledge links* (each considered a single *item of knowledge*), established between them via synaptic strengthening. The entire axonal bundle of all unidirectional knowledge links between two modules is termed a *knowledge base*. Knowledge bases compose the vast majority of cortical white matter. Confabulation Theory predicts that knowledge links must be implemented in vast quantities for cognition to be useful, which is consistent with known neuroscience; white matter is the single largest structure in the human brain. The average adult human is postulated to possess billions of knowledge links, most of which are usually established in childhood.

target symbols to which it is connected through knowledge links, where the *strength* of a knowledge link (a quantity which varies over a limited dynamic range) determines the amount of input excitation that a target symbol receives. Therefore, a knowledge link is an 'association' between two cell assemblies, as Hebb postulated, albeit with a much higher level of neuronal complexity than he envisioned. In consonance with the pairwise *associationist* doctrine established by Aristotle and his colleagues 2360 years ago and built up further by a series of leading thinkers on human cognition over the past 500 years, Confabulation Theory contends that such knowledge links—formed on the basis of symbol pair co-occurrence—are the only type of knowledge used (or needed) in cognition.

In particular, in mammals, knowledge links are formed over two time-scales [Squire, 2004]. Instantaneous knowledge links are formed indirectly by linking each learnable pair of co-active symbols via the hippocampus, entorhinal cortex, and related portions of perirhinal/parahippocampal cortex. Over many subsequent sleep periods this indirect knowledge link may be consolidated into a direct cortico-cortical knowledge link from one symbol to another symbol (i.e., no longer through the hippocampus). This unidirectional consolidated cortico-cortical knowledge link between two symbols will typically last for decades, even if it is not used. Knowledge links that are used last for life. The collection of all unidirectional knowledge links connecting a particular source module to a particular target module is termed a *knowledge base*.

Figure 2.4 illustrates an example of some knowledge links that may have been formed by experiencing a red apple. Here, five modules are each expressing a symbol describing one attribute of the apple. In the center, the symbol representing the English name of the apple is active. Above that, the symbol representing the apple's skin texture is active. To the right, the apple's visual color is active. And to the left and at the bottom the motor chewing process for an apple and the gustatory sensation of the apple are active. When an apple is currently present in the mental world, it *is* its collection of knowledge-link-connected symbols which are currently active in many modules. There is no "binding problem" [von der Malsburg, 1981], because all of these symbols are mutually "bound" by their previously established pairwise knowledge links.

Confabulation Theory proposes that the mathematics of cognition relies on the formation, strength, and use of these knowledge links. The strength of a single knowledge link is logarithmically related to the conditional probability $p(\beta|\epsilon)$, where β represents the occurrence of source symbol β and ϵ the occurrence of the target symbol ϵ (see

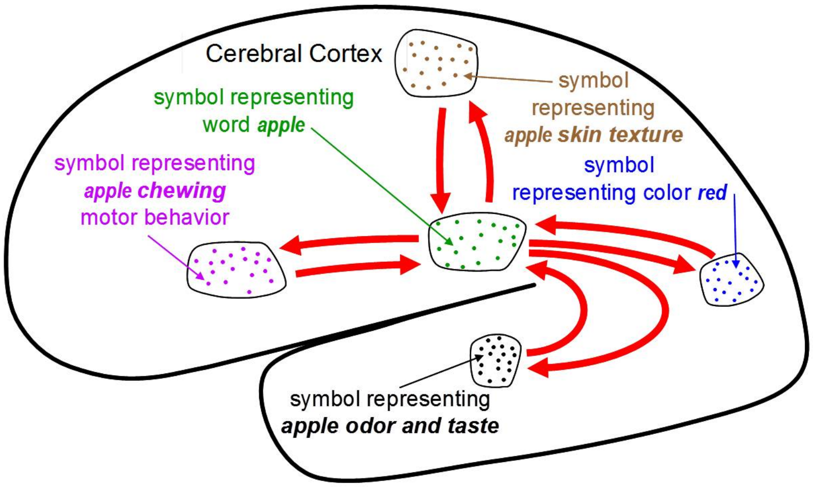

Cerebral Cortex
symbol representing word *apple*
symbol representing *apple* ***skin texture***
symbol representing ***apple chewing*** motor behavior
symbol representing color ***red***
symbol representing ***apple odor and taste***

Figure 2.4: Billions of pairs of symbols are connected via knowledge links.
The set of all knowledge links joining symbols belonging to one specific *source* module to symbols belonging to one specific *target* module is termed a *knowledge base*. In the human brain, knowledge bases take the form of huge bundles of axons termed *fascicles*, which together make up a large portion of each cerebral hemisphere's ipsilateral white matter. Each module also typically has a knowledge base to its contralateral 'twin' module (and perhaps to a few others near its twin)—which together constitute the *corpus callosum* fascicle linking the two cerebral hemispheres. Here, reciprocal knowledge links (red arrows), only some of which are shown, connect various symbols representing different attributes of an apple pairwise with each other.

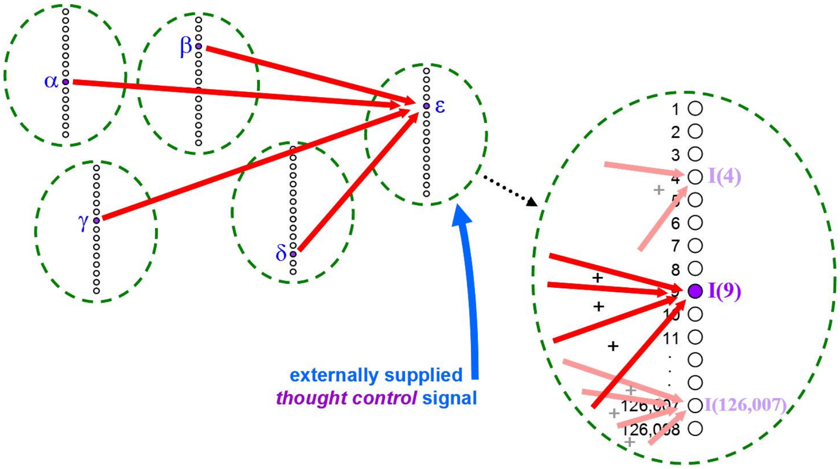


Figure 2.5: *Confabulation*—the only information-processing operation used in cognition.
Here, a concrete example involving five thalamocortical modules is shown (for simplicity, each module is illustrated as a dashed green oval with a list of that module's symbols inside it). During a confabulation, active symbols (α, β, γ, δ) in four *source* modules shown on the left send excitation through knowledge links to symbols in a fifth target module (shown on the right). Each confabulation on every module is controlled by a graded analog *thought control* signal (analogous to the motor neuron input signal that contracts a muscle). The *conclusion* of a confabulation operation will ultimately be the symbol receiving the most *input excitation* I (symbol 9 shown on the right). See text for more details.

Figure 2.5). Importantly this quantity is estimated by dividing the number of times β and ϵ co-occur by the number of total occurrences of the target symbol ϵ. Biologically, this implies that a target symbol (composed of neurons) has a relatively fixed total strength of incoming knowledge links (synapses) that it can physically support, and that the total strength of all incoming knowledge links to a single target symbol is limited. The brain, therefore, cannot form an arbitrary number of strengthened knowledge links, which explains the need to use temporary knowledge link formation and an entirely dedicated brain region (the hippocampus) to determine which knowledge links should be consolidated and become permanent. Such a simple biological constraint on neurons and the support of synapses may have enabled the exploitation of the underlying mathematics necessary for cognition.

A major question arises as to whether co-occurrence knowledge of this sort can be sufficient to account for human and animal "intelligence". Below, in Section 4, we will see that they are.

2.3.3 Confabulation: the universal basic operation of thought

The vague notion that cognition employs some sort of "information-processing" has been around for millennia. Today, the understanding of the exact nature of this "cognitive information-pro the first neuroscientist). Confabulation Theory states explicitly that cognition involves only one information-processing operation—*confabulation*: a simple controlled winner-take-all competition between symbols on the basis of their total input excitation received from knowledge links.

Figure 2.5 illustrates a confabulation. The four source modules on the left each have a single active symbol in them: α, β, γ, and δ. Each of these active source symbols delivers input excitation to many symbols (often hundreds) in the fifth, target module on the right through knowledge links. The state of the fifth module, which is about to undergo *confabulation*, is shown enlarged on the far right (red arrows depict individual knowledge links). For illustration, symbol 4 of this module is receiving two active knowledge links, whereas symbols 9 and 126,007 are receiving knowledge links from all four symbols α, β, γ, and . Each knowledge link is delivering a certain quantity of input excitation to the neurons of its target symbols. The input from the thought control signal (blue arrow) causes the module to contract, as a result the number of active target symbols decrease. If the manipulation of this thought control signal allowed only two symbols to be active, then symbols 9 and 126,007 would be active (since they have the most *input excitation*) and symbol 4 would end up being shut off by the competition, and thus made inactive. If the *thought control signal* manipulations then caused only a single symbol to be active, then symbol 9, having the most input excitation, would remain active and symbol 126,007 would be shut off, resulting in a single conclusion symbol: number 9.

The input excitations arriving at symbol k from different knowledge links are *summed* to yield the *total input excitation for symbol k*, *I(k)* (this summation is noted by the plus signs between the knowledge links in the enlarged illustration of module five). As discussed in detail by Hecht-Nielsen [2007], this *additive knowledge combination* property of thought is what enables the vast information-processing power and flexibility of human cerebral cortex. Note that knowledge links are not neuron to neuron connections, but rather symbol to symbol connections (i.e., many neurons to many neurons); therefore, many hundreds to thousands of synapses may transmit input excitation from a single source symbol to a single target symbol, enabling accurate additive combination even in the presence of large background noise or individual synaptic failure.

We emphasize that a thalamocortical module does not undergo a confabulation operation unless commanded to do so, in the same way a muscle contracts only when commanded to do so by its motorneuron input [Lieber, 2002]. Upon being commanded to contract its list (by a deliberately supplied *thought control signal*, illustrated by a blue arrow in Figure 2.5), each symbol of the fifth module competes with all others for exclusive activity. During this competition the number of active symbols being considered decreases in proportion to the thought control signal strength (thus a confabulation is a "contraction of a list of symbols"). Since the timing of this contraction is controlled, coordinating the parallel convergence of many modules to a final state may itself involve a significant amount of learning. This learned coordinated control of convergence is termed a *thought process*. Upon converging to a final conclusion, the neurons representing the symbol with the largest input intensity I (in the example of Figure 2.5, symbol 9) are highly active and all other symbol-representing neurons are not. This "winner-take-all" competition is called a *confabulation*, and the winning symbol is termed its *conclusion*.

It may seem mysterious that mere neurons can implement controlled, winner-take-all symbol competition. Within a module, connections between the neural fields in the module's cortical patch and its paired thalamic region constitute a *neuronal attractor network* [Hecht-Nielsen, 2007], the state of which evolves through cortex-thalamus-cortex oscillations and is modulated by the *thought control signal*. Each collection of neurons representing a symbol is a stable state of the attractor network. A thalamocortical module can be held constant with a single active symbol or multiple partially (or fully) active symbols by means of this cortex-thalamus-cortex oscillation. During the oscillation, additional context can be applied through knowledge links to influence the competition. In this way, modules can be made to converge slowly or quickly, and the number of active symbols at any one time can be made to grow or contract to the symbol with the greatest input excitation. In behavioral experiments, subjects can temporarily retain a finite set of sensory domain specific information, which has been termed *working memory* [Monsell, 1984]. We propose that the underlying neural mechanisms of working memory is a controlled continuous thalamocortical oscillation of a single (or possibly multiple) symbols in a single module. Working memory is the controlled cortex-thalamus-cortex oscillation maintaining symbol activation(s) in a single module. Each module can implement working memory specific to its attribute, thereby distributing working memory throughout the cortex.

Confabulation is hypothesized to be the only information-processing operation of thinking. In the Figure 2.5 example, there is only one confabulation taking place. Ordinarily, confabulations on multiple modules take place together (with the involved modules acting as source and target simultaneously), with convergence to the winning symbols slowed somewhat to allow mutual knowledge-link-mediated interaction ("comparing notes" in order to arrive at a mutually consistent *confabulation consensus* of final conclusions). In such a *multiconfabulation*, millions of relevant *items of knowledge* (i.e., knowledge links), each emanating from a viable candidate conclusion, are employed in parallel in a "swirling" convergence process. Multiconfabulation is a key mechanism enabling the enormous information-processing power and flexibility of thought [Hecht-Nielsen, 2007]. As an analogy between movement and thought, a biceps contraction is to a single confabulation, as the elegant movements of a ballerina are to multiconfabulation.

Confabulation seems quite alien in comparison to existing concepts in neuroscience, computational intelligence, neural networks, computer science, traditional AI, and philosophy. For example, computers typically follow the Turing paradigm: when commanded via a specific, digital, instruction code they execute a pre-defined mathematical instruction on specified variables. Thalamocortical modules, on the other hand, have only one information-processing "instruction"—confabulation. Further, the thought control signal delivered to the confabulating module from outside the cerebral cortex, is not digital, but analog (and often very dynamic). Yet the result of each completed module confabulation is digital: a single symbol.

A natural question arises as to where the thought control signal originates. The most likely source of the thought control signal is a small area of the thalamus (VM/VAmc) close to the mammothalamic tract, which projects diffusely to layer 1 of virtually the entire cerebral cortex [Herkenham, 1980]. Early electrophysiology experiments showed that stimulation of this thalamic area caused an immediate activation of almost the entire cerebral cortex [Hanbery & Jasper, 1953] as would be expected from a central thought control signal. Although the thalamic intralaminar nuclei had for decades largely been the focus of the layer 1 nonspecific projection [Jones, 2007], we now know that these intralaminar nuclei predominantly target layers 5 and 6 of the cortex [Herkenham, 1980; Jones, 2007]. From the Confabulation Theory perspective, the intralaminar nuclei are quite likely involved with the behavioral triggering of action commands discussed in the next section. In addition to layer 1 projections, the VM/VAmc nucleus of the thalamus also receives projections from both the basal ganglia and cerebellum (both highly involved in movement) giving this small thalamic area all the necessary axonal connections to function as the “alpha motor neurons of thought”.

2.3.4 Action commands: skill knowledge and the origin of behavior

One of the most obvious aspects of brain function (and therefore one of the most consistently ignored) is that animals typically launch many behaviors every second they are awake. Most of these are *microbehaviors* (small corrective modifications or addenda to ongoing behaviors), but typically, major new behaviors are launched many times per hour, often predicated on newly emerged events. Beyond simple reflexes (e.g., knee jerk) and autonomic reactions (e.g., digestion), no understanding of how and why behaviors neurologically originate currently exists.

Confabulation Theory proposes the “conclusion → action” principle (see Figure 2.6): every time a confabulation operation on any thalamocortical module reaches a conclusion, an associated set of *action commands* are launched from a specific set of neurons within the module. Action commands arise from a neural field within the module (probably the layer 5 pyramidal neurons) that send axons towards subcortical structures. These action commands either cause the launch of *behaviors* (movements and/or thoughts) immediately (when originating from layer 5b subcortical projections) or they cause the immediate consideration of *suggested behaviors* for further evaluation (when originating from layer 5a projections to the basal ganglia). Confabulation Theory postulates that all non-reflexive and non-autonomic behaviors arise in this manner.

The mapping between symbols and action commands represents a different type of learning product, termed *skill knowledge*, that requires rehearsal and practice. As opposed to cognitive knowledge of facts and events (stored by knowledge links), skill knowledge is not directly consciously accessible [Squire, 2004]. Skill knowledge is a learned association from the conclusion symbol neural field to the action command neural field within a thalamocortical module.

The neuroanatomical location and physiological properties of skill knowledge is very different from cognitive knowledge. First, as opposed to the module-to-module (symbol-to-symbol) nature of knowledge links, the learned mapping from symbols to action commands lies entirely within a thalamocortical module. Second, unlike a cognitive knowledge link, which may be extremely robust if consolidated over many nights of

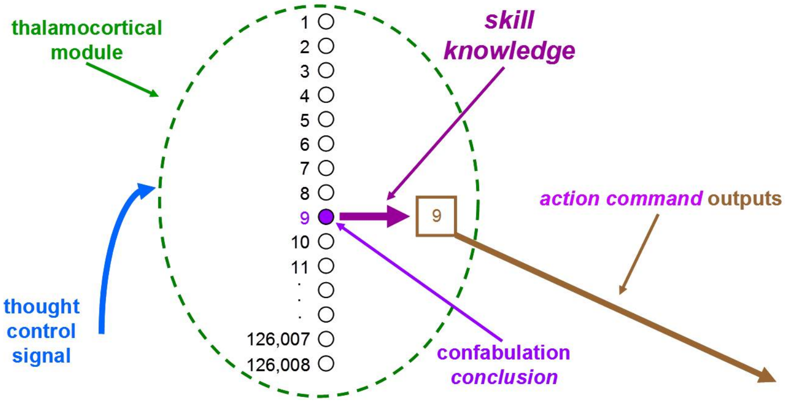


Figure 2.6: The *conclusion action principle*: hypothesized to be the origin of all non-reflexive and non-autonomic behavior.
Here, a thalamocortical module (illustrated abstractly, in consonance with Figure 2.5, as an oval containing a list of the module's symbols) has successfully completed a confabulation operation (under control of its externally supplied thought control signal) and reached a conclusion (symbol number 9, as in Figure 2.5). Whenever a module completes a confabulation and reaches a conclusion it immediately causes a set of *action command* outputs to be launched (these outputs proceed to subcortical nuclei). The action command outputs that are launched are those which have been previously associated with the conclusion symbol via a subcortically managed *skill-learning* process (distinct from cortical knowledge link learning). The "conclusion → action" principle is the fourth and last of the key elements of Confabulation Theory.

sleep, skill knowledge is often fragile and short-lived. The impermanence of skill knowledge is required for *rehearsal learning* of skills (like playing a musical instrument), where gradually more competent skill knowledge needs to supplant earlier, less perfected, skill knowledge. Finally, there are separate learning mechanisms for each of these types of knowledge. Whereas the learning of cognitive knowledge requires the hippocampus and its related medial temporal lobe, the learning of skill knowledge requires other subcortical structures such as the basal ganglia, intralaminar thalamus, and basal forebrain. This is clearly a topic richly deserving of extensive new neuroscience research.

The application of skill knowledge to the launching of action commands is not part of cognitive information processing per se (it comes into play only after each thalamocortical information processing operation has completed its job of reaching a conclusion). However, thought processes are dependent upon the *thought control sequences* coordinating confabulations in many thalamocortical modules. In the same way that movement sequences (actually, *postural goal* sequences) are learned, stored, and recalled, so are thought control sequences. These thought control sequences are controlled directly by action commands launched by thalamocortical modules. Therefore, thought (confabulation) begets action (action commands) and action begets thought in an endless cycle during wakefulness. The homunculus hiding behind a curtain pulling the control levers of the brain and body is thus exorcised.

2.4 Confabulation Theory Experiments

Confabulation Theory offers a unified approach to achieving the holy grail of Artificial Intelligence: a fully integrated intelligent system of human-level capacity.

To glimpse this potential, consider the capabilities of the simple *confabulation architecture* (see Figure 2.7). This particular architecture allows sets of three consecutive sentences from the same paragraph of a well-written newspaper story to be represented in terms of symbols. At the bottom level, each module has 63,008 symbols, representing the most common words and punctuation of English. When a sentence is entered, the symbol representing the corresponding word of the sentence is activated in each module. Words and punctuations are entered in order from left to right, one per module, and each module only has one active symbol. Modules to the right of each sentence's ending period have no active symbols. The modules of the second and third levels of the architecture have symbols representing words, word phrases, and punctuation.

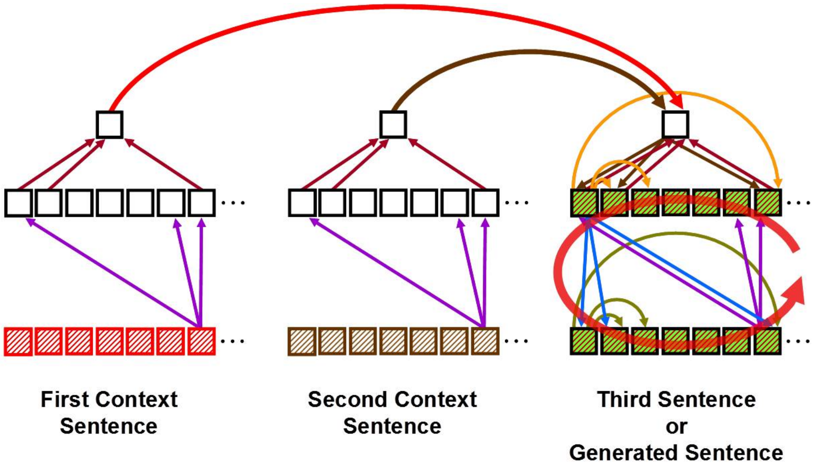


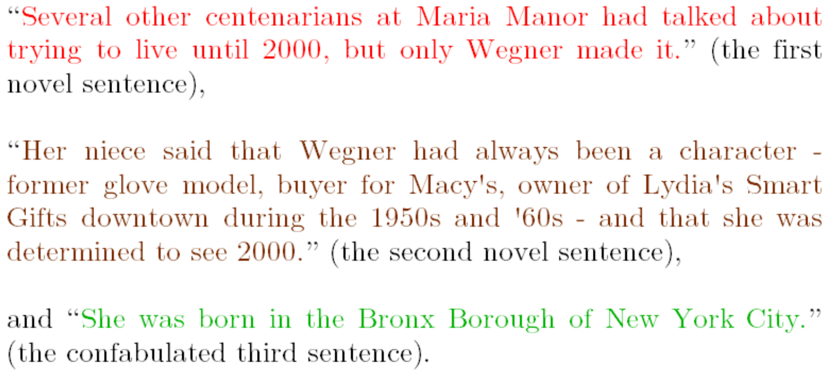
Figure 2.7: A Confabulation Theory architecture for sentence generation.
This confabulation architecture (implemented on a computer) consists of hundreds of modules (each indicated by a square—only a few of which are shown) and thousands of knowledge bases (each illustrated by an arrow connecting one module to another—again, only a few of which are shown). This particular architecture likely captures elements of thalamocortical module connectivity in the human brain, but should not be viewed as a reproduction of known connectivity.

As tens of millions of such well-written sentence triples from 1990's-vintage newspaper stories are entered into this confabulation architecture and symbols co-occur on the various connected modules, billions of knowledge links arise. Although this architecture is implemented on a computer, it is important to note that the formation of these knowledge links is consistent with the known anatomy and physiology of the human brain [Abeles, 1991; Barbas & Hilgetag, 2002; Braitenberg & Schüz, 1998; de No, 1943].

Once this architecture has completed this "reading" exposure (to a huge amount of text), its "intelligence" can be explored. Consecutive pairs of novel sentences (ones not seen during learning) are read into the modules of the system's first and second sentences (the "context sentences"). The modules of the third sentence are then commanded to confabulate. The multiconfabulation swirling of that thought process (see the red arrow in Figure 2.7) represents coordinated confabulations in many modules, which are interacting and mutually converging to single symbols. As each module converges to a single symbol the result is a plausible, although entirely made up (i.e., *confabulated*), sequence of words and punctuations in place of a third sentence.

We emphasize that the storage of the knowledge links are consistent with anatomy, the convergent confabulation operation functions identically in each module and can be biologically implemented by a thalamocortical neuronal module, and the coordination of the confabulations in the multiple modules requires no more neural circuitry than is used to control the coordination of muscles. Therefore, the simulation of this architecture is extremely biologically consistent and should be viewed as a basic simulation of a human thought process.

As an example of the results of this simulated thought processes, if the two novel consecutive context sentences (obtained from the Detroit Free Press and never before seen by the architecture) entered are:

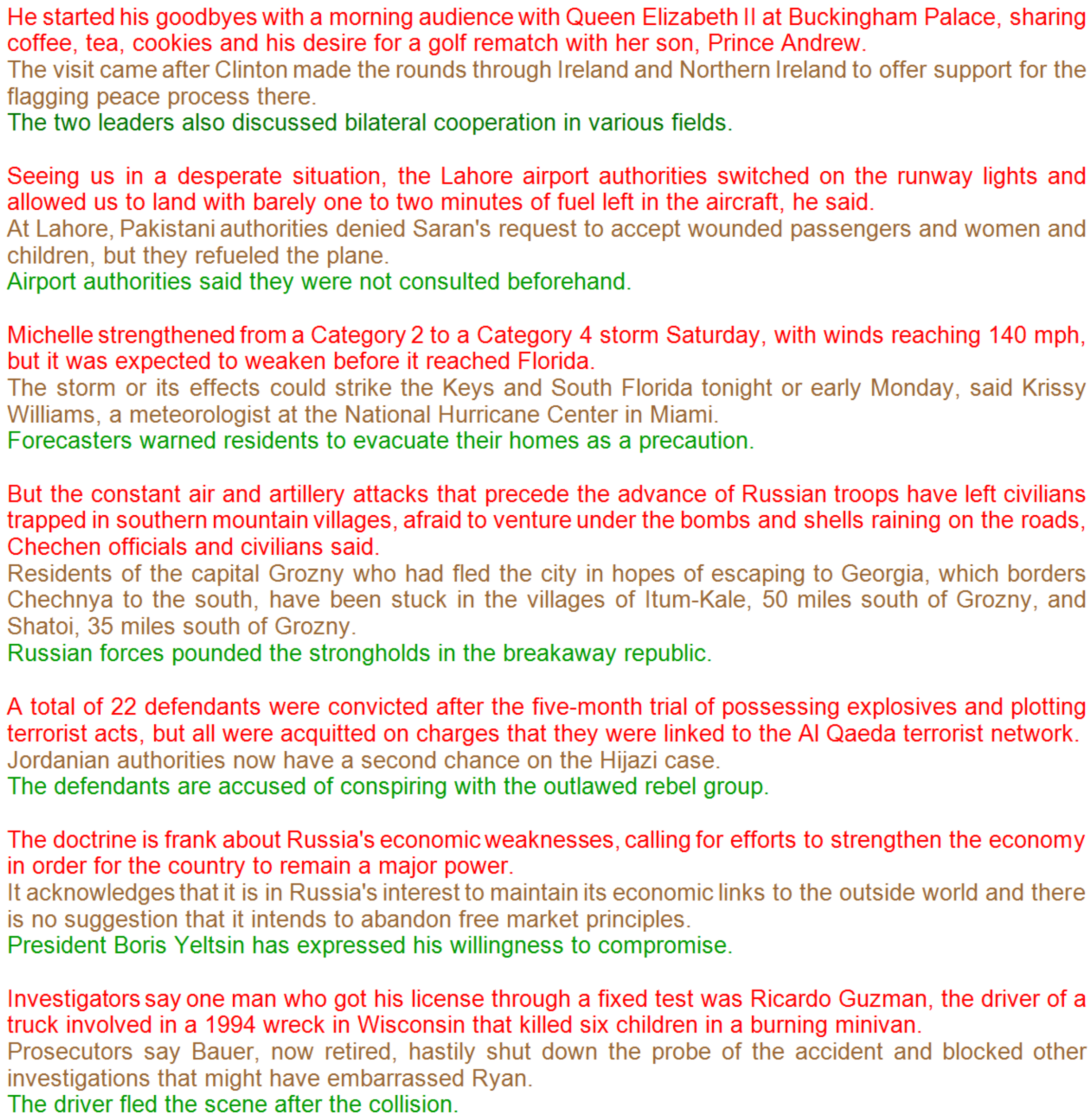


> "Several other centenarians at Maria Manor had talked about trying to live until 2000, but only Wegner made it." (the first novel sentence),
>
> "Her niece said that Wegner had always been a character - former glove model, buyer for Macy's, owner of Lydia's Smart Gifts downtown during the 1950s and '60s - and that she was determined to see 2000." (the second novel sentence),
>
> and "She was born in the Bronx Borough of New York City." (the confabulated third sentence).

Using the same color scheme, Figure 2.8 presents more examples of the operation of the confabulation architecture of Figure 2.7 (a good fraction of the outputs from randomly chosen fresh consecutive sentence pairs are of this high quality).

These results suggest that the computer simulation of this confabulation architecture must somehow be applying a deep knowledge and understanding of the general functioning of the world. The architecture links context across two previous sentences and applies that context to generate a cogent third sentence. Additionally, the third sentence produced is a grammatically correct, well structured English sentence, yet there are no rules for language in the system. In fact, the identical architecture, will produce cogent third sentences in any language given training data from that language. Interestingly, when born, all humans have the same grossly fixed brain architecture, yet each can learn any language provided that they are continually exposed to it. The emergence of grammatically correct and cogent language production in a biologically consistent architecture provides significant evidence that Confabulation Theory is in fact describing the complete fundamental mechanisms of human (mammalian) brain cognitive function.

Another example of a practical application of Confabulation Theory is in speech



He started his goodbyes with a morning audience with Queen Elizabeth II at Buckingham Palace, sharing coffee, tea, cookies and his desire for a golf rematch with her son, Prince Andrew.
The visit came after Clinton made the rounds through Ireland and Northern Ireland to offer support for the flagging peace process there.
The two leaders also discussed bilateral cooperation in various fields.

Seeing us in a desperate situation, the Lahore airport authorities switched on the runway lights and allowed us to land with barely one to two minutes of fuel left in the aircraft, he said.
At Lahore, Pakistani authorities denied Saran's request to accept wounded passengers and women and children, but they refueled the plane.
Airport authorities said they were not consulted beforehand.

Michelle strengthened from a Category 2 to a Category 4 storm Saturday, with winds reaching 140 mph, but it was expected to weaken before it reached Florida.
The storm or its effects could strike the Keys and South Florida tonight or early Monday, said Krissy Williams, a meteorologist at the National Hurricane Center in Miami.
Forecasters warned residents to evacuate their homes as a precaution.

But the constant air and artillery attacks that precede the advance of Russian troops have left civilians trapped in southern mountain villages, afraid to venture under the bombs and shells raining on the roads, Chechen officials and civilians said.
Residents of the capital Grozny who had fled the city in hopes of escaping to Georgia, which borders Chechnya to the south, have been stuck in the villages of Itum-Kale, 50 miles south of Grozny, and Shatoi, 35 miles south of Grozny.
Russian forces pounded the strongholds in the breakaway republic.

A total of 22 defendants were convicted after the five-month trial of possessing explosives and plotting terrorist acts, but all were acquitted on charges that they were linked to the Al Qaeda terrorist network.
Jordanian authorities now have a second chance on the Hijazi case.
The defendants are accused of conspiring with the outlawed rebel group.

The doctrine is frank about Russia's economic weaknesses, calling for efforts to strengthen the economy in order for the country to remain a major power.
It acknowledges that it is in Russia's interest to maintain its economic links to the outside world and there is no suggestion that it intends to abandon free market principles.
President Boris Yeltsin has expressed his willingness to compromise.

Investigators say one man who got his license through a fixed test was Ricardo Guzman, the driver of a truck involved in a 1994 wreck in Wisconsin that killed six children in a burning minivan.
Prosecutors say Bauer, now retired, hastily shut down the probe of the accident and blocked other investigations that might have embarrassed Ryan.
The driver fled the scene after the collision.

Figure 2.8: Results obtained from using the confabulation architecture after being exposed to tens of millions of triples of consecutive sentences from 1990's era news stories. Here, the first two sentences of each triple (shown, respectively, in red and brown) were consecutive sentences obtained from the same paragraph of a novel news story from the Detroit Free Press, which the architecture had not seen during its learning exposure. The third sentence (shown in green) was the sentence produced by the confabulation architecture.

The National Corn Growers Association says Gore is likely to have an ear of corn following him too if EPA sides with California officials, who oppose using ethanol.
Ten days before the Iowa caucuses, Gore was more than 20 points ahead of Bradley in various Iowa presidential polls.
Gore's aides said they would not have any problems.

The incident threatens relations between the Americans and Kosovo civilians, whom the peacekeepers were sent to protect after the 78-day NATO bombing campaign.
We don't want them here to give us security if they are going to do this, said Muharram Samakova, a neighbor of the girl's family.
NATO has struck a military airfield near Pale.

Now, I must admit that I'm not so sure the Palestinians really wanted to reach a framework agreement, Eran said Tuesday.
Eran wondered aloud whether the Palestinian strategy might be to negotiate as much land as possible in the remaining transfers, then declare statehood unilaterally – as the Palestinians have threatened to do before when talks bog down.
Netanyahu said the Palestinians would be barred from jobs in Israel.

The shortage has been attributed to rapid expansion of the prison system, low pay, a booming economy that makes the prospect of spending the day guarding convicts less attractive, and the risks of dealing with inmates who seem to be getting meaner and more violent.
Prison officials are scrambling to keep penitentiaries staffed, recruiting at schools and from the Internet.
Prison officials are still debating what they have to do.

Outside investigators announced the conclusions Tuesday as NASA's top scientist confirmed that the agency will cancel plans to launch a robot spacecraft in 2001 on a mission to land on Mars and indefinitely postpone all future launches to Mars, with one exception: a 2001 mission.
With only its aging Mars Global Surveyor in orbit around Mars, the agency is reassessing its entire approach to the exploration of the planet after losing all four of its spacecraft bound for Mars last year – a package totaling $360 million.
Mars Global Surveyor will be mapping out the planet.

However, despite his acquittal by the Senate, Clinton still faces a continuing investigation by Independent Counsel Robert, who has said he has hired additional prosecutors and is considering whether to indict Clinton after he leaves office.
Clinton said that I wouldn't be surprised by anything that happens but I'm not interested in being pardoned.
Starr is investigating the Clintons' Whitewater affair in Arkansas.

In one violent showdown in front of the Treasury Building a block from the White House, a few hundred demonstrators charged a barricade and faced a counter-assault from police swinging billy clubs and squirting pepper spray.
Closer to the IMF building police discharged a canister of bright green ammonia gas to disperse a crowd surrounding a police bus on G Street, near George Washington University.
The protesters hurled stones at the riot shields.

Figure 2.9: Additional results obtained from using the confabulation architecture after being exposed to triples of consecutive sentences.

understanding. A different confabulation architecture has demonstrated nearly perfect recognition accuracy on novel speech recordings [Hecht-Nielsen, 2007]. Again, billions of knowledge links are first learned by the architecture and then used in a confabulation simulation to understand speech.

One striking feature about these confabulation architectures is the extremely large quantity of knowledge (one knowledge link is a single *item of knowledge*) they employ and the effectiveness with which confabulation architectures exploit this knowledge in demonstrating "intelligence". Furthermore, this performance is achieved in a biologically plausible way and lacks traditional "software" or "algorithms". Since language, speech recognition, and even visual processing systems have been implemented with nothing more than modules, symbols, knowledge links, and thought control signals, we know a wide variety of cognitive tasks can be carried out by confabulation architectures. The cerebral cortex similarly seems to perform all human capabilities without significant variation in its fundamental structure between cortical areas.

More sophisticated thought processes, involving interactions between many sensory and behavioral modalities, are possible with confabulation architectures. Such fixed cognitive tasks (along with movements) can dynamically interact and be selectively activated by hierarchies of modules [Hecht-Nielsen, 2007]. Thus, enormously powerful ensembles of thought processes and movement processes can be rapidly selected and integrated.

Chapter 2, in full, is a modified reprint of the material as it appears in Physics of Life Reviews. Solari, Soren; Smith, Andrew T.; Minnett, Rupert C.J.; Hecht-Nielsen, Robert, Elsevier, 2008. The dissertation author was an investigator and author of this material.

2.5 Chapter References

Abeles, M. (1991). *Corticonics: neural circuits of the cerebral cortex.* Cambridge University Press.

Barbas, H. & Hilgetag, C. C. (2002). Rules relating connections to cortical structure in primate prefrontal cortex. *Neurocomputing*, 44-46, 301 – 308.

Barnes, J. (1984). *The complete works of Aristotle: the revised Oxford translation.* Bollingen Series. Princeton University Press.

Braitenberg, V. & Schüz, A. (1998). *Cortex: statistics and geometry of neuronal connectivity.* Studies Brain Function Series. Springer.

Brodmann, K. (1909). *Brodmann's "Localisation in the Cerebral Cortex"*. Smith-Gordon, London.

Catani, M. & Ffytche, D. H. (2005). The rises and falls of disconnection syndromes. *Brain*, 128(10), 2224 – 2239.

de No, L. R. (1943). The cerebral cortex: architecture, intracortical connections and motor projections. In J. Fulton (Ed.), *Physiology of the nervous system* (pp. 274 – 301). Oxford University Press.

Finger, S. (2001). *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press.

Hanbery, J. & Jasper, H. (1953). Independence of diffuse thalamo-cortical projection system shown by specific nuclear destructions. *Journal of Neurophysiology*, 16(3), 252–271.

Hebb, D. (2002). *The organization of behavior: a neuropsychological theory*. L. Erlbaum Associates.

Hecht-Nielsen, R. (2007). *Confabulation Theory*. Springer.

Herkenham, M. (1980). Laminar organization of thalamic projections to the rat neocortex. *Science*, 207, 532 – 535.

Jones, E. (2007). *The Thalamus*, volume 1 - 2 of *The Thalamus 2 volume set*. Cambridge University Press.

Lieber, R. (2002). *Skeletal muscle structure, function & plasticity: the physiological basis of rehabilitation*. Lippincott Williams & Wilkins.

Monsell, S. (1984). Components of working memory underlying verbal skills: A "distributed capacities" view. *International symposium on attention and performance*, 10, 327 – 350.

Northcutt, R. G. & Kaas, J. H. (1995). The emergence and evolution of mammalian neocortex. *Trends in Neurosciences*, 18(9), 373 – 379.

Penfield, W. & Rasmussen, T. (1968). *The cerebral cortex of man: A clinical study of localization of function*. Hafner Publishing Company.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, 82(3), 171 – 177.

Striedter, G. (2005). *Principles of brain evolution*. Sinauer Associates.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex*, 13(1), 90–99.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, 311(5761), 670–674.

von der Malsburg, C. (1981). *The Correlation Theory of Brain Function, Internal Report 81-2*. Technical report, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.

Chapter 3

Mammalian Vision

Vision is such a tantalizing sense. Evidence of its function lies right in front of our eyes every day, yet, despite this wealth of personal experience that the vast majority of humans have ever had with visual stimuli, how we perceive electromagnetic waves in our visible spectrum is still very poorly understood. In statistical machine learning there have been many attempts over the last several decades to provide computers with working vision, but most have been increasingly deviating from a far superior design, of which there are currently billions of working examples at our disposal: vertebrate vision. The solution to almost all computer vision problems lies in the return to a principled and fundamental understanding of nature's solution. Furthermore, such a level of understanding of vertebrate, or specifically human, vision has the potential to expedite breakthroughs in ophthalmology, neurology, and sensory systems neuroscience in general.

3.1 Retina

The retina lies on the rear inner surface of the eye onto which an inverted image of the visual field is projected through the cornea and lens. In vertebrates, the retina is a 0.5 mm thick sheet of three layers of neurons spread over a disc between 30 and 40 mm in diameter [Kolb et al., 2001]. The deepest layer consists of the photoreceptors (classified as rods or cones). The middle layer effectively combines neighboring photoreceptors and projects to the most shallow layer. The shallow layer contains ganglion cells which fire action potentials and project their axons through the optic nerve to the lateral geniculate

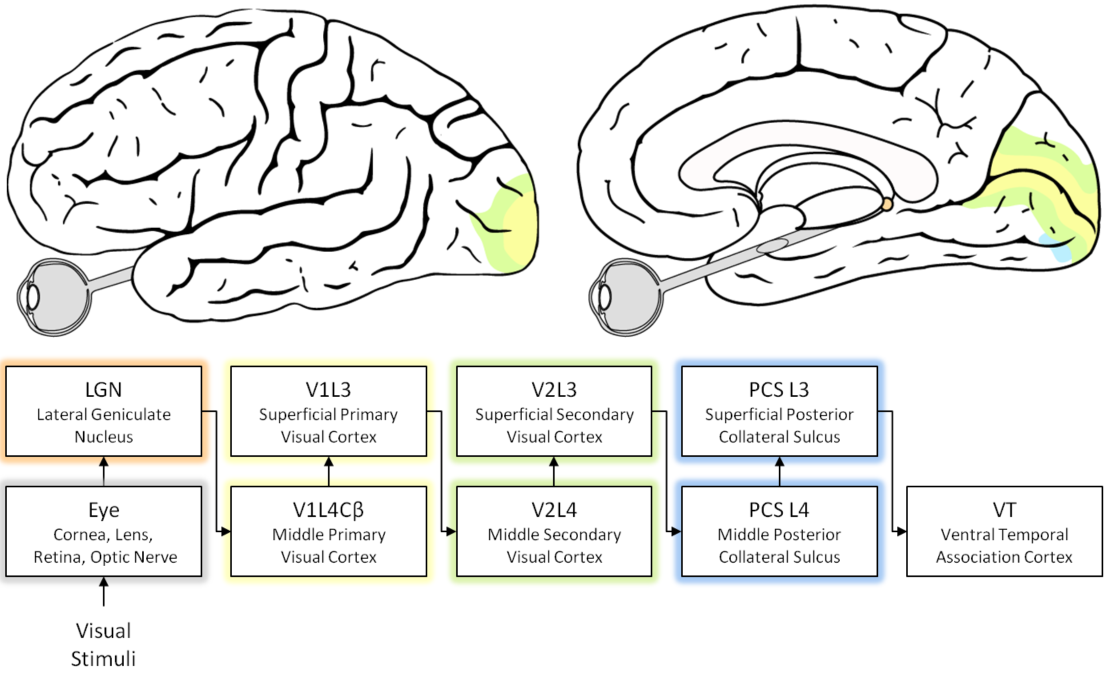

LGN
Lateral Geniculate Nucleus
V1L3
Superficial Primary Visual Cortex
V2L3
Superficial Secondary Visual Cortex
PCS L3
Superficial Posterior Collateral Sulcus
Eye
Cornea, Lens, Retina, Optic Nerve
V1L4Cβ
Middle Primary Visual Cortex
V2L4
Middle Secondary Visual Cortex
PCS L4
Middle Posterior Collateral Sulcus
VT
Ventral Temporal Association Cortex
Visual Stimuli

Figure 3.1: The neural basis for mammalian texture classification.
The posterior collateral sulcus (pCoS) appears to be implicated specifically in texture recognition, and not form recognition [Cavina-Pratesi et al., 2010]. The dominant feedforward visual pathway to the pCoS involves visual stimuli passing through the optics in the eye (grey), projecting onto the retina (grey), transmitting through the optic nerve (grey) to the lateral geniculate nuclei (LGN; orange), radiating to the primary visual cortex (V1; yellow), followed by the secondary visual cortex (V2; green), and then exciting the pCoS (blue). The ventral temporal association cortex (VT) is probably implicated in making the association between the pCoS activity and a semantic label for classification.

nuclei (LGN) of the thalamus (see Section 3.2). These ganglion cells send their axons through the retinal sheet, in a 3 mm^2 area called the optic disc, to project through the optic nerve. Consequently, vertebrates have a "blind spot" in their vision where this occurs since there is an absence of photoreceptors. Cephalopods have a similarly organized retina, except that the three layers are reversed, with the photoreceptors on the shallow layer and the ganglion cells on the deep layer, allowing for a projection of their axons into the optic nerve without any interruption of the sheet of photoreceptors and without a "blind spot".

The fovea, located in the center of the visual field (spanning roughly 2 degrees of eccentricity, or roughly two thumbnail widths at arm's length for an adult human [Hubel, 1995]) is a specialized area of the retina with a high density of cone photoreceptors, which are arranged in roughly a hexagonal grid and capable of high acuity vision and have color sensitivity (trichromatic in humans, some primates and some marsupials [Arrese et al., 2005], and dichromatic in other mammals), and an absence of rod photoreceptors. Although the fovea only occupies roughly 1% of the surface area of the retina, through cortical magnification, roughly 50% of the primary visual cortex is dedicated to processing these high-acuity cone photoreceptor responses.

In primates, there appear to be over a dozen types of ganglion cells in the retina, three of which compose roughly 88% of the population: the midget ganglion retina cells, which project to the LGN parvocellular cells, account for roughly 70%, the parasol ganglion retina cells, which project to the LGN magnocellular cells, account for roughly 10%, and the bistratified ganglion retina cells, which project to the LGN koniocellular cells, account for roughly 8% [Nassi & Callaway, 2009] (see Figure 3.2). The physiology in which the other 12% of the ganglion cells are implicated is currently not clear, but may involve blinking, gaze and pupilary control, and diurnal rhythmns.

The primary goal of sensory neurophysiology is to characterize the functional relationship between a stimulus and the neuronal response. In the case of experimental visual neurophysiology, this consists of presenting the biological model with a series of visual stimuli and recording from single or multiple neurons at a given stage of the visual processing stream, often initially to discover the stimulus that maximally excites the neuron(s) and then perturbing the stimulus to record the deviation from maximal excitation. Experimental studies of the visual system have varied greatly in the types of visual stimuli presented and methods with which the neuronal responses are recorded.

Spatial integration of the stimulus begins as early as the ganglion cells in the retina. Each ganglion cell can be characterized by the stimulus, the location and extent of which in the visual field is called the receptive field (see Figure 3.5) that most strongly excites the neuron. Receptive fields in the retina are often characterized by an on-center or off-center and an isotropic annulus (e.g. a difference of Gaussians, or DoG, filter).

3.2 Lateral Geniculate Nucleus

Most of the ganglion cells in the retina (roughly 90% in primates [Perry & Cowey, 1985]) project contralaterally and in a hemifield organization through the optic nerve and chiasm to a pair of lateral geniculate nuclei (LGN) on the posterior surface of the thalamus. The remainder of the ganglion cells project to other subcortical nuclei (e.g. the superior colliculus). The human and primate LGN is organized into six layers, enumerated 1 through 6 from the deepest to the most superficial. Layers 1 and 2 contain magnocellular magnocellular (M cells in humans or Y cells in primates). Layers 3 through 4 contain parvocellular cells (P cells in humans or X cells in primates). Koniocellular cells (K cells in humans or W cells in primates) are dispersed between and within these layers [Nassi & Callaway, 2009] (see Figure 3.2). In this dissertation, we concentrate on the parvocellular pathway, which is implicated in higher-acuity vision and is driven primarily by foveal stimuli.

The receptive field of LGN neurons, like the ganglion cells in the retina, is often characterized by an on-center or off-center and a nearly isotropic annulus expressing very weak orientation (anisotropy in the receptive field) and motion direction selectivity [Xu et al., 2002]. The on and off channel receptive field features remain separate from retina through LGN [Sherk & Horton, 1984; Reid & Alonso, 1995] and are roughly equal in the number of neurons with each preference [Krüger & Fischer, 1975; Kremers et al., 1993].

Due to their similar receptive fields to retinal ganglion cells, the LGN is often oversimplified as a relay of retinal activations to primary visual cortex, but is likely to be performing much more complex visual computation with its lateral and cortical feedback connectivity [Sillito & Jones, 2002]. This is particularly evident when more complex natural scene stimuli are presented to the subject. Synthetic stimuli (e.g. drifting gratings) elicit responses in LGN neurons that can be predicted fairly well with a linear DoG filter model (roughly 78% of the variance can be explained [Mante, 2005]). However, more complex natural scene stimuli (e.g. a video of what a cat sees walking through the grass) elicit responses that fail to be predicted well with a linear model (up to about 46% of the variance can be explained) [Mante, 2005]. Incorporating more complex nonlinear computations into the model (luminance and contrast adaptation mediated by lateral connectivity; see Figure 4.9) recovers this loss of explained variance with natural scene stimuli (roughly 78% of the variance can be explained) [Mante, 2005].

3.3 Striate Cerebral Cortex

The mammalian cerebral cortex is essentially a thin (on average less than 1 mm thick in mice to roughly 2 mm thick in humans [von Economo & Triarhou, 2009]) folded sheet of grey matter (containing the neurons) covering two hemispheres of white matter (containing the interareal fascicle projections). Hystological studies (e.g. golgi body staining [Ramón Y Cajal, 1899] and Nissl staining [Brodmann, 1909; von Economo & Koskinas, 1925; von Economo, 1929; von Economo & Koskinas, 2008]) of the cerebral cortex have identified six layers, enumerated 1 through 6 from the most superficial to the deepest; however, these six laminar cortical layers are not clearly distinct in all areas of the cortex, vary greatly in relative thickness (e.g. the primary sensory areas, compared with the association areas, and compared with the primary motor area [Solari, 2009; Solari & Stoner, 2011]), and are often subdivided further. These layers have also been distinguished due to their afferent (feedforward) and efferent (feedback) subcortical, intra-areal and interareal connectivity.

Over half of the cerebral cortex in non-human primates is devoted to visual processing and consists of at least 26 distinguishable regions [Sereno et al., 1995]. Of all of the cortical areas, primary visual cortex (V1 or Brodmann Area 17 [Brodmann, 1909]) is greatly overrepresented in the neuroscience literature due to its accessibility on the gyri surrounding the calcarine sulcus and its reliably and easily excited pyramidal cells using synthetic stimuli (e.g. bars and drifting gratings). Furthermore, in humans, V1 can be identified by the naked eye from the stria of Gennari formed by the myelinated axons terminating in layer 4, hence V1's alias: the *striate* cortex for its stripes.

The intracortical V1 circuit is complex and involves many collateral and feedback projections, but the dominant feedforward parvocellular pathway through V1 involves thalamocortical afferents combining LGN receptive fields and terminating in V1 layer 4Cβ, whose pyramidal cells then project to V1 layer 3 (often referred to as layer 2/3 due the unclear physiological boundary in anaesthetised subjects which clearly segregates,

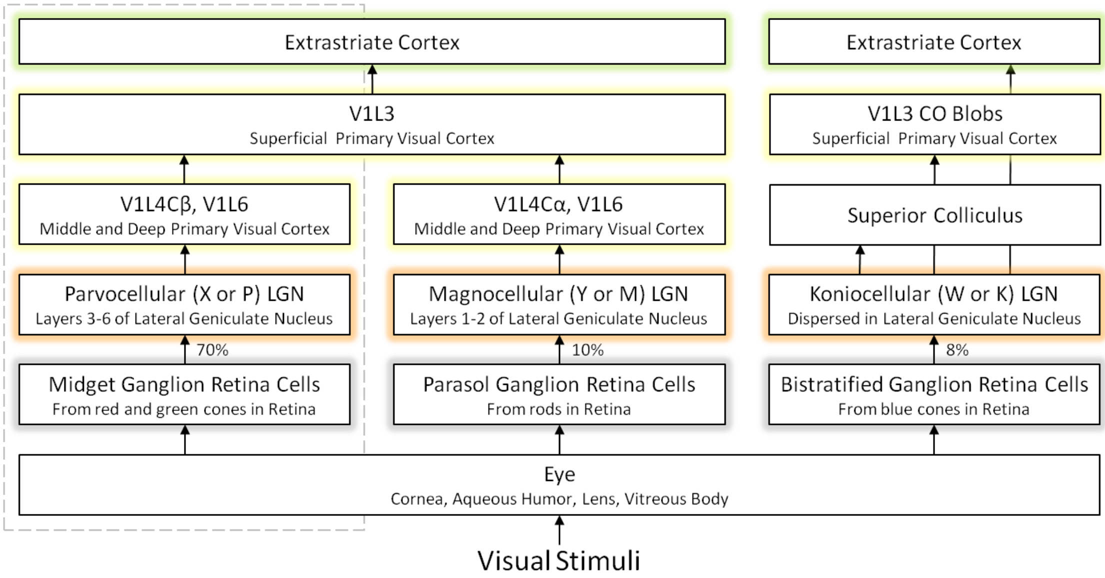

Extrastriate Cortex
Extrastriate Cortex
V1L3
Superficial Primary Visual Cortex
V1L3 CO Blobs
Superficial Primary Visual Cortex
V1L4Cβ, V1L6
Middle and Deep Primary Visual Cortex
V1L4Cα, V1L6
Middle and Deep Primary Visual Cortex
Superior Colliculus
Parvocellular (X or P) LGN
Layers 3-6 of Lateral Geniculate Nucleus
Magnocellular (Y or M) LGN
Layers 1-2 of Lateral Geniculate Nucleus
Koniocellular (W or K) LGN
Dispersed in Lateral Geniculate Nucleus
70%
10%
8%
Midget Ganglion Retina Cells
From red and green cones in Retina
Parasol Ganglion Retina Cells
From rods in Retina
Bistratified Ganglion Retina Cells
From blue cones in Retina
Eye
Cornea, Aqueous Humor, Lens, Vitreous Body
Visual Stimuli

Figure 3.2: Parallel pathways in the mammalian visual processing stream.
The early mammalian visual processing stream can be roughly decomposed into three dominant parallel pathways (the vision research in this dissertation concentrates on the parvocellular pathway, outlined in a dashed bounding box). The long (red) and medium (green) wavelength cone photoreceptors (mainly in the fovea) drive the midget cells (roughly 70% of the retinal ganglion cells [Nassi & Callaway, 2009]), which project through LGN parvocellular cells to V1 layer 4Cβ pyramidal cells. This parvocellular pathway is slower to respond, has smaller receptive fields, is insensitive to contrast, and prefers low temporal and high spatial frequencies. The rod photoreceptors (mainly outside of the fovea) drive the parasol cells (roughly 10% of the retinal ganglion cells [Nassi & Callaway, 2009]), which project through LGN magnocellular cells to V1 layer 4Cα pyramidal cells. This magnocellular pathway complements the parvocellular pathway and is faster to respond, has larger receptive fields, is sensitive to contrast, and prefers high temporal and low spatial frequencies. The short (blue) wavelength cone photoreceptors drive the bistratified cells (roughly 8% of the retinal ganglion cells [Nassi & Callaway, 2009]), which project through the LGN koniocellular cells to the superior colliculus, V1 layer 3B cytochrome oxidase blobs and V1 layer 4A and possibly directly to extrastriate cortex [Hendry & Reid, 2000]). The koniocellular pathway appears to be implicated in low acuity vision and blind sight. The remaining 12% of the retinal ganglion cells may be implicated in blinking, gaze and pupilary control, and diurnal rhythmns. Beyond V1 L4, the parallel pathways are not nearly as separable and appear to follow a bipartite projection to V2 [Sincich & Horton, 2005], rather than the originally proposed tripartite projection of color, form, and motion [Livingstone & Hubel, 1988].

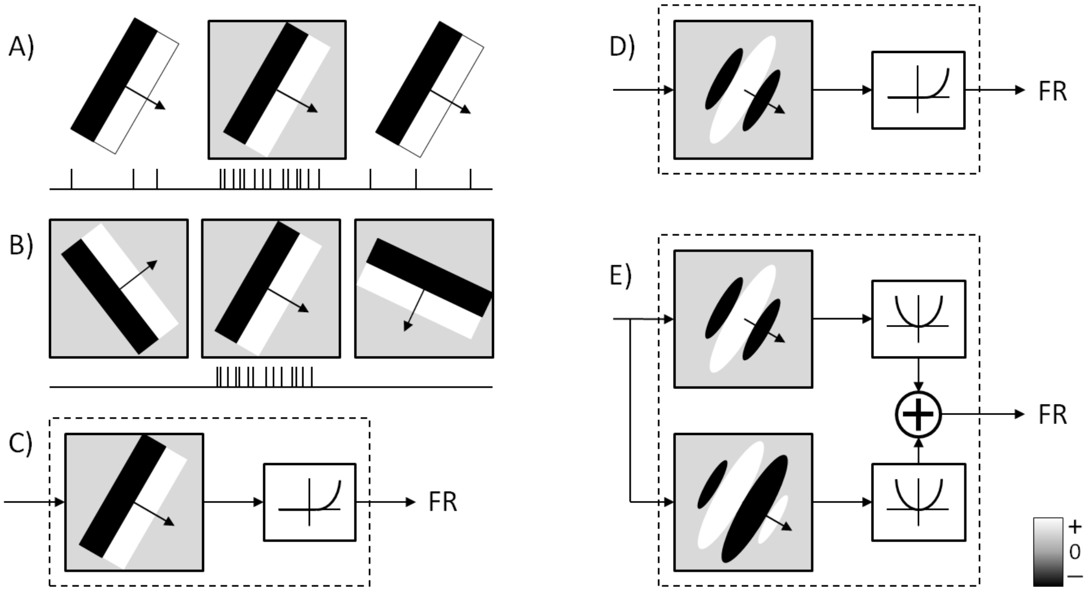


Figure 3.3: The standard linear model of visual neurons.
David Hubel and Torsten Wiesel discovered, almost accidentally (drawing stimuli by hand on slides for a projector screen failed to excite the neuron, but the motion of removing the slide from the projector, thereby presenting a sliding bar stimulus, haphazardly managed to excite the cell), the V1 simple cell receptive field [Hubel & Wiesel, 1962] and consequently won the 1981 Nobel Prize. The neuron's receptive field extent (grey boxes) can be coarsely estimated by the boundary at which the stimulus begins to excite the neuron. **A)** A bar stimulus, oriented correctly, will only drive the neuron to fire (spike trains are shown below the stimuli) when presented within the neuron's receptive field. **B)** A bar stimulus will drive the neuron to fire when presented in the receptive field at the optimal orientation and may suppress the neuron from firing when presented at a suboptimal orientation. **C)** Computational models based on the simple cells described by Hubel & Wiesel [1962] often involve treating the optimal stimulus (receptive field) as a linear filter, followed by an activation function, to approximate the neuron's instantaneous firing rate (FR). **D)** Daugman [1985] estimated the spatial properties of the receptive field more explicitly and parameterized them as a two-dimensional Gabor filters ([Gabor, 1946]), which were experimentally confirmed by Jones et al. [1987]. **E)** Adelson & Bergen [1985] proposed the energy of the response of a quadrature pair of Gabor filters as a computational model for complex cells.

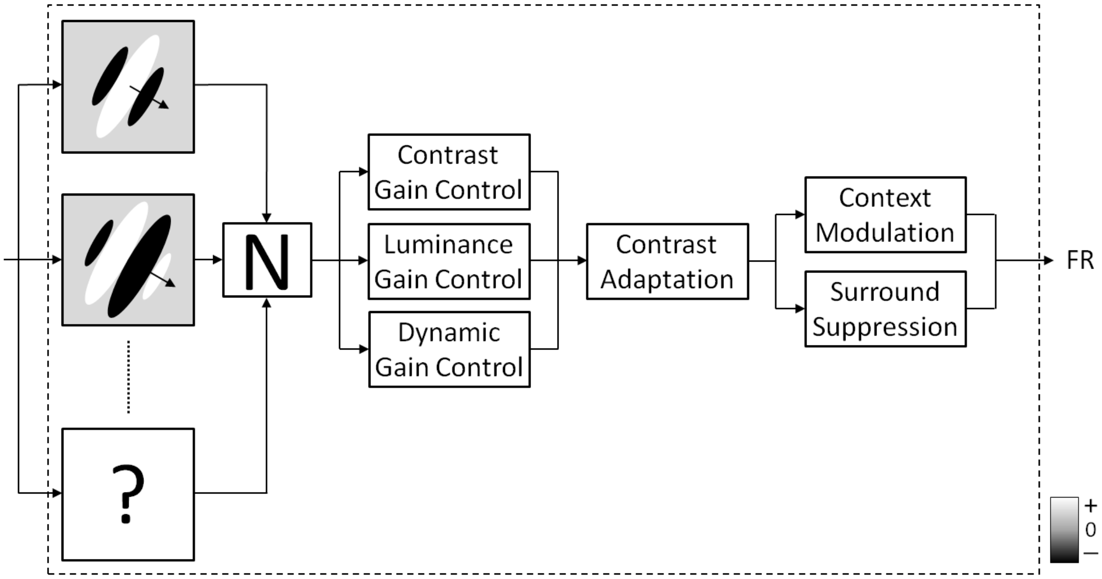


Figure 3.4: The nonlinear model of visual neurons.
Despite its popularity in computer vision, the standard linear model of primary visual cortex (V1) is incomplete (see Figure 3.3 D and E). When this linear model is presented with natural stimuli (e.g. a video of what a cat sees walking through the grass), rather than synthetic stimuli (e.g. bars and drifting gratings), it fails to predict V1 neuron responses. Rust & Movshon [2005] convincingly argue that the standard model should be updated to include the nonlinearities that have been observed in V1 and driven with additional nonlinearly-combined filters, all of which should be discovered with more complex and carefully crafted synthetic stimuli (e.g. [Tanaka & Ohzawa, 2009]) and then tested with natural stimuli. Adapted from [Rust & Movshon, 2005].

consistently with other cortical areas [Solari & Stoner, 2011], into a feedforward layer 3 and feedback layer 2 [Gur & Snodderly, 2008]).

V1 layer 4 pyramidal cells have been investigated intensely since David Hubel and Torsten Wiesel discovered the simple cell edge-detecting receptive field [Hubel & Wiesel, 1962] (see Figure 3.3). This strong orientation selectivity exhibited by many simple cells in V1 layer 4 and simple and complex cells in layer 3 is almost always organized in a semiregular, smoothly-varying orientation selectivity retinotopic map. These orientation maps are reproducible simply with feedforward and lateral connectivity [McKinstry & Guest, 1997, 2001]; however, they are not found in rodents or lagomorphs, which appear not to suffer from any related visual deficits [Van Hooser et al., 2005]. Consequently, orientation column organization may not be necessary in V1 and may only be a solution to the retinotopically limited arborization of isotropic lateral connectivity.

Hubel & Wiesel [1962] described most pyramidal neurons in V1 as belonging to one of two classes: simple cells (those whose receptive field consists of distinct excitatory and suppressive regions and are common in layer 4) and complex cells (those that are not simple cells and are common in layer 3). Computationally, these simple cells are often modeled as linear Gabor filters [Gabor, 1946; Jones et al., 1987; Daugman, 1985] and the complex cells are often modeled as the energy of a quadrature pair of linear Gabor filters [Adelson & Bergen, 1985] (see Figure 3.3 D and E). Evidence now suggests that these two classes (which may be prominent due to a selection bias) form two ends of a continuum mediated by the degree of lateral inhibition and feedback influencing the cells [Priebe et al., 2004; Mata & Ringach, 2005]. Similarly to LGN, when V1 layer 3 and 4 pyramidal neurons are presented with natural scene stimuli, their responses exhibit complex nonlinearities (e.g. luminance and contrast gain control, cross-orientation suppression, iso-orientation surround suppression, co-linear long-range facilitation in layer 3, etc.). However, nonlinear models designed to predict these responses are still incomplete [Rust & Movshon, 2005]: roughly 84% of the variance in V1 responses to drifting gratings can be explained by linear models [Carandini et al., 1997], but when presented with natural scene stimuli, roughly 21% can be explained by linear models and only roughly 40% by nonlinear models [David & Gallant, 2005].

3.4 Extrastriate Cerebral Cortex

Extrastriate cortex consists of many cortical areas implicated in visual processing. Although drastically oversimplified, these areas are sometimes divided into a ventral ("what") pathway for processing visual content and a dorsal ("where") pathway for processing spatial information. The primate ventral pathway projects from V1 to V2 to V4 to inferotemporal cortex (IT). This pathway can be approximated by a feedfoward model due to the anatomically similar, but physiologically asymmetrical, interareal connectivity. For example, there is no obvious difference in the axonal or synaptic anatomy between the V1 to V2 projection and the V2 to V1 projection, yet V1 can drive V2 and not vice-versa, perhaps due to a tightly temporally synchronized feedforward projection and a temporally diffuse feedback projection [Anderson & Martin, 2009]. As the visual stimulus is processed through the ventral pathway, it appears to be processed similarly in each area with the receptive fields becoming larger (e.g. V2 surround mechanisms appear to be the same as in V1, but spatially scaled up by a factor of 2 [Shushruth et al., 2009]), but beyond V1, the physiological differences between cortical layers in poorly understood

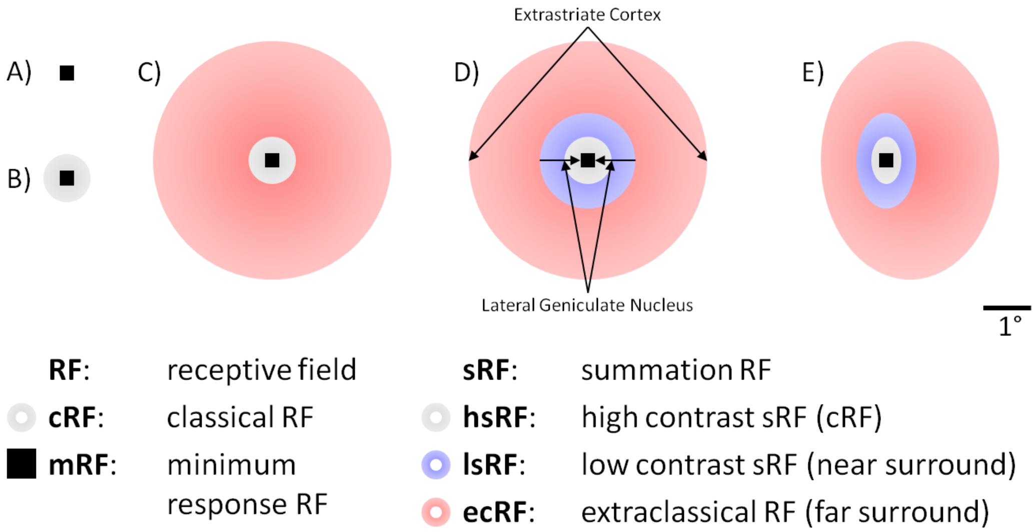

Extrastriate Cortex
A)
B)
C)
D)
E)
Lateral Geniculate Nucleus
1°
RF: receptive field
cRF: classical RF
mRF: minimum response RF
sRF: summation RF
hsRF: high contrast sRF (cRF)
lsRF: low contrast sRF (near surround)
ecRF: extraclassical RF (far surround)

Figure 3.5: Receptive fields in the visual processing stream.
The receptive field (RF) extent of a neuron is the region of the visual field that excites a neuron. **A)** Moving the optimal stimulus in and out of the RF to estimate the minimum response receptive field (mRF) boundaries [Hubel & Wiesel, 1962] tends to underestimate the RF's spatial extent (∼ 0.5° diameter at ∼ 5° eccentricity [Angelucci et al., 2002] . **B)** Estimating the classical receptive field (cRF) size through reverse correlation methods (e.g. spike triggered averaging) is more accurate [Jones et al., 1987]. **C)** Although stimuli outside of the cRF, but inside the extraclassical RF (ecRF; diameter at ∼ 5° eccentricity [Angelucci et al., 2002]) cannot sufficiently depolarize the neuron past the spiking threshold, they can modulate the response (e.g. iso-orientation surround suppression). **D)** Low contrast stimuli result in a larger cRF (∼ 2° diameter at ∼ 5° eccentricity), perhaps mediated by horizontal monosynaptic extent, than with high contrast stimuli (∼ 1° diameter at ∼ 5° eccentricity), mediated by geniculocortical afferents [Angelucci et al., 2002]. The extrastriate (V2) feedback also appears to be coextensive with, and perhaps the source of, the ecRF [Angelucci et al., 2002; Schwabe et al., 2006]. **E)** More complex synthetic stimuli (with higher-order features) estimate anisotropic and nonconcentric facilitatory and suppressive regions whose major axes are parallel and not correlated with the orientation selectivity of the neuron [Tanaka & Ohzawa, 2009].

[Shipp, 2007]. There is increasing evidence that texture, color and form are processed independently in V2 and V4, before being combined in areas for complex stimuli like faces and places in IT [Cavina-Pratesi et al., 2010]. V2 and the posterior collateral sulcus (pCoS), a human homologue in the medial occipital lobe of part of the primate V4, specifically appear to be implicated in texture perception [Cavina-Pratesi et al., 2010].

Chapter 3, in part, has been submitted for publication of the material. Minnett, Rupert C.J.; Hecht-Nielsen, Robert. The dissertation author was the primary investigator and author of this material.

3.5 Chapter References

Adelson, E. H. & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A: Optic*, 2(2), 284–299.

Anderson, J. C. & Martin, K. A. C. (2009). The synaptic connections between cortical areas v1 and v2 in macaque monkey. *Journal of Neuroscience*, 29(36), 11283–11293.

Angelucci, A., Levitt, J. B., Walton, E. J. S., Hupe, J.-M., Bullier, J., & Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22(19), 8633–8646.

Arrese, C. A., Oddy, A. Y., Runham, P. B., Hart, N. S., Shand, J., Hunt, D. M., & Beazley, L. D. (2005). Cone topography and spectral sensitivity in two potentially trichromatic marsupials, the quokka (setonix brachyurus) and quenda (isoodon obesulus). *Proceedings of the Royal Society B Biological Sciences*, 272(1565), 791–796.

Brodmann, K. (1909). *Brodmann's "Localisation in the Cerebral Cortex"*. Smith-Gordon, London.

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21), 8621–8644.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010). Separate processing of texture and form in the ventral stream: evidence from fmri and visual agnosia. *Cerebral Cortex*, 20(2), 433–446.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A Optics and image science*, 2(7), 1160–1169.

David, S. V. & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network*, 16(2-3), 239–260.

Gabor, D. (1946). Theory of communication. *Communication Theory*, 93(26), 429–457.

Gur, M. & Snodderly, D. M. (2008). Physiological differences between neurons in layer 2 and layer 3 of primary visual cortex (v1) of alert macaque monkeys. *The Journal of Physiology*, 586(Pt 9), 2293–2306.

Hendry, S. H. & Reid, R. C. (2000). The koniocellular pathway in primate vision. *Annual Review of Neuroscience*, 23(1), 127–153.

Hubel, D. (1995). *Eye, brain, and vision*. Scientific American Library series. Scientific American Library.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154.2.

Jones, J. P., Stepnoski, A., & Palmer, L. A. (1987). The two-dimensional spectral structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6), 1233–1258.

Kolb, H., Nelson, R., Ahnelt, P., & Cuenca, N. (2001). Cellular organization of the vertebrate retina. *Progress in Brain Research*, 131, 3–26.

Kremers, J., Lee, B. B., Pokorny, J., & Smith, V. C. (1993). Responses of macaque ganglion cells and human observers to compound periodic waveforms. *Vision Research*, 33(14), 1997–2011.

Krüger, J. & Fischer, B. (1975). Symmetry between the visual b- and d-systems and equivalence of center and surround: studies of light increment and decrement in retinal and geniculate neurons of the cat. *Biological Cybernetics*, 20(3-4), 223–236.

Livingstone, M. & Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240(4853), 740–749.

Mante, V. (2005). *Gain controls based on luminance and contrast in the early visual system*. PhD thesis, Die Eidgenössische Technische Hochschule Zürich.

Mata, M. L. & Ringach, D. L. (2005). Spatial overlap of on and off subregions and its relation to response modulation ratio in macaque primary visual cortex. *Journal of Neurophysiology*, 93(2), 919–928.

McKinstry, J. & Guest, C. (1997). *Self-organizing map develops V1 organization given biologically realistic input*, volume 1, (pp. 338–343). IEEE Service Center.

McKinstry, J. & Guest, C. (2001). Long range connections in primary visual cortex: a large scale model applied to edge detection in gray-scale images. In *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, volume 2 (pp. 843 –847 vol.2).

Nassi, J. J. & Callaway, E. M. (2009). Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, 10(5), 360–72.

Perry, V. H. & Cowey, A. (1985). The ganglion cell and cone distributions in the monkey's retina: implications for central magnification factors. *Vision Research*, 25(12), 1795–1810.

Priebe, N., Mechler, F., Carandini, M., & Ferster, D. (2004). The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience*, 7(10), 1113–22.

Ramón Y Cajal, S. (1899). Estudios sobre la corteza cerebral humana i. corteza visual. *Revista Trimestral Micrográfica*, 4.

Reid, R. C. & Alonso, J. M. (1995). 1995 nature publishing group. *Nature*, 378(6554), 281–283.

Rust, N. C. & Movshon, A. J. (2005). In praise of artifice. *Nature Neuroscience*, 8(12), 1647 – 1650.

Schwabe, L., Obermayer, K., Angelucci, A., & Bressloff, P. C. (2006). The role of feedback in shaping the extra-classical receptive field of cortical neurons: a recurrent network model. *Journal of Neuroscience*, 26(36), 9117–9129.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889–93.

Sherk, H. & Horton, J. C. (1984). Receptive field properties in the catÊŒs area 17 in the absence of on-center geniculate input. *Journal of Neuroscience*, 4(2), 381–393.

Shipp, S. (2007). Structure and function of the cerebral cortex. *Current Biology*, 17(12), R443 – R449.

Sillito, A. M. & Jones, H. E. (2002). Corticothalamic interactions in the transfer of visual information. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 357(1428), 1739–1752.

Sincich, L. C. & Horton, J. C. (2005). The circuitry of v1 and v2: integration of color, form, and motion. *Annual Review of Neuroscience*, 28(1), 303–326.

Solari, S. V. H. (2009). *A unified anatomical theory and computational model of cognitive information processing in the mammalian brain and the introduction of DNA reco codes.* PhD thesis, University of California, San Diego.

Solari, S. V. H. & Stoner, R. M. (2011). Cognitive consilience: Primate non-primary neuroanatomical circuits underlying cognition. *Frontiers in Neuroanatomy*, 5.

Tanaka, H. & Ohzawa, I. (2009). Surround suppression of v1 neurons mediates orientation-based representation of high-order visual features. *J Neurophysiol*, 101(3), 1444–62.

Van Hooser, S. D., Heimel, J. A., Chung, S., Nelson, S. B., & Toth, L. J. (2005). Orientation selectivity without orientation maps in visual cortex of a highly visual mammal. *Journal of Neuroscience*, 25(1), 19–28.

von Economo, C. (1929). The cytoarchitectonics of the human cerebral cortex. *Journal of Anatomy*, 63(Pt 3), 389.

von Economo, C. & Koskinas, G. N. (1925). *Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen.* Springer Verlag, Vienna.

von Economo, C. & Koskinas, G. N. (2008). *Atlas of Cytoarchitectonics of the Adult Human Cerebral Cortex* (translated, revised and edited by L.C. Triarhou). Karger, Basel.

von Economo, C. & Triarhou, L. C. (2009). Cellular structure of the human cerebral cortex. *Brain*, 133(3), 38–49.

Xu, X., Ichida, J., Shostak, Y., Bonds, A. B., & Casagrande, V. A. (2002). Are primate lateral geniculate nucleus (lgn) cells really sensitive to orientation or direction? *Visual Neuroscience*, 19(1), 97–108.

Chapter 4

Texture Classification

Texture classification is a particularly good, yet relatively underappreciated, experimental paradigm for testing computational vision models with the goal of understanding mammalian vision. Historically, object recognition has been a far more popular task to attempt to perform in computer vision, perhaps because of its obvious commercial value, but neuroscience clearly indicates that object recognition is performed significantly far downstream in the ventral visual processing stream (mainly in various areas of the fusiform gyrus of the inferotemporal cortex). The neural basis of texture recognition, on the other hand, has recently been identified in human patients with visual agnosia as the posterior collateral sulcus (pCoS) [Cavina-Pratesi et al., 2010], which is roughly part of the V4 macaque homologue and is posterior (upstream in the ventral visual processing stream) from the fusiform gyrus. Texture recognition, therefore, is probably a more primitive task than object recognition and more appropriate for elucidating the biological mechanisms involved in mammalian early vision perception.

4.1 Existing Methods

The experiments that have been performed with computational models of texture classification can be roughly organized into two groups: biologically inspired methods and non-biologically inspired methods. The biologically inspired models attempt to draw insights from monumental neuroscience results and create computationally tractable algorithms. The non-biologically inspired methods seek to optimize classification accuracy using tools developed in computer vision which are theoretically advantageous and often

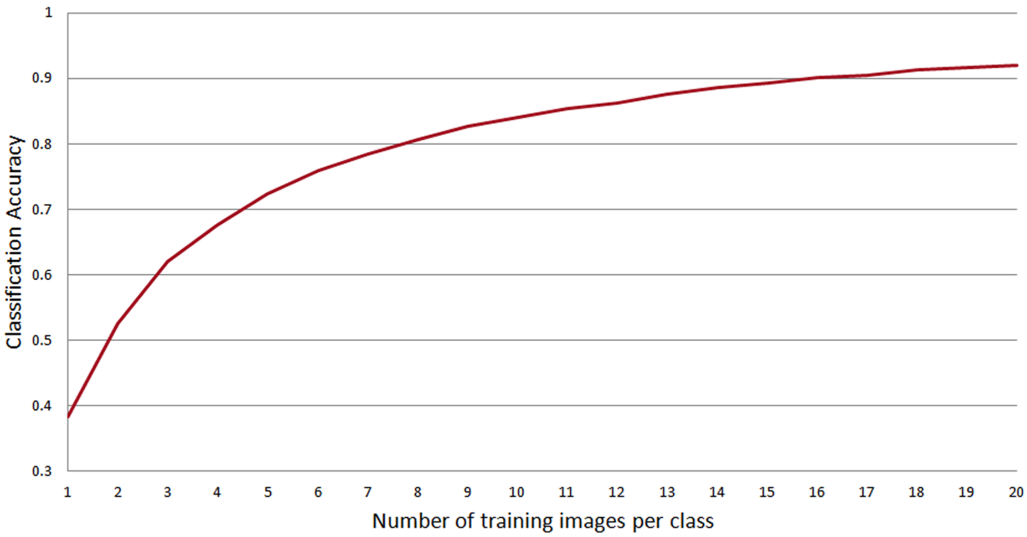


Figure 4.1: Texture classification experimental paradigm typical results.
The typical supervised learning computational experimental paradigm involves exposing the model to a portion of the labeled dataset for training, predicting the labels on the rest of the dataset, and reporting the accuracy of the prediction as the proportion of the dataset used for training is varied (the Monte Carlo variation). In this case the classification accuracy is 38.4% with only one training image per class and increases monotonically to 92% with twenty training images per class. Repeated random sub-sampling validation (typically over 10 to 100 trials) is implemented at each number of training images per class and the mean and standard deviation are reported, although the standard deviation is often only reported for the best performing training proportion step.

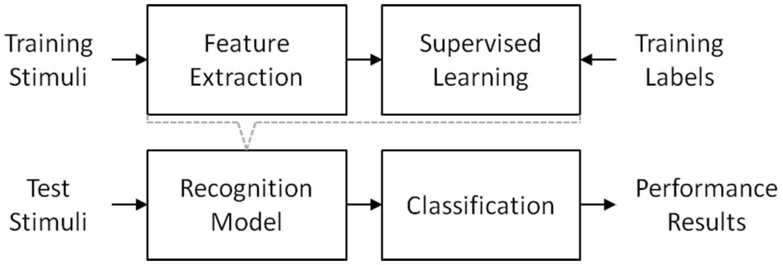


Figure 4.2: Texture classification experimental paradigm typical supervised learning model.
Existing methods of computational texture classification involve supervised learning with extracted features from training samples, followed by a classifier operating on the test samples. The goal of the feature extraction step is to extract information with sufficient predictive power and discard noise that could detract from the classification performance. The classification step can be implemented in many ways, some of which are more biologically plausible that others.

easier to implement programmatically, but may bear no resemblance to nature's solution to the problem.

4.1.1 Biologically Inspired Features

The scale-invariant feature transform (SIFT) [Lowe, 1999] and its rotationally invariant counterpart (RIFT) [Lazebnik et al., 2004; Mikolajczyk & Schmid, 2004] are inspired by some of the observed properties in inferotemporal cortex (IT), parts of which are often implicated in object recognition. Although there has been significant success in applying these feature detectors and their variants to many computer vision tasks [Lazebnik et al., 2004; Serre et al., 2005; Zhang et al., 2006], the algorithm roughly consists of identifying key points as robustly as possible, heuristically removing key points that fail to match criteria, and assigning orientations and scales to these key points to form descriptor vectors. This approach, unfortunately, does not help to reverse engineer the mammalian solution to invariant feature extraction which could hold the clues to improved performance and consists of a hierarchy of repeated feature extractions and invariance recombinations [Fukushima, 1975, 2008].

4.1.2 Non-Biologically Inspired Features

There are also several non-biologically inspired features intended for texture classification. Most of these feature extraction methods consist of calculating image luminance statistics, but alone often fail to be robust to noise and fail to be invariant affine transformations. The grey-level co-occurrence matrix (GLCM) [Haralick et al., 1973], or sometimes referred to as the spatial grey-level dependence matrix (SGLDM), is a method for capturing pair-wise luminance statistics from the entire image. Local binary patterns (LBP) [Ojala et al., 1994] are a fairly simple rotationally and luminance invariant method of extracting features by comparing neighboring pixels in a radial fashion and reporting a positive or negative luminance differential. Patch statistics approaches typically involve extracting the joint distribution of luminance values with compact neighborhoods to build a texton library [Varma & Zisserman, 2003]. Spin images [Johnson, 1997; Johnson & Hebert, 1999] are a feature extraction method designed to be invariant to rigid transformations for 3-D surface matching and registration, which alone do not directly benefit texture classification, but have been successfully combined scale-invariant feature transforms for this task [Lazebnik et al., 2005].

4.1.3 Biologically Inspired Classifiers

K-nearest neighbors (KNN) [Cover & Hart, 1967]is a popular instance-based classifier that labels test samples based on the label of a winner-take-all competition amongst the k nearest (in the chosen metric) neighboring training samples. Training a KNN classifier is relatively computationally inexpensive, but can require large amounts of storage and can be computationally expensive during testing. The number of neighbors is a free parameter that must be chosen heuristically or optimized with cross-validation techniques. Variations of KNN involve weighting training samples, methods for breaking competition ties, and choosing or learning different or multiple distance metrics. Although KNN was not biologically inspired, with simple distance metrics (e.g. cosine distance) it is likely to be biologically plausible.

Artificial neural networks (ANN), specifically multilayer perceptrons (MLP) [Rosenblatt, 1961], configured as classifiers, and their variants such as learning vector quantization (LVQ) [Kohonen, 1995] and self-organizing maps (SOMs) [Kohonen, 1998], are also likely to be biologically plausible and are certainly biologically inspired. MLPs can be computationally expensive to train, but are very inexpensive to test and, depending on their implementation, can be susceptible to complications during training (see Chapter 5).

4.1.4 Non-Biologically Inspired Classifiers

There are many other classifiers, often designed for specific applications, but are most likely not biologically plausible. The most popular and successful of which, at least in computational texture classification, is the support vector machine (SVM) [Cortes & Vapnik, 1995], a maximum-margin binary classification algorithm which depends on a kernel function to map the input space into a linearly separable rerepresentation for discrimination. Some adjustments can be made to make the algorithm more biologically relevant (e.g. incremental and decremental online learning [Cauwenberghs & Poggio, 2001]); however, choosing, crafting, or learning the correct kernel function for a specific classification task is still an active field of research. SVMs are designed for discrimination performance and are not optimized for generalization, which can pose problems if the training data differ substantially, even if consistently, from the test data [Hayman et al., 2004].

4.2 Our Approach

Rather than follow the traditional supervised learning paradigm, we draw inspiration from the biological solution: unsupervised learning followed by supervised learning (see Figure 4.4). Infant humans are exposed to a plethora of visual stimuli long before they are taught by a caretaker to label those stimuli. Similarly, our system is first exposed to a corpus of unlabeled natural scene stimuli (see Section 4.4) about which it learns statistical properties and develops a hierarchy of features. Unlike the previously published computational texture classification algorithms, which largely consist of a feature extraction step followed by a classifier, our approach consists of a hierarchy of repeated feature extraction and invariance before classification.

4.3 Texture Datasets

Several texture datasets have been used as the stimuli for texture classification experiments in the past, each with their advantages and limitations. The earlier computational texture classification experiments, in the 1980s, were performed using the Brodatz dataset [Brodatz, 1981], followed by more complex datasets published and adopted in the two successive decades. These datasets vary tremendously in their complexity and the extent of interclass and interclass variations. The Columbia-Utrecht Reflectance and

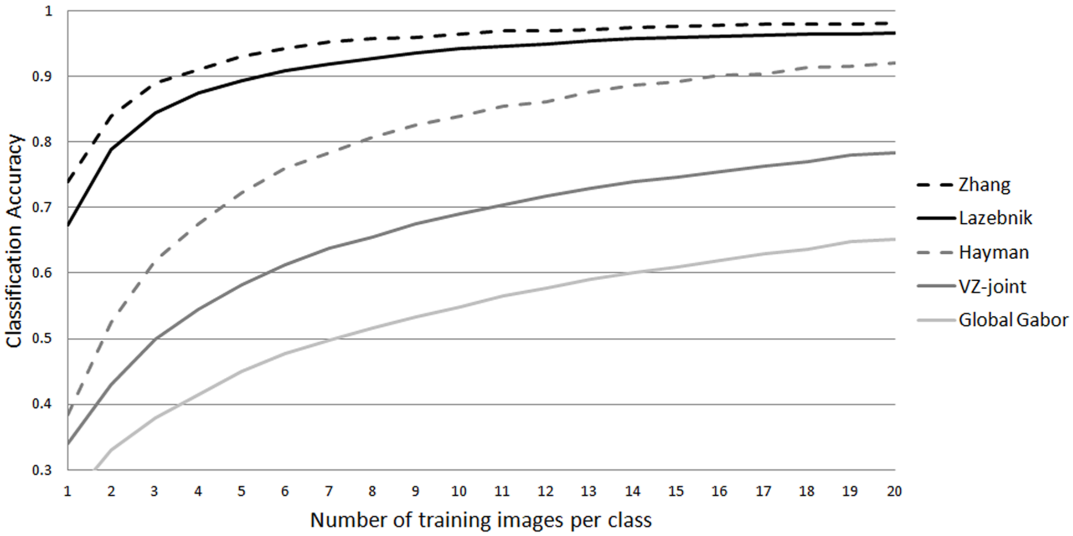


Figure 4.3: Previously published classification accuracy on the UIUCTex dataset. Classification accuracy results on the UIUCTex dataset varying the number of training images per class and reporting the mean over repeated random sub-sampling validation (N=100) are reproduced from [Zhang et al., 2006]. The "Global Gabor" method [Ma & Manjunath, 1996] uses a Gabor filter bank (six orientations and four spatial scales) for feature extraction and a KNN classifier with a Mahalanobis distance metric. The "VZ-joint" method [Varma & Zisserman, 2003] creates histograms from pixel patches for feature extraction and a KNN classifier with a χ^2 distance metric. The "Hayman" method [Hayman et al., 2004] uses a VZ-MR8 filter bank [Varma & Zisserman, 2002] for feature extraction and a support vector machine (SVM) [Cortes & Vapnik, 1995] classifier with a radial basis function (RBF) kernel and a χ^2 distance metric. The "Lazebnik" method [Lazebnik et al., 2005] uses affine invariant Harris-Laplace and Laplacian detectors [Lindeberg & Gårding, 1997; Mikolajczyk & Schmid, 2004], spin images [Johnson, 1997; Johnson & Hebert, 1999] and the rotationally invariant feature transform (RIFT) [Lazebnik et al., 2004; Mikolajczyk & Schmid, 2004] descriptors with a KNN classifier the Earth Mover's Distance (EMD) [Rubner et al., 2000] metric. The "Zhang" method [Zhang et al., 2006] uses scale and rotation invariant Harris-Laplace and Laplacian detectors, spin images and RIFT detectors with an SVM classifier and the EMD kernel.

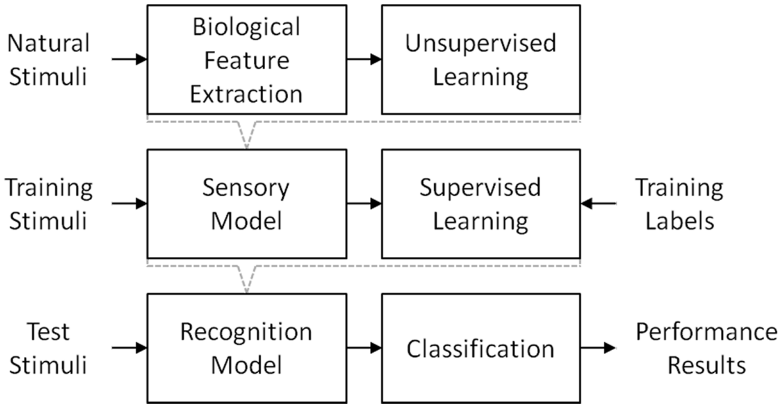


Figure 4.4: Our texture classification experimental paradigm unsupervised learning model.
The biological solution for visual processing involves unsupervised learning from unlabeled stimuli, followed by supervised learning from labeled data. Following this approach, especially in multimodal context, has already demonstrated improved performance over supervised learning alone [de Sa, 1998]. Our approach to texture classification involves presenting the model with unlabeled natural scene stimuli prior to supervised learning on the labeled texture dataset.

Texture Database (CUReT) [Dana et al., 1999] introduces the complicated effects of specularities and shadowing under varying illumination directions. The University of Illinois at Urbana-Champaign Texture Dataset (UIUCTex) [?] introduces very large spatial scale variations and non-rigid deformations, making it the most challenging of the three datasets.

4.3.1 Brodatz Dataset

The Brodatz Dataset[1] is a collection of 111 (typically 512 by 512 pixels to 632 by 632 pixels) scans of texture photograph negatives from the book. These 111 images (textures classes) are often divided into sub-images on a 3 by 3 grid (typically 128 by 128 to 215 by 215 pixels depending on the class image source) creating a dataset with 999 texture samples. Some of the textures (e.g. D23 and D27) appear to be pictures of the same texture at different spatial scales and some (e.g. D101 and D102) simply appear to be negatives of one another. Models that perform well on this dataset do not necessarily demonstrate scale invariant texture recognition (and are in fact discouraged to do so due to textures D23 and D27), yet this has been a popular dataset for many computational

[1] The Brodatz dataset is currently available for download at http://www.ux.uis.no/∼tranden/brodatz.html.

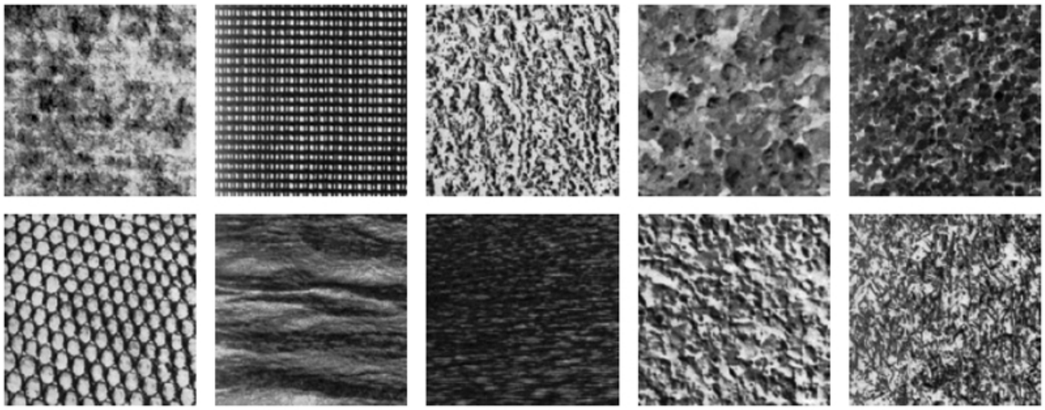


Figure 4.5: Samples from the Brodatz Dataset.
Ten of the 111 textures photographs in the Brodatz Dataset.

texture classification experiments.

4.3.2 CUReT Dataset

The CUReT Dataset[2] is a large dataset of color images of 61 physical samples (texture classes) photographed under carefully controlled lighting conditions. These textures were photographed 205 times (texture samples) from the same focal length at various viewing angles and illumination directions. Due to the large azimuthal angles for many of these texture samples, the dataset is often refined to retain only 92 of the 205 texture samples. Also, due to these images consisting of both texture samples in the foreground and a black background in the periphery, the samples are often cropped from the original 640 by 480 pixels to the central 200 by 200 pixels[3]. However, this dataset is limited only to in-plane rotations and by the physical texture samples chosen.

4.3.3 UIUCTex Dataset

The UIUCTex Dataset[4] consists of 25 physical textures (e.g. the bark of a tree, a brick pathway, a patterned cloth, etc.) photographed 40 times at largely varying viewing angles and focal lengths. This dataset of 1000 640 by 480 pixel texture samples (see Appendix B for a complete list of all of the samples) is unique and particularly challenging

[2] The CUReT dataset is currently available for download at http://www1.cs.columbia.edu/CAVE/exclude/curet/dataComp/.

[3] The cropped CUReT dataset is currently available for download at http://www.robots.ox.ac.uk/∼vgg/research/texclass/data/curetcol.zip.

[4] The UIUCTex dataset is currently available for download at http://www.cs.unc.edu/∼lazebnik/research/uiuc_texture_dataset.zip.

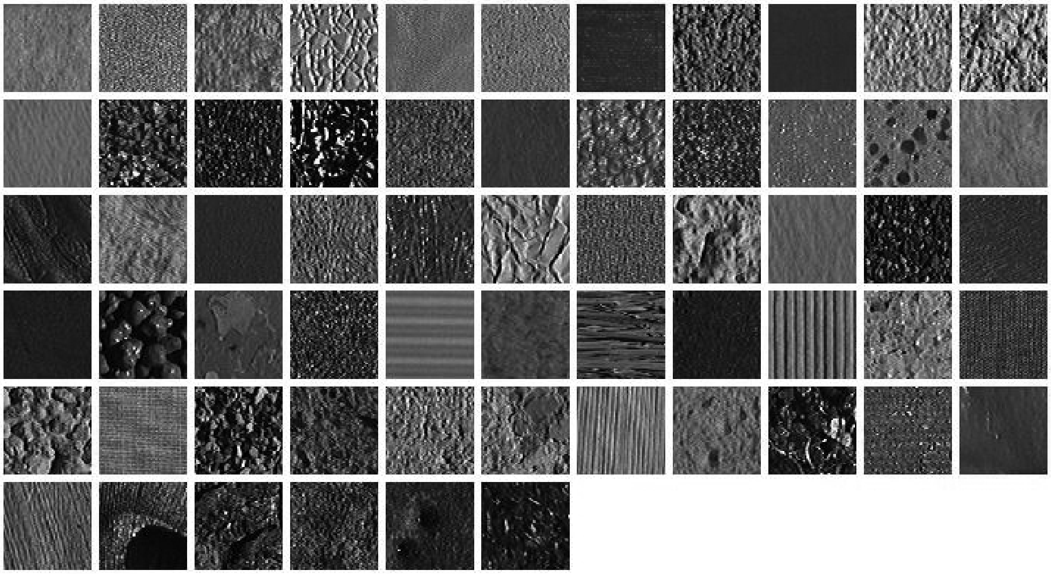


Figure 4.6: Samples from the CUReT Dataset.
One texture sample from each of the 61 textures in the cropped CUReT dataset.

due to the size (307.2 million pixels compared to 224 million in the cropped CUReT dataset and 48 million in the Brodatz dataset), the non-rigid deformations (fabrics draped over staircases), and the large range of focal lengths. Consequently, there is a much larger interclass and intraclass variation than the other datasets and in some recent studies it has been avoided, in favor of the less complex CUReT and Brodatz datasets, on the grounds that “the task of classification in [the UIUCTex] database is beyond the capability of the proposed method” [Yin et al., 2009].

4.4 UCSD Natural Scenes Dataset

The UCSD Natural Scenes Dataset consists of 100 8 megapixel (3264 by 2448 pixel) grayscale photographs (see Appendix A for a complete list of all of the photographs) of scenery we took with a Canon digital camera while walking around the UCSD campus. These images are not labeled with segmented textures, making them inappropriate as stimuli for a texture classification task, but are a large corpus on which to train a vision system prior to the classification task.

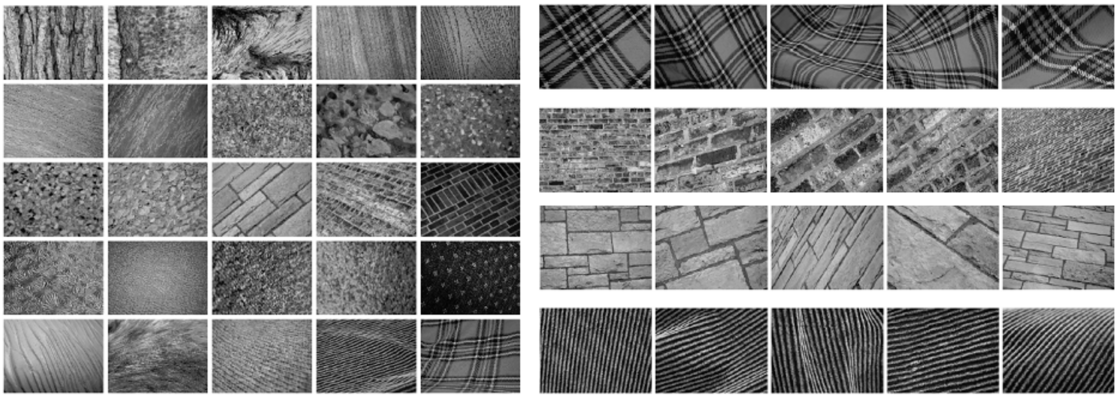


Figure 4.7: Samples from the UIUCTex Dataset.
On the left are single textures samples from each of the 25 textures to illustrate interclass variance. On the right are five texture samples from four of the textures to illustrate intraclass variance.

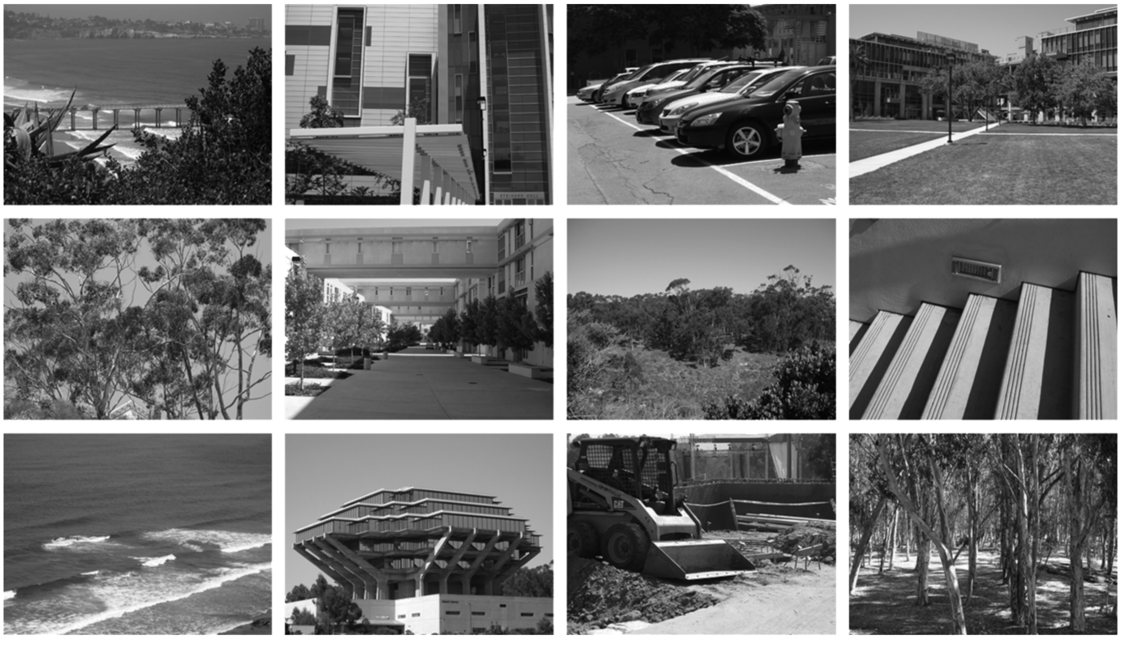


Figure 4.8: Samples from the UCSD Natural Scenes Dataset.
Twelve of the 100 photographs of scenery around the UCSD campus that compose the dataset of unlabeled natural scenes.

4.5 Computational Model

Traditionally, computer vision algorithms attempting to be biologically plausible models of mammalian cortical vision processing have focused on modeling primary visual cortex responses. This is largely because of the wealth of literature on the primary visual cortex dating back to the original Hubel and Wiesel experiments [Hubel & Wiesel, 1962]. These linear models (see Section 3.3) typically consist of a Gabor [Gabor, 1946; Daugman, 1985] filter bank (sinusoids at various orientations and spatial frequencies modulated by a Gaussian filter envelope), which predicts the response of striate cortex layer 4 pyramidal cells to these ideal synthetic stimuli (drifting gratings) fairly well (about 84% of the variance in the neuronal responses is explained). However, when subjects are presented with natural stimuli (e.g. a video of what a cat sees walking through the grass), these models of primary visual cortical neurons perform poorly (only about 21% of the variance in the neuronal responses is explained). This substantial decrease in predictive power with natural stimuli, as opposed to synthetic stimuli, could be explained by the complex nonlinearities that are prevalent in all stages of the visual processing (including as early as the retina [Enroth-Cugell & Lennie, 1975]). Many attempts (beginning with David Heeger's divisive contrast adaptation [Heeger, 1992; Carandini & Heeger, 1994]) have been made to incorporate some of these nonlinearities into the models and improve their predictive power of neuronal responses, but even advanced nonlinear system identification (NLSI) approaches can only explain about 40% of the variance in striate cortex pyramidal cell responses to natural stimuli [David & Gallant, 2005].

Alternatively, predicting responses of neurons in earlier stages of the visual processing stream does not appear to be prone to the same problems. Early models of the lateral geniculate nucleus (LGN) typically consist of an opposing pair (on-center and off-center) of isotropic difference of Gaussians (DoG) or Laplacian of the Gaussian (LoG) linear filters. These linear models predict the response of parvocellular and magnocellular LGN neurons to ideal synthetic stimuli reasonably well (about 78% of the variance is explained), but fail to predict their response to more complex natural stimuli (about 46% of the variance is explained) [Mante, 2005]. Unlike nonlinear models of V1, though, nonlinear models of LGN, which include luminance and contrast adaptation, are much more robust to the complexity of the stimuli, suggesting that most of the nonlinearities are accounted. Nonlinear models of LGN can explain up to 96% of the variance in responses to ideal synthetic stimuli and also explain about 78% of the variance in the responses

Table 4.1: Explained variance of neuron responses with linear and nonlinear models.

	Linear Model		**Nonlinear Model**	
Stimuli:	Synthetic	Natural	Synthetic	Natural
V1:	84%	21%	93%	40%
LGN:	78%	39 - 46%	84 – 96%	78%

Linear models of striate cortex (V1) typically involve convolving the stimulus with a receptive field filter determined by spike triggered averaging, followed by a Poisson process spike train generator. When presented with ideal synthetic stimuli (drifting gratings), these models can explain about 84% of the variance in the spiking neuronal response [Carandini et al., 1997]. However, when presented with more complex natural scene stimuli, these models can only explain about 21% of the variance [David & Gallant, 2005]. Nonlinear models of striate cortex do not perform much better (93% of the variance is explained with synthetic stimuli [Carandini et al., 1997] and about 40% of the variance is explained with natural stimuli [David & Gallant, 2005]). This suggests that nonlinear approaches to modeling the striate cortex are still incomplete. The lateral geniculate nucleus (LGN), however, appears to be modeled far more robustly. Linear models of LGN also fail to explain responses to natural stimuli (78% of the variance is explained with synthetic stimuli and up to 46% of the variance is explained with natural stimuli [Mante, 2005]), but nonlinear models, which include luminance and contrast adaptation, recover this loss in predictive power (up to 96% of the variance is explained with synthetic stimuli and about 78% of the variance is explained with natural stimuli [Mante, 2005]).

to complex natural scene stimuli. This convincing evidence suggests that a hierarchical computational model of the mammalian visual system should begin with and build upon a model proved to explain biological phenomena, the LGN nonlinear model.

4.5.1 Retinothalamic Model

We begin our hierarchical computational model of the mammalian visual system for texture recognition by modeling the transfer function between the stimulus presented to the eye and the response observed in the LGN [Bonin et al., 2005]. One could independently model the retina and the thalamus; however, the evidence presented in recent experiments (see Table 4.1) suggests that the retinothalamic transfer function can be modeled sufficiently accurately and robustly with a nonlinear LGN model (see Figure 4.9). This model consists of five stages: receptive field linear filtering, local luminance adaptation, local contrast adaptation, surround suppression, and relative firing rate activation.

The receptive fields of parvocellular LGN neurons are well characterized by a pair of opposing difference of Gaussian (DoG) filters:

$$\begin{aligned}
\mathbf{G}(\mathbf{x},\mathbf{y}) &= \frac{1}{2\pi\sigma^2} e^{-\mathbf{x}^2-\mathbf{y}^2/2\sigma^2}, && (4.1)\\
\mathbf{M}\left(\mathbf{F}(\mathbf{x},\mathbf{y})\right) &= \mathbf{F}(\mathbf{x},\mathbf{y}) - \frac{1}{n(\mathbf{x})n(\mathbf{y})} \sum_{\mathbf{x}} \sum_{\mathbf{y}} \mathbf{F}(\mathbf{x},\mathbf{y}), && (4.2)\\
\mathbf{N}\left(\mathbf{F}(\mathbf{x},\mathbf{y})\right) &= \frac{\mathbf{F}(\mathbf{x},\mathbf{y})}{\|\mathbf{F}(\mathbf{x},\mathbf{y})\|_2}, && (4.3)\\
\widehat{\mathbf{D}}_{k=on}(\mathbf{x},\mathbf{y}) &= \mathbf{G}(\mathbf{x},\mathbf{y},\sigma_{center}) - \beta\mathbf{G}(\mathbf{x},\mathbf{y},\sigma_{annulus}), && (4.4)\\
\mathbf{D}_k(\mathbf{x},\mathbf{y}) &= \mathbf{N}\left(\mathbf{M}\left(\widehat{\mathbf{D}}_k(\mathbf{x},\mathbf{y})\right)\right), && (4.5)\\
\mathbf{D}_{k=off}(\mathbf{x},\mathbf{y}) &= -\mathbf{D}_{k=on}(\mathbf{x},\mathbf{y}), && (4.6)\\
\mathbf{R}_k^{LGN}(\mathbf{i},\mathbf{j}) &= \mathbf{D}_k^{LGN}(\mathbf{x},\mathbf{y}) \circ \mathbf{X}^{LGN}(\mathbf{i},\mathbf{j}), && (4.7)
\end{aligned}$$

where $\mathbf{x}$ and $\mathbf{y}$ index the filter, $\mathbf{i}$ and $\mathbf{j}$ index the input stimulus or output response, $\mathbf{G}(\bullet)$ is the two-dimensional isotropic Gaussian with standard deviation σ, $\mathbf{M}(\bullet)$ enforces feedfoward balanced excitation and inhibition (by having a mean of zero), $n(\bullet)$ is the cardinality of its argument, $\mathbf{N}(\bullet)$ L_2 normalizes the filter, $\mathbf{D}_{on}(\bullet)$ and $\mathbf{D}_{off}(\bullet)$ are the on-center and off-center DoG balanced and normalized filters with β annulus strengths, and $\mathbf{R}(\bullet)$ is the two-dimensional correlation of the DoG filter with the stimulus, $\mathbf{X}(\bullet)$.

The responses from these filters are divisively suppressed by the local luminance in the stimulus:

$$\mathbf{L}_k^{LGN}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{R}_k^{LGN}(\mathbf{i},\mathbf{j})}{\mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{luminance}^{LGN}) \circ \mathbf{X}(\mathbf{i},\mathbf{j})}. \qquad (4.8)$$

The luminance normalized response is then divisively suppressed by the local contrast in the luminance normalized response cross filters (the statistical independence of luminance and contrast observed in natural stimuli allows us to cascade these operations safely [Mante et al., 2005]):

$$\mathbf{C}_k^{LGN}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{L}_k^{LGN}(\mathbf{i},\mathbf{j})}{\sqrt{\sum_{k'=k} \mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{contrast}^{LGN}) \circ \mathbf{L}_{k'}^{LGN}(\mathbf{i},\mathbf{j})^2}}. \tag{4.9}$$

The luminance and contrast normalized response is then activated to compute the relative (to spontaneous) firing rate (RFR) of the k^{th} cortical feature column, $\mathbf{Y}_k(\mathbf{i},\mathbf{j})$:

$$\begin{aligned} \mathbf{A}\left(\mathbf{F}(\mathbf{x},\mathbf{y})\right) &= \frac{1.1}{1+e^{-8\mathbf{F}(\mathbf{x},\mathbf{y})+2.30258}} - 0.1, & (4.10)\\ \mathbf{Y}_k^{LGN}(\mathbf{i},\mathbf{j}) &= \mathbf{A}\left(\mathbf{C}_k^{LGN}(\mathbf{i},\mathbf{j})\right), & (4.11) \end{aligned}$$

where $\mathbf{A}(\bullet)$ is a sigmoidal mapping of no response, $\mathbf{F}(\bullet) = 0$, to 0 (spontaneous firing rate), a large positive response to 1 (maximally excited firing rate), and a large negative response to $-.1$ (maximally suppressed firing rate, or not spiking).

4.5.2 Striate Visual Cortex Model

The second stage of our computational vision model, the striate visual cortex (V1) input layer (4Cβ) model (see Figure 4.11), is implemented similarly to our LGN nonlinear model. Unlike the LGN model, though, the filter weights are not parametrically defined, because doing so would render the LGN model unnecessary and could lead to issues of failing to explain the variance of the V1L4Cβ pyramidal response to natural scene stimuli observed in biology (see Section 4.1). Instead these weights are learned in an unsupervised fashion from natural scene stimuli using a variation of the BCM rule [Bienenstock et al., 1982]. The BCM learning rule is derived from the classical Hebbian rule [Hebb, 1949], but introduces a modification threshold, θ, to address the problems of weight instability prevalent in Hebbian learning. The BCM rule is also supported by over twenty years of neuroscience validation in multiple modalities [Shouval et al., 1997; Castellani et al., 2001; Yeung et al., 2004] and was shown to be, with a few reasonable assumptions, a firing rate analog to spike-timing dependent plasticity (STDP) [Izhikevich & Desai, 2003]. A variation of BCM learning, which is designed to model early synaptic development, is ABS learning, which models online learning and does not require the neuron to fire before the synaptic weights are updated [Artola & Singer, 1987, 1993]. BCM learning, in the

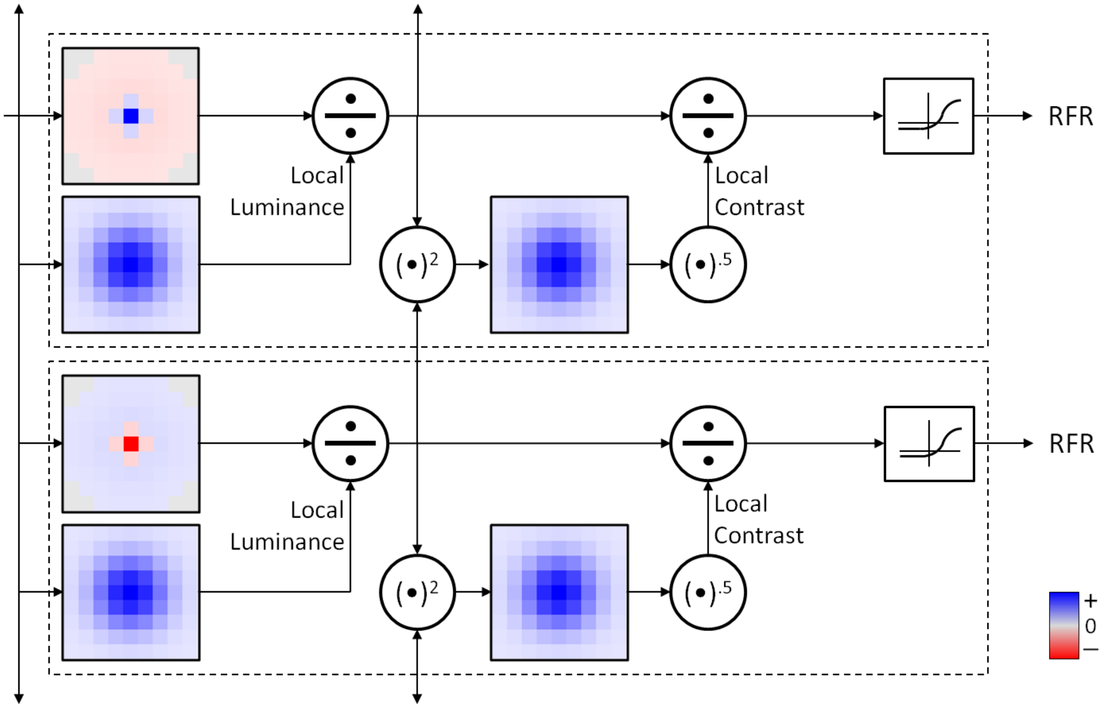


Figure 4.9: Lateral geniculate nucleus thalamic feature column model with luminance and contrast adaptation.
Each dashed box represents the nonlinear model for a single thalamic feature column type (on-center in the upper half of the figure and off-center in the bottom half of the figure), which is distributed retinotopically over the stimulus. The stimulus is linearly filtered by the DoG filter synaptic weights, divisively normalized by the local luminance, divisively normalized by the local contrast, and activated with a sigmoidal mapping to calculate the feature column's relative firing rate.

context of our model, can be described with the following equations:

$$\begin{aligned}
\mathbf{X}^{V1L4}(\mathbf{i},\mathbf{j}) &= \mathbf{Y}_n^{LGN}(\mathbf{i},\mathbf{j}), && (4.12)\\
\mathbf{R}_k^{V1L4}(\mathbf{i},\mathbf{j}) &= \mathbf{W}_k^{V1L4}(\mathbf{x},\mathbf{y}) \circ \mathbf{X}^{V1L4}(\mathbf{i},\mathbf{j}), && (4.13)\\
\theta_k &= \left\langle \mathbf{R}_k^{V1L4}(\mathbf{x},\mathbf{y})_k^2 \right\rangle, && (4.14)\\
\triangle\mathbf{W}_k^{V1L4}(\mathbf{x},\mathbf{y}) &= \alpha\mathbf{R}_k^{V1L4}(\mathbf{x},\mathbf{y})\left(\mathbf{R}_k^{V1L4}(\mathbf{x},\mathbf{y}) - \theta_k\right)\frac{\mathbf{X}^{V1L4}(\mathbf{x},\mathbf{y})}{\theta_k}, && (4.15)
\end{aligned}$$

where $\mathbf{x}$ and $\mathbf{y}$ index the filter or winning input or response patch, $\mathbf{i}$ and $\mathbf{j}$ index the input or response, $\mathbf{X}(\bullet)$ is now the input to the V1 model from LGN model's n^{th} thalamic feature column activation, $\mathbf{R}(\bullet)$ is the two-dimensional correlation of the k^{th} cortical feature column's weights, $\mathbf{W}_k(\bullet)$, with the input, θ_k is the modification threshold of the k^{th} cortical feature column, $\langle\bullet\rangle$ denotes a time average over the history of the cortical feature column's responses, and α is the learning rate. The weights are initialized randomly with a uniform distribution, only the most responsive cortical feature column has its weights updated with each presentation during learning, and they are balanced with $\mathbf{M}(\bullet)$, and normalized with $\mathbf{N}(\bullet)$ after each update. BCM learning, though, does not prevent one cortical feature column's weights from dominating and preventing other cortical feature columns from updating. To prevent cortical feature columns from never winning, we implemented a solution (similar to the rebalancing in [Pinto et al., 2009]) that reinitializes the weights of the least winning filters to those of the most winning filters with additive noise to force competition in frequently observed regions of the input space. Adjusting these weights with unsupervised learning on the UCSD Natural Scenes Dataset until they are stable ($\theta_k = 1 - \delta, \delta \to 0$) yields filters qualitatively similar to the receptive fields found in biology (see figure 4.10).

The luminance and contrast adaptation nonlinearities evident in LGN are also found in the responses measured in the input layers of primary visual cortex, namely V1 layer 4C [Rust & Movshon, 2005]. Additionally, these neurons are characterized by an extraclassical receptive field. Although stimuli presented in a neuron's extraclassical receptive field cannot drive the neuron alone, they certainly modulate its response (see Section 3.4) [Carandini et al., 2005]. This physiological phenomenon is implemented with the luminance and contrast normalized response being divisively suppressed by the

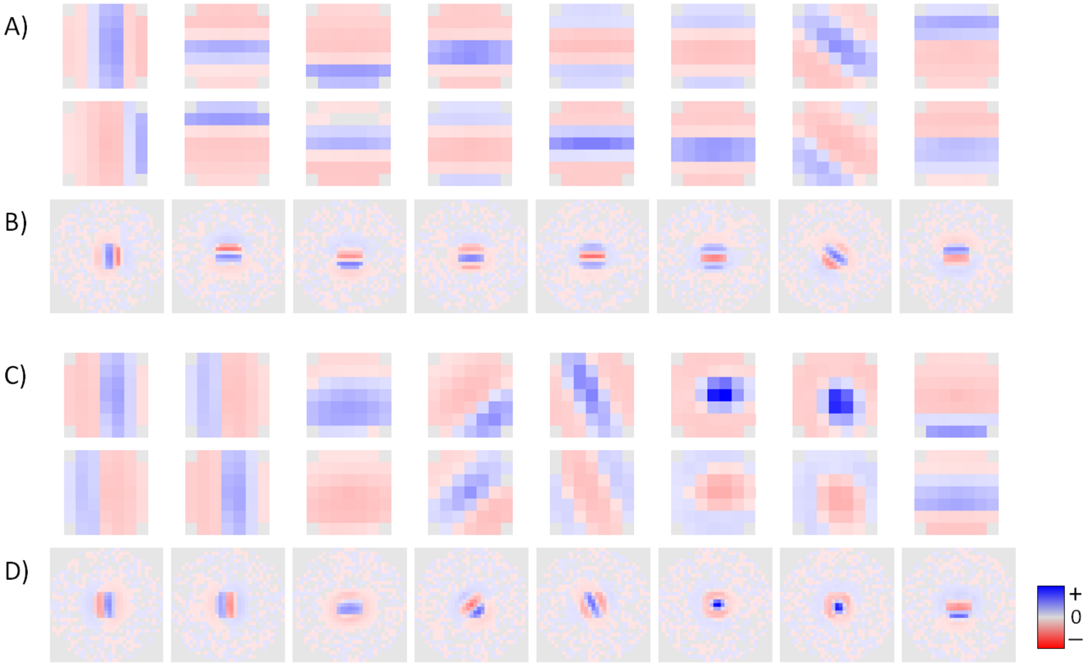


Figure 4.10: A subset of primary visual cortex feature column input layer 4Cβ synaptic weights and their linear reconstructions.
A) and **C)** Synaptic weights from LGN on-center (top row) and off-center (bottom row) feature columns to 16, 8 in A) and 8 in B), of the 100 V1 feature columns' layer 4Cβ are displayed. **B)** and **D)** Reverse correlation (the non-spiking analog of spike-triggered averaging) is performed with random stimuli presented through the LGN nonlinearies and the synaptic weight filters. The reconstructed linear components (receptive fields), shown below each of the corresponding synaptic weights, exhibit properties often observed in V1 layer 4Cβ recordings: balanced inhibition, small variations in spatial frequency, most with strong orientation selectivity, and some with almost isotropic receptive fields.

activation of other surround responses, which is then activated similarly to the LGN model:

$$\mathbf{L}_k^{V1L4}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{R}_k^{V1L4}(\mathbf{i},\mathbf{j})}{\mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{luminance}^{V1L4}) \circ \mathbf{X}^{V1L4}(\mathbf{i},\mathbf{j})}, \tag{4.16}$$

$$\mathbf{C}_k^{V1L4}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{L}_k^{V1L4}(\mathbf{i},\mathbf{j})}{\sqrt{\sum_{k'=k} \mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{contrast}^{V1L4}) \circ \mathbf{L}_{k'}^{V1L4}(\mathbf{i},\mathbf{j})^2}}, \tag{4.17}$$

$$\mathbf{V}_k^{V1L4}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{C}_k^{V1L4}(\mathbf{i},\mathbf{j})}{\mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{surround}^{V1L4}) \circ \mathbf{Y}_k^{V1L4}(\mathbf{i},\mathbf{j})}, \tag{4.18}$$

$$\mathbf{Y}_k^{V1L4}(\mathbf{i},\mathbf{j}) = \mathbf{A}(\mathbf{V}_k^{V1L4}(\mathbf{i},\mathbf{j})). \tag{4.19}$$

It should be noted that this recurrent network topology of activating the column k, followed by applying the surround suppression using that activation, can be applied iteratively; however, in practice, the model appears to approach sufficiently close to a stable state after only one iteration.

The output layer of the V1 model introduces local spatial and slight orientation invariance and consists of a maximum operation over a small neighborhood within each column. This operation produces responses similar to those of complex cells observed in V1 (the excitatory and inhibitory regions of the receptive field are no longer distinct, orientation selectivity is largely still prevalent, and spatial selectivity is relaxed) and is again followed by luminance and contrast adaptation nonlinearities and surround suppression:

$$\mathbf{X}_k^{V1L3}(\mathbf{i},\mathbf{j}) = \mathbf{Y}_k^{V1L4}(\mathbf{i},\mathbf{j}), \tag{4.20}$$

$$\mathbf{R}_k^{V1L3}(\mathbf{i},\mathbf{j}) = \underset{\mathbf{i}',\mathbf{j}'}{max}\left(\mathbf{X}_k^{V1L3}(\mathbf{i},\mathbf{j})\right), \tag{4.21}$$

$$\mathbf{L}_k^{V1L3}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{R}_k^{V1L3}(\mathbf{i},\mathbf{j})}{\mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{luminance}) \circ \mathbf{X}_k^{V1L3}(\mathbf{i},\mathbf{j})}, \tag{4.22}$$

$$\mathbf{C}_k^{V1L3}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{L}_k^{V1L3}(\mathbf{i},\mathbf{j})}{\sqrt{\sum_{k'=k} \mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{contrast}) \circ \mathbf{L}_{k'}^{V1L3}(\mathbf{i},\mathbf{j})^2}}, \tag{4.23}$$

$$\mathbf{V}_k^{V1L3}(\mathbf{i},\mathbf{j}) = \frac{\mathbf{C}_k^{V1L3}(\mathbf{i},\mathbf{j})}{\mathbf{G}(\mathbf{x},\mathbf{y};\sigma_{surround}) \circ \mathbf{Y}_k^{V1L3}(\mathbf{i},\mathbf{j})}, \tag{4.24}$$

$$\mathbf{Y}_k^{V1L3}(\mathbf{i},\mathbf{j}) = \mathbf{A}\left(\mathbf{V}_k^{V1L3}(\mathbf{i},\mathbf{j})\right), \tag{4.25}$$

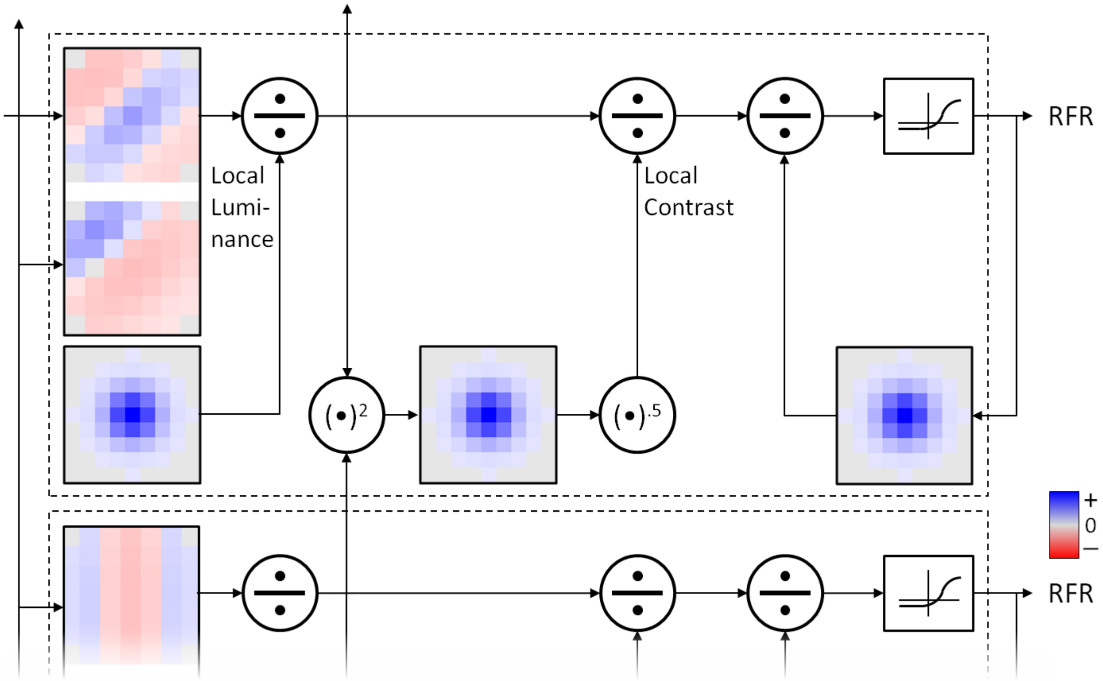


Figure 4.11: Primary visual cortex feature column input layer 4Cβ model with surround nonlinearities.
Each dashed box represents the nonlinear model for a single cortical feature column type, which is distributed retinotopically over the stimulus. The stimulus is linearly filtered by the synaptic weights (see Figure 4.10 A), divisively normalized by the local luminance, divisively normalized by the local contrast, divisively suppressed by the surround activation and activated with a sigmoidal mapping to calculate the feature column's relative firing rate.

where $\mathbf{i}'$ and $\mathbf{j}'$ index regions in the neighborhood of $\mathbf{Y}_k(\mathbf{i}, \mathbf{j})$.

4.5.3 Extrastriate Visual Cortex Model

Secondary visual cortex (V2) differs anatomically from primary visual cortex (V1) in that each cytochrome oxidase stripe appears to have a full retinotopic mapping (compared with the single retinotopic mapping in V1) [Sereno et al., 1995; Sincich & Horton, 2002, 2005]. However, physiologically, similar functionality has been identified in both areas. For example, V2 surround mechanisms appear to be the same as in V1, but spatially scaled up by a factor of 2 [Shushruth et al., 2009]. Again, we rely on unsupervised learning constrained by the statistics of natural scene stimuli to develop the feedforward linear weights between a downsampled (by a factor of 2) activation in the output layer

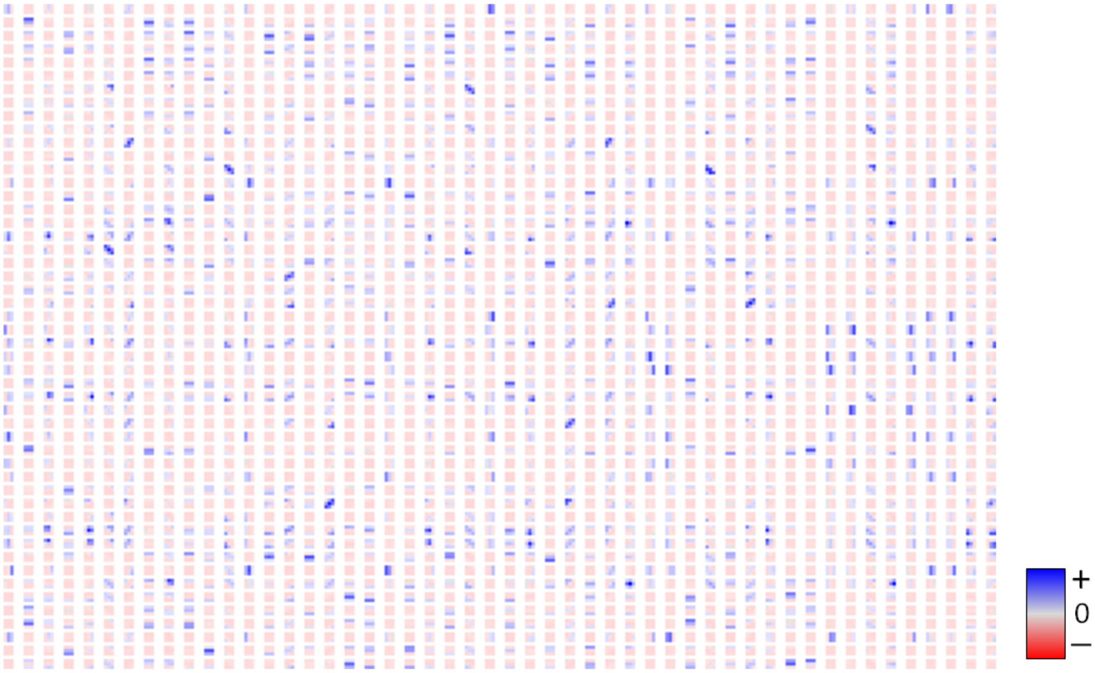


Figure 4.12: A subset of secondary visual cortex layer 4 synaptic weights. Synaptic weights from 50 of the 100 V1 cortical feature columns' layer 3 to 50 of the 250 V2 cortical features columns' layer 4. Consistently with observations in neuroscience [Carandini et al., 2005], the projections appear to be implementing sparse linear combinations of V1 responses over a small neighborhood.

of the V1 model, $\mathbf{Z}_k(\mathbf{i}, \mathbf{j})$, and in the input layer of the V2 model, followed by the same surround nonlinearities (see Figure 4.11) and maximum pooling operation. Although the physiology of V2 is not currently very well defined, partially because ideal synthetic stimuli to drive these neurons are difficult to construct and test, it does appear to be implementing sparse linear combinations of responses in V1 over a small neighborhood [Carandini et al., 2005]. Our unsupervised learning method results in exactly these types of projections (see Figure 4.12).

4.5.4 Examples of the Computational Early Visual Processing Model

In the following figures we demonstrate the feedforward flow of information through the computational model of the early mammalian visual processing stream given natural scene stimuli after unsupervised learning from the UCSD Natural Scenes Dataset (see Section 4.4 and Appendix A). Each figure displays a single LGN feature column (either on-center or off-center), a single V1 feature column (of the 100 V1 feature columns trained), and a single V2 cortical feature column (of the 250 feature columns trained).

The stimulus is first processed sequentially with a linear filter for the LGN feature column (upper response in the LGN feature column of each figure) and with luminance and contrast adaptation (lower response in the LGN feature column of each figure, see Figure 4.9 for details). The activation of the LGN feature column is then processed with a linear filter for the V1 feature column (upper left response in the V1 feature column of each figure), with luminance and contrast adaptation and surround suppression (lower left response in the V1 feature column of each figure, see Figure 4.11 for details), with a maximum pooling operation (upper right response in the V1 feature column of each figure), and again with luminance and contrast adaptation and surround suppression (lower right response in the V1 feature column of each figure). The activation of the V1 feature column is then downsampled and processed in the V2 feature column identically to the V1 feature column. Clipping (the response spans a subset of the stimulus with a gap around the edges) can be observed in each subsequent stage of the model and is implemented to avoid computing responses from incomplete feedforward projections. Downsampling can be observed in the increasing granularity of the pixelations in each subsequent stage of the model. Note the importance of the surround nonlinearities in imposing sparsity in the output (i.e. the activation pattern in V1L3 and V2L3 is much more dense before the surround nonlinearities than after) and sparsity in the learned afferent projection (see Figure 4.12), which is critical for efficient neural coding [Olshausen & Field, 2004] .

4.6 Psychophysical Experiment

We hypothesized that human texture classification accuracy will significantly outperform state-of-the-art computational vision models in this task. To investigate this hypothesis, we constructed a psychophysical experiment and data collection instrument (see Figure 4.16) to closely replicate the computational texture classification experimental paradigm (using the UIUCTex Dataset, see Section 4.3.3 and Appendix B) and obtained UCSD Institutional Review Board (IRB) approval for human subject testing (project proposal #090618). Human subjects (N=22 male; N=8 female; UCSD graduate students and associates) were recruited and the experiment was conducted in the same room under fluorescent lighting, on the same computer (Dell Precision T7400), and the same monitor (Dell E248WFP; 400 candelas/meter^2 brightness; 1,000:1 contrast ratio).

The experiment consisted of asking the human subjects to match new test texture samples to training texture samples they had already been presented in an Adobe Flash

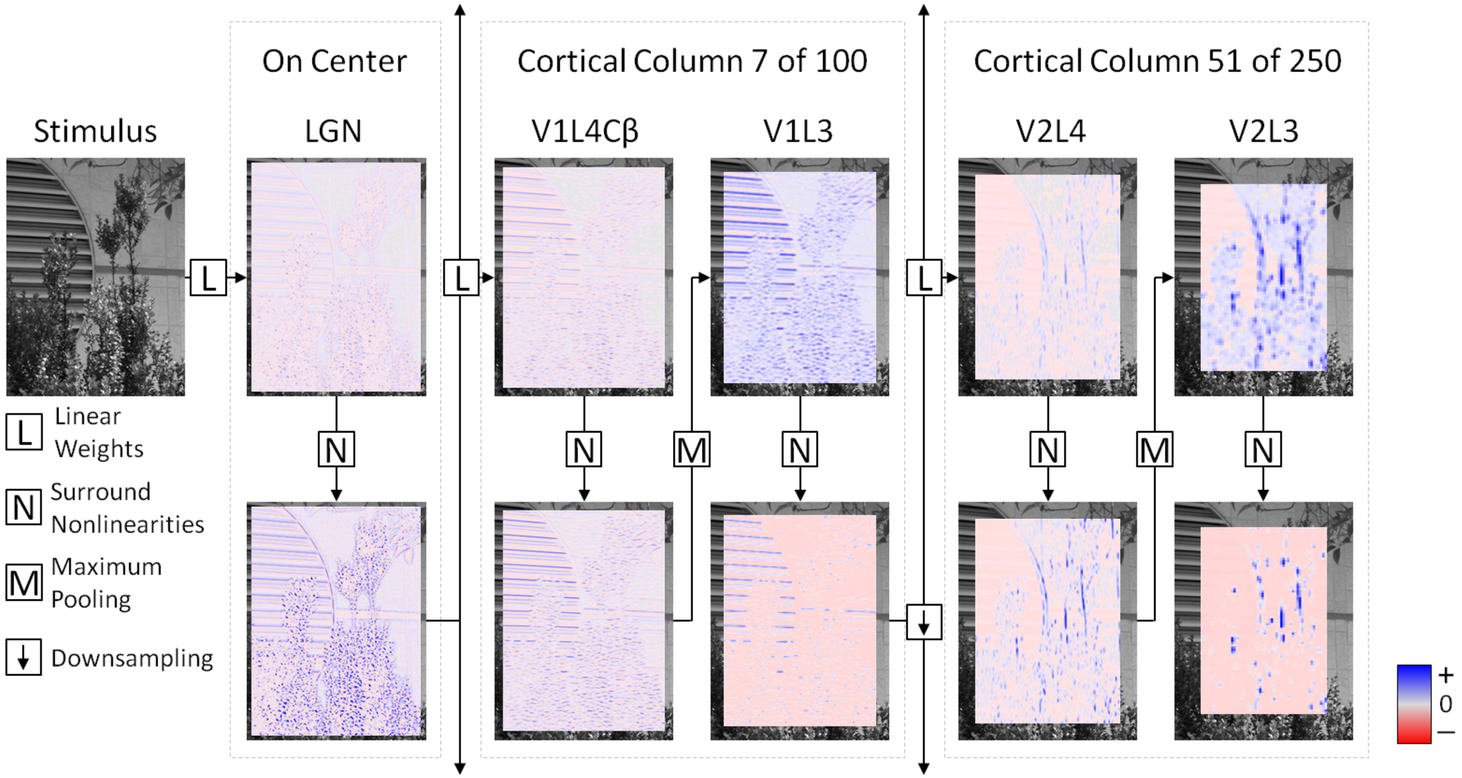


Figure 4.13: Information flow through the early visual processing model after unsupervised learning (example 1).
A natural scene stimulus from the UCSD Natural Scenes Data (in this case it is the upper left quadrant of the IMG_0039.jpg photograph, see Appendix A) is presented to the on-center feature column of the LGN model. Both LGN feature columns are then presented to feature column 7 of the trained 100 V1 columns (which learned selectivity for horizontal edges), processed with luminance and contrast adaptation and surround suppression (see Figure 4.11) in the input layer V1L4Cβ, pooled with the maximum operator into the output layer V1L3 and reprocessed with luminance and contrast adaptation and surround suppression. All 100 V1 feature columns are then downsampled and presented to feature column 51 of the trained 250 V2 columns (which learned selectivity for vertical dark stripes with end-stopping).

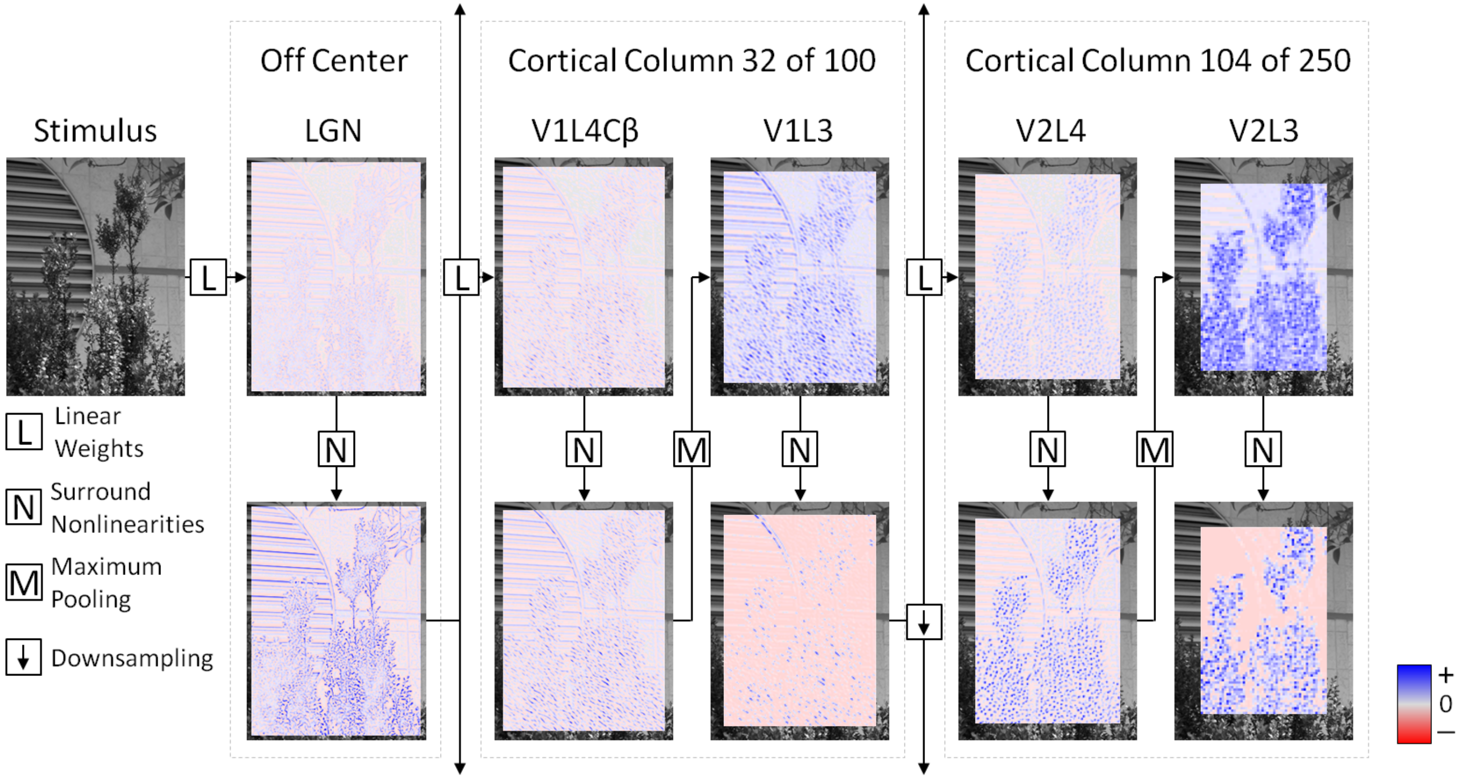


Figure 4.14: Information flow through the early visual processing model after unsupervised learning (example 2). The same stimulus used in Figure 4.13 is now presented to the off-center feature column of the LGN model, to feature column 32 of the V1 model (which learned selectivity for diagonal upper left to lower right edges), and to feature column 104 of the V2 model (which learned selectivity for high frequency and dense blobs like leaves).

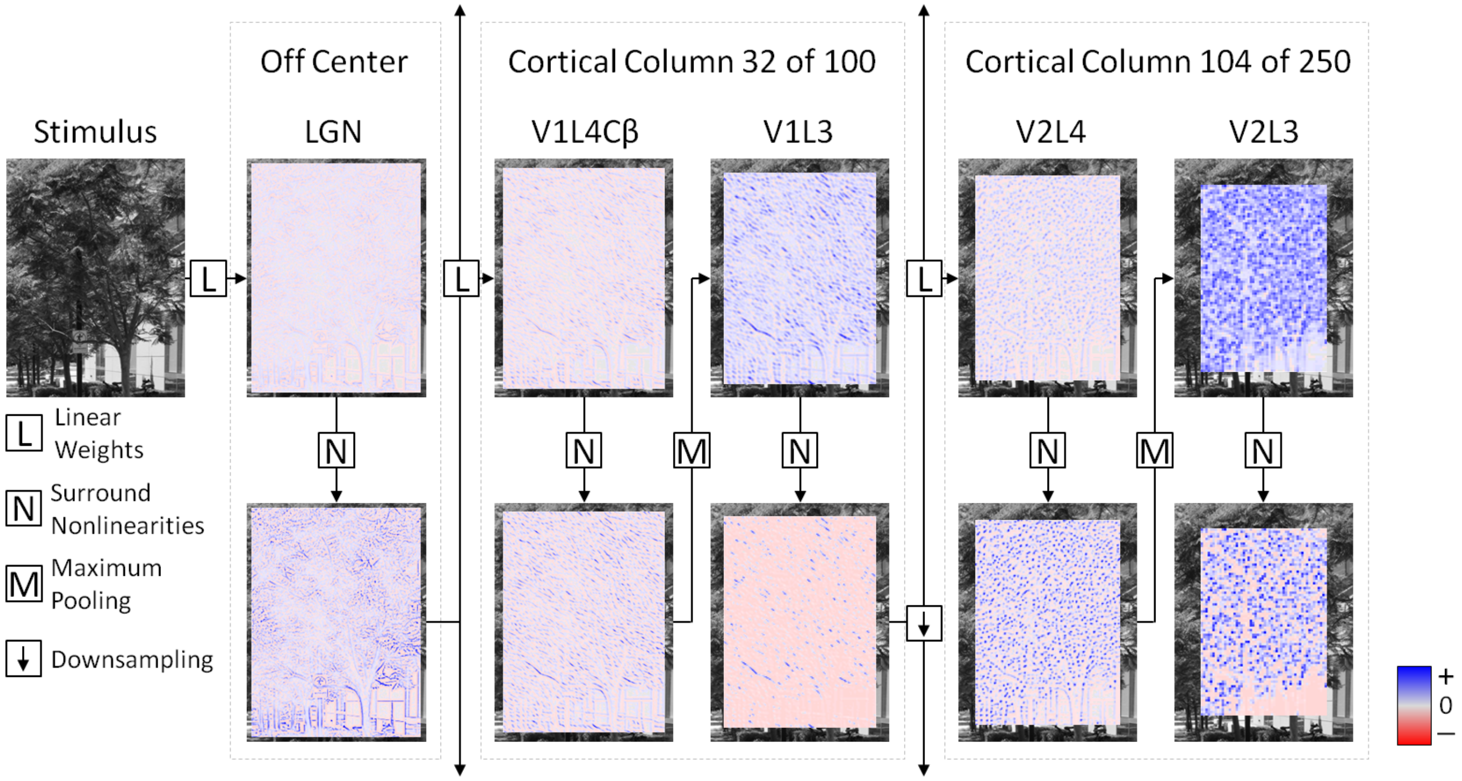
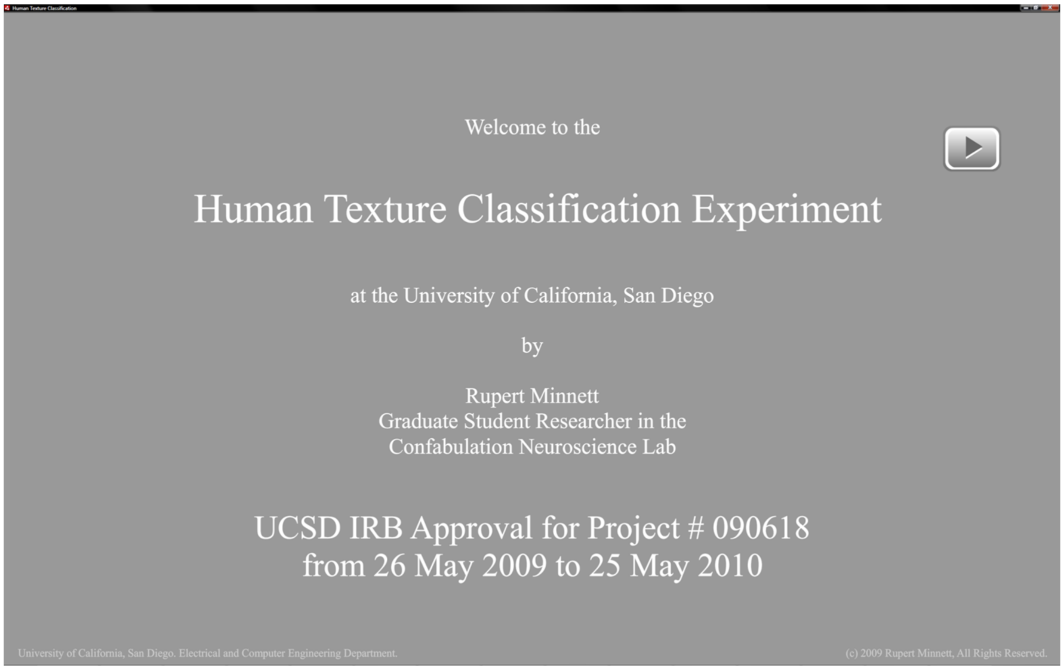


Figure 4.15: Information flow through the early visual processing model after unsupervised learning (example 3). The same feature columns used in Figure 4.14 are now presented with a new stimulus from the UCSD Natural Scenes Data (the upper right quadrant of the IMG_0069.jpg photograph, see Appendix A).



Welcome to the

Human Texture Classification Experiment

at the University of California, San Diego

by

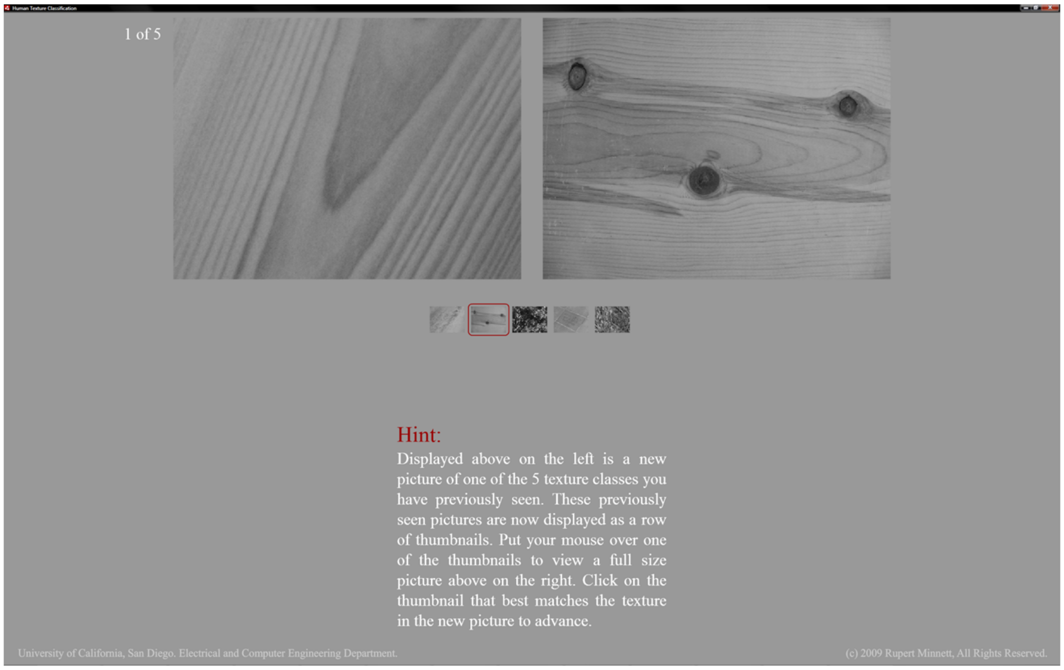
Rupert Minnett
Graduate Student Researcher in the
Confabulation Neuroscience Lab

UCSD IRB Approval for Project # 090618
from 26 May 2009 to 25 May 2010

University of California, San Diego. Electrical and Computer Engineering Department.

(c) 2009 Rupert Minnett, All Rights Reserved.

Figure 4.16: The human texture classification data collection instrument welcome page.



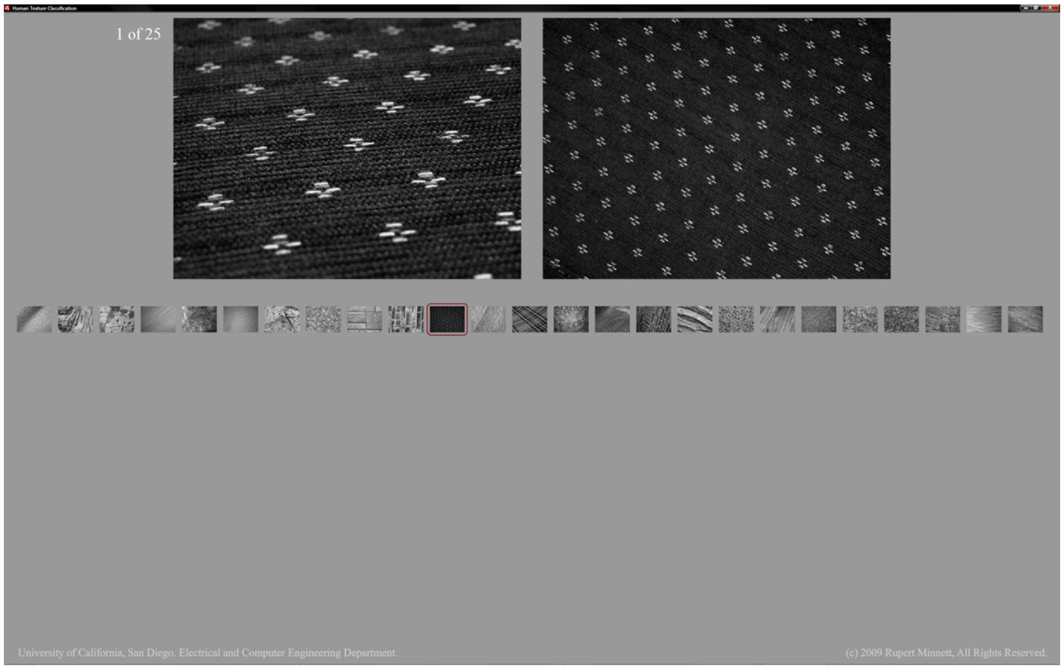
Hint:

Displayed above on the left is a new picture of one of the 5 texture classes you have previously seen. These previously seen pictures are now displayed as a row of thumbnails. Put your mouse over one of the thumbnails to view a full size picture above on the right. Click on the thumbnail that best matches the texture in the new picture to advance.

University of California, San Diego. Electrical and Computer Engineering Department.

(c) 2009 Rupert Minnett, All Rights Reserved.

Figure 4.17: The human texture classification data collection instruction pages.



University of California, San Diego. Electrical and Computer Engineering Department.

(c) 2009 Rupert Minnett, All Rights Reserved.

Figure 4.18: The human texture classification data collection instrument during the first presentation of the first round.

application we developed. To familiarize the subjects with the data collection instrument, example textures we collect with a Canon digital camera (see Figure C) were presented along with instructions for using the interface and performing the texture classification task (see Figure 4.17). Human subjects, unlike computers, whose memory can be erased and the experiment repeated over multiple trials, could not be tested on texture samples they had already been presented. To address this, we began the experiment by presenting one texture sample (chosen randomly) from each of the 25 classes (presented in a random order) in the UIUCTex Dataset to the subject, forming the first training set. Subsequently, a new texture sample (chosen randomly) from each of the 25 classes (presented in a random order) was presented and the subject was asked to click on the previously presented texture they perceived it matched (see Figure 4.18). These texture samples were then treated as training samples and the testing continued with two training samples per class. This process was repeated until testing was complete with ten training samples per class (see Figure 4.19) and took between roughly 23 and 46 (mean of 34) minutes for the subjects to complete.

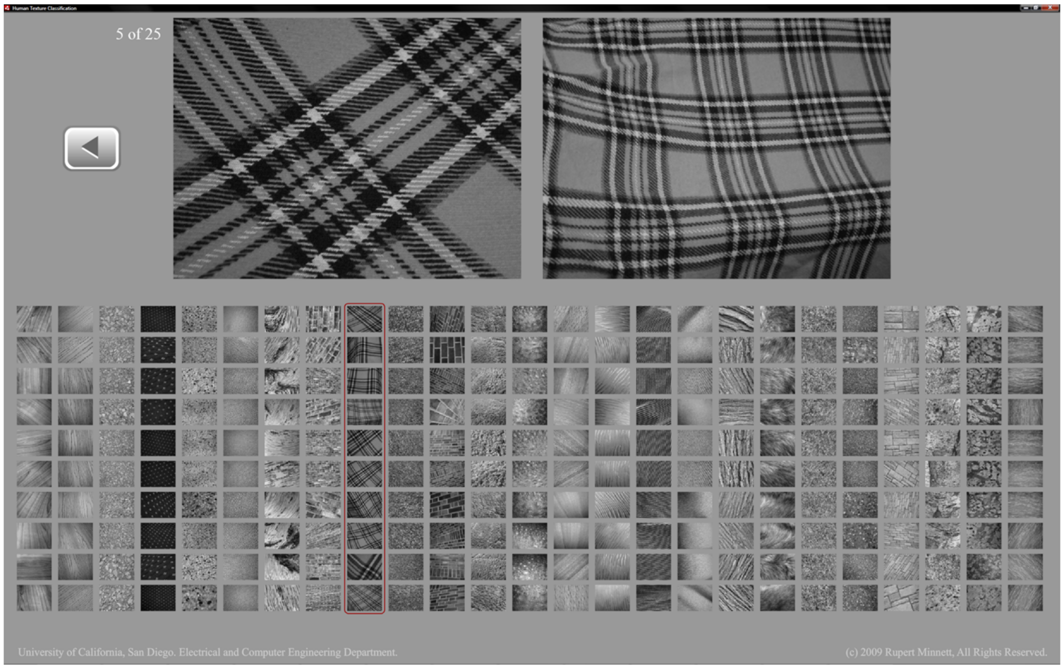

Human Texture Classification
5 of 25
University of California, San Diego. Electrical and Computer Engineering Department.
(c) 2009 Rupert Minnett, All Rights Reserved.

Figure 4.19: The human texture classification data collection instrument during the fifth presentation of the tenth round.

4.7 Human Results

The results from the human subject psychophysical experiment are compiled into trials (N=30, one for each subject) for comparison with the previously published computational results (see Figure 4.20). The human subjects, to avoid fatigue, stopped the experiment after being presented with ten training images per class. The computational results report classification accuracy with up to twenty training images per class and over repeated random sub-sampling validation (N=100 trials).

4.8 Computational Results

Using our computational model exposed to the UCSD Natural Scenes Dataset (see Section 4.4 and Appendix A) and trained with unsupervised learning, we perform the texture classification task for direct comparison with the previously published computational results. The dataset is partitioned into a training set and test set, the size of the training set is varied from one to twenty training image samples per class, and the mean and standard deviation of the classification accuracy on the remaining test set is reported over repeated random sub-sampling validation (N=100 trials).

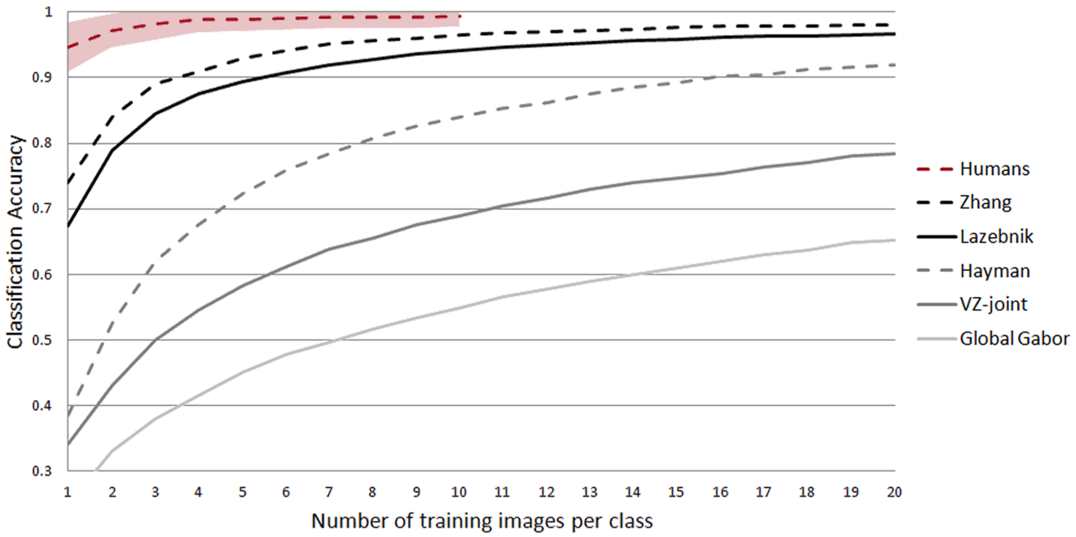


Figure 4.20: Human classification accuracy on the UIUCTex dataset.
The human subject psychophysical experiment classification accuracy is presented in the context of previously published computational results (see Figure 4.3 for a description of these algorithms). The red dotted line is the mean over the experiment trials (N=30, one for each subject) and the translucent red background spans the distribution of classification accuracies up to ±1 standard deviation. Statistical significance cannot be tested since the standard deviations are not reported for less than 20 training images per class [Zhang et al., 2006]. However, the human subjects clearly outperform state-of-the-art computational vision algorithms by a large margin. Previously published computational results are reproduced from [Zhang et al., 2006].

The model, similarly to adult humans, is already trained with unsupervised learning on stimuli that are not part of the texture dataset. The unsupervised learning allows the model to develop invariant feature representations from the statistics of the stimuli it is presented. The supervised learning in the posterior collateral sulcus (pCoS) is implemented by training an extrastriate visual cortex model with $k = 100$ cortical feature columns (see Section 4.5.3) for each class, c, by presenting the training texture samples as stimuli:

$$\mathbf{Z}_k^c(\mathbf{i},\mathbf{j}) \quad = \quad \mathbf{A}\left(\mathbf{V}_k^c(\mathbf{i},\mathbf{j})\right), \tag{4.26}$$

where $\mathbf{Z}\left(\bullet\right)$ is the s k^{th} cortical feature column output layer 3 activation in the texture class c pCoS model.

Each model, therefore, develops features consistent with the statistics of the texture class being presented (see Figure 4.21). This approach, due largely to the spatial and rotational invariance of the features learned in the early visual processing stream model with unsupervised learning, is capable of becoming selective for a single texture class.

The mapping from each of the trained pCoS extrastriate visual cortex model responses with each of the test texture samples to the class label can be implemented with any standard machine learning classifier, although some are less likely to be biologically plausible than others (see Section 4.1.4). We apply a biologically plausible winner-take-all (WTA) classifier operating on the most activated feature column in each of the pCoS models to label the test texture sample with the class used to train that pCoS model:

$$\hat{c}, \hat{k} \quad = \quad \underset{c,k}{argmax}\left(\sum_{\mathbf{i}}\sum_{\mathbf{j}}\mathbf{Z}_k^c(\mathbf{i},\mathbf{j})\right), \tag{4.27}$$

where $\hat{c}$ indexes the pCoS model (and by association, the class label) whose $\hat{k}^{th}$ cortical feature column generated the greatest response summed over the stimulus. When tested on the UIUCTex Dataset, the human texture classification accuracy extremely significantly ($p < 0.0001$ at both one and ten training images per class, Welch's unpaired

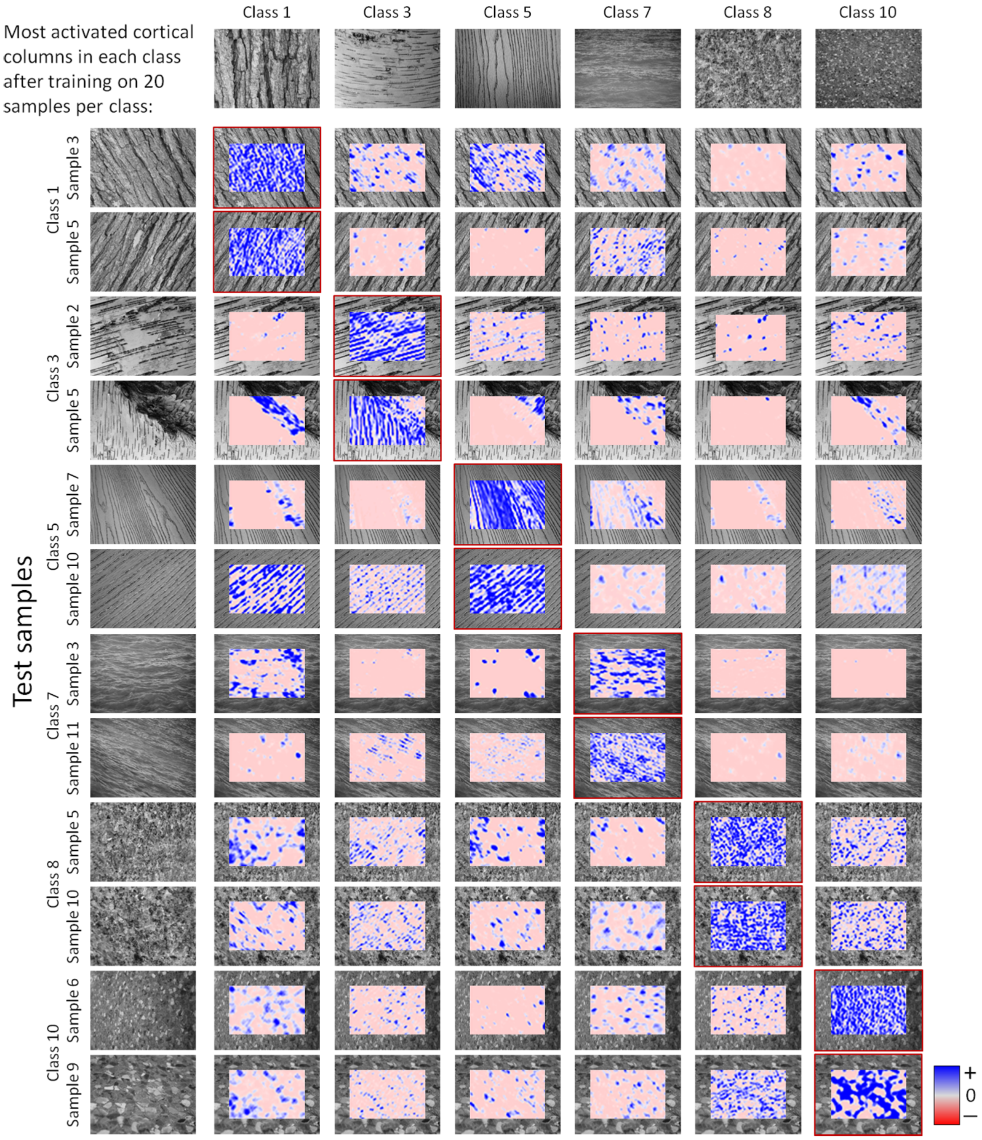


Figure 4.21: pCoS responses during the texture classification task.
Here we present a subset of the pCoS responses after supervised learning on twenty training images per class. Two test samples from each of six texture classes are displayed in the rows. The column headers display one of the twenty texture samples used during training for these six classes. Each column displays responses from the pCoS model trained with training samples from that class. When a test sample is presented to the model and the supervised learning is trained on the correct texture class training samples, a relatively large response is observed in the most activated pCoS cortical feature column (e.g. the upper left response compared with the other responses in that row). Using a simple winner-take-all (WTA) classifier, the test sample is assigned the class label of the pCoS model whose most activated cortical feature column generated the greatest response summed over the stimulus.

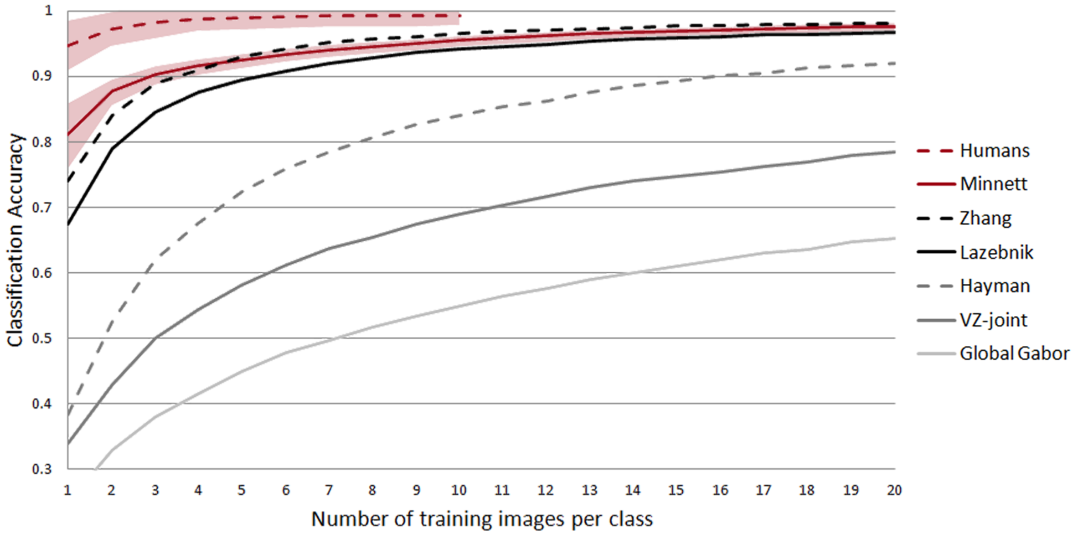


Figure 4.22: Human and computational classification accuracy on the UIUCTex dataset. Our computational texture classification accuracy is presented in the context of the human texture classification accuracy we recorded and previously published computational results (see Figure 4.3 for a description of these algorithms). The red solid line is the mean over the random sub-sampling validation (N=100) and the translucent red background spans the distribution of classification accuracies up to ±1 standard deviation. Previously published computational results are reproduced from [Zhang et al., 2006].

t-test) outperforms our computational texture classification accuracy. Our computational texture classification accuracy significantly ($p < 0.0001$, unpaired *t*-test assuming equal variances) outperforms the previously published state-of-the-art results [Zhang et al., 2006] with one training image per class, which do not significantly outperform our computational texture classification accuracy ($p = 0.07$) with twenty training images per class. We further investigate the performance of our computational texture classification accuracy by degrading the model (see Figure 4.23) and performing the classification task on other texture datasets (see Figure 4.24).

4.9 Conclusion

The human texture classification accuracy very significantly ($p < 0.0001$, Welch's unpaired *t*-test) outperforms our computational texture classification experiment, confirming our hypothesis. Although these exact texture image samples are novel to the human subjects, adults have a vast wealth of prior exposure to similar textures, provid-

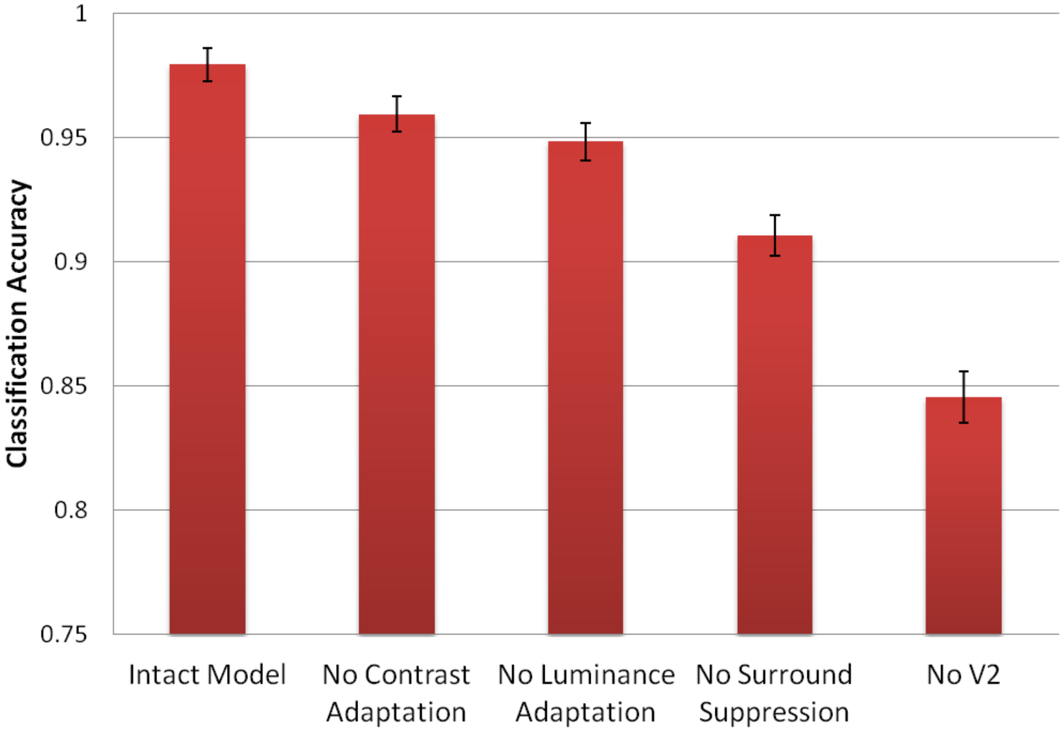


Figure 4.23: Model degradation classification accuracy on the UIUCTex dataset. The intact model is degraded by removing individual components and reporting the classification accuracy on the UIUCTex dataset with 20 training images per class. Removing the contrast adaptation from all feature column models has the least impact, followed by removing the luminance adaptation. Removing surround suppression from the cortical feature column model has a very significant impact on the performance and removing the secondary visual cortex (V2) model entirely has an extremely significant impact on the performance. The classification accuracy error bars express ±1 standard deviation over the random sub-sampling validation (N=100). The loss in classification accuracy after each degradation is statistically significant ($p < 0.0001$, Welch's unpaired *t*-test).

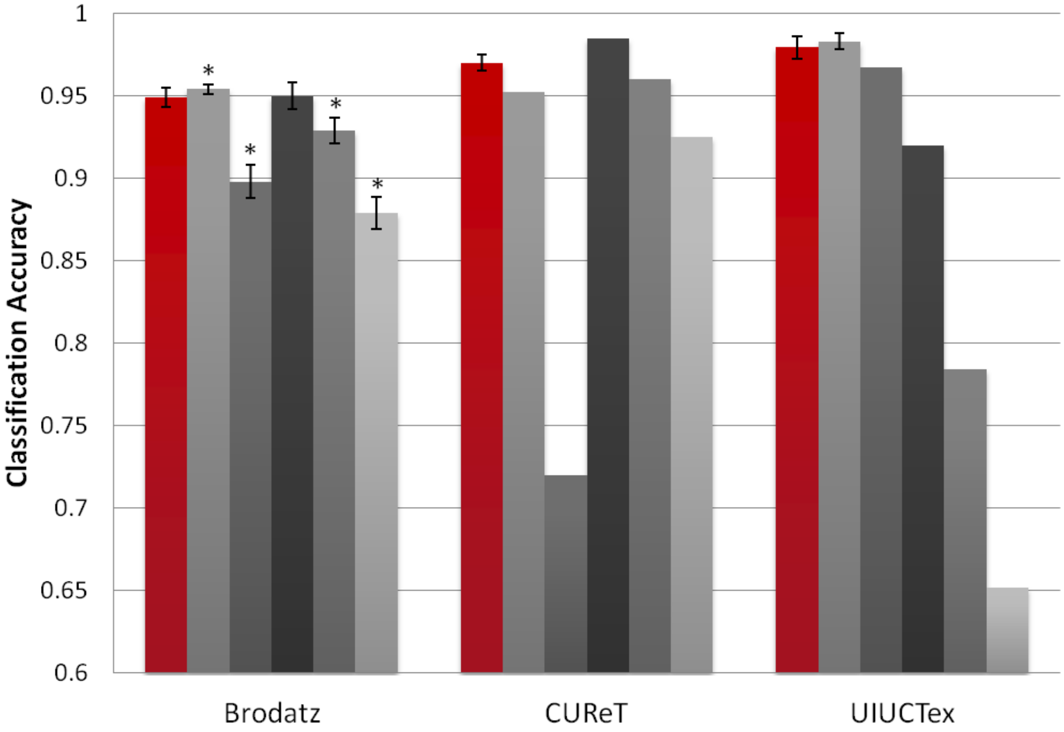


Figure 4.24: Texture dataset classification accuracy comparison.
Our computational texture classification accuracy (red) is presented in the context of previously published computational results ([Zhang et al., 2006], [Lazebnik et al., 2005], [Hayman et al., 2004], [Varma & Zisserman, 2003], and [Ma & Manjunath, 1996], respectively) on the Brodatz (3 of 9 training sub-images per class), CUReT (46 of 92 training images per class) and UIUCTex (20 training images per class) datasets (see Section 4.3). Although our model does not outperform the best approach for each dataset, which is often tailored for performance on that dataset (e.g. key point descriptor approaches do well on the UIUCTex dataset and the patch and global statistics approaches do well on the Brodatz dataset), our approach appears to report the most consistently competitive results. Where reported, the classification accuracy error bars express ±1 standard deviation over the random sub-sampling validation (N=100). The performance improvement of Hayman et al. [2004] over our model is not statistically significant ($p = 0.32$, Welch's unpaired t-test), but the performance improvement of Zhang et al. [2006] is ($p < 0.0001$, Welch's unpaired t-test), as is our performance improvement over the other three algorithms ($p < 0.0001$, Welch's unpaired t-test). Previously published computational results are reproduced from [Zhang et al., 2006].

ing them with semantic knowledge about the physical photographed sample (e.g. wood or cloth as opposed to texture number 1 or 25). Incorporating that knowledge into computational models, at least for this texture classification experiment, could improve their performance further.

In the experimental results we present here, computational models clearly are not currently capable of matching human texture classification performance. However, the results of the human texture classification experiment are not intended to be directly compared with computational results, but, rather, to provide an upper bound on reasonable performance for a computational model that incorporates adult human semantic knowledge about textures and adequately encapsulates the mammalian visual processing stream physiology. Our computational approach is more computationally expensive and complex than previously published methods, but may provide invaluable insight into the biological mechanism of performing the texture classification task and may begin a new direction for research striving to close the performance gap between computational and human vision.

Chapter 4, in part, has been submitted for publication of the material. Minnett, Rupert C.J.; Hecht-Nielsen, Robert. The dissertation author was the primary investigator and author of this material.

4.10 Chapter References

Artola, A. & Singer, W. (1987). Long-term potentiation and nmda receptors in rat visual cortex. *Nature*, 330(6149), 649–652.

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

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

Bonin, V., Mante, V., & Carandini, M. (2005). The suppressive field of neurons in lateral geniculate nucleus. *Journal of Neuroscience*, 25(47), 10844–10856.

Brodatz, P. (1981). *Textures: A Photographic Album for Artists and Designers.* Peter Smith Publisher, Incorporated.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, 25(46), 10577–97.

Carandini, M. & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, 264(5163), 1333–1336.

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21), 8621–8644.

Castellani, G. C., Quinlan, E. M., Cooper, L. N., & Shouval, H. Z. (2001). A biophysical model of bidirectional synaptic plasticity: Dependence on ampa and nmda receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 98(22), 12772–12777.

Cauwenberghs, G. & Poggio, T. (2001). Incremental and decremental support vector machine learning. *Advances in neural information processing systems*, 13, 409–415.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010). Separate processing of texture and form in the ventral stream: evidence from fmri and visual agnosia. *Cerebral Cortex*, 20(2), 433–446.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

Dana, K. J., Van Ginneken, B., Nayar, S. K., & Koenderink, J. J. (1999). Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics*, 18(1), 1–34.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A Optics and image science*, 2(7), 1160–1169.

David, S. V. & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network*, 16(2-3), 239–260.

de Sa, V. R. (1998). Category learning through multimodality sensing. *Neural Computation*, 10(5), 1097–1117.

Enroth-Cugell, C. & Lennie, P. (1975). The control of retinal ganglion cell discharge by receptive field surrounds. *The Journal of Physiology*, 247(3), 551–578.

Fukushima, K. (1975). Cognitron: a self organizing multilayered neural network. *Biological Cybernetics*, 20, 121.

Fukushima, K. (2008). Recent advances in the neocognitron. *Neural Information Processing*, 4984, 1041–1050.

Gabor, D. (1946). Theory of communication. *Communication Theory*, 93(26), 429–457.

Haralick, R. M., Dinstein, I., & Shanmugam, K. (1973). Textural features for image classification. *Ieee Transactions On Systems Man And Cybernetics*, 3(6), 610–621.

Hayman, E., Caputo, B., Fritz, M., & Eklundh, J.-o. (2004). On the significance of real-world conditions for material classification. *Most*, (pp. 253–266).

Hebb, D. O. (1949). *The organization of behavior*, volume 911. Wiley.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2), 181–197.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154.2.

Izhikevich, E. M. & Desai, N. S. (2003). Relating stdp to bcm. *Neural Computation*, 15(7), 1511–23.

Johnson, A. E. (1997). *Spin-images: A representation for 3-d surface matching.* PhD thesis, Carnegie Mellon University.

Johnson, A. E. & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5), 433–449.

Kohonen, T. (1995). Learning vector quantization. In M. Arbib (Ed.), *The Handbook of Brain. Theory and Neural Networks.* (pp. 537–540). MIT Press, Cambridge, MA.

Kohonen, T. (1998). Self-organizing maps of symbol strings. *Neurocomputing*, 21(1-3), 19–30.

Lazebnik, S., Schmid, C., & Ponce, J. (2004). Semi-Local Affine Parts for Object Recognition. In *British Machine Vision Conference*.

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8), 1265 –1278.

Lindeberg, T. & Gårding, J. (1997). Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6), 415–434.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In J. Tsotsos (Ed.), *Seventh IEEE International Conference on Computer Vision*, volume 2 (pp. 1150–1157 vol.2).: Ieee.

Ma, W. Y. & Manjunath, B. S. (1996). Texture features and learning similarity. *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (pp. 425–430).

Mante, V. (2005). *Gain controls based on luminance and contrast in the early visual system*. PhD thesis, Die Eidgenössische Technische Hochschule Zürich.

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S., & Carandini, M. (2005). Independence of luminance and contrast in natural scenes and in the early visual system. *Nature Neuroscience*, 8(12), 1690–1697.

Mikolajczyk, K. & Schmid, C. (2004). Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1), 63–86.

Ojala, T., Pietikainen, M., & Harwood, D. (1994). *Performance evaluation of texture measures with classification based on Kullback discrimination of distributions*, volume 1, (pp. 582–585). IEEE.

Olshausen, B. A. & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4), 481–487.

Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, 5(11), 12.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Ithica: Cornell Aeronautical Laboratory.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), 99–121.

Rust, N. C. & Movshon, A. J. (2005). In praise of artifice. *Nature Neuroscience*, 8(12), 1647 – 1650.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889–93.

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., & Poggio, T. (2005). A theory of object recognition : Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *Artificial Intelligence*, (December), 0–130.

Shouval, H., Intrator, N., & Cooper, L. N. (1997). Bcm network develops orientation selectivity and ocular dominance in natural scene environment. *Vision Research*, 37(23), 3339–3342.

Shushruth, S., Ichida, J. M., Levitt, J. B., & Angelucci, A. (2009). Comparison of spatial summation properties of neurons in macaque v1 and v2. *Journal of Neurophysiology*, 102(4), 2069–2083.

Sincich, L. C. & Horton, J. C. (2002). Divided by cytochrome oxidase: a map of the projections from v1 to v2 in macaques. *Science*, 295(5560), 1734–7.

Sincich, L. C. & Horton, J. C. (2005). The circuitry of v1 and v2: integration of color, form, and motion. *Annual Review of Neuroscience*, 28(1), 303–326.

Varma, M. & Zisserman, A. (2002). Classifying images of materials: Achieving viewpoint and illumination independence. *Lecture Notes in Computer Science*, 3, 255–271.

Varma, M. & Zisserman, A. (2003). Texture classification: Are filter banks necessary ? a review of the vz classifier. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2003 Proceedings*, 2, II–691–8.

Yeung, L. C., Shouval, H. Z., Blais, B. S., & Cooper, L. N. (2004). Synaptic homeostasis and input selectivity follow from a calcium-dependent plasticity model. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41), 14943–14948.

Yin, Q., Kim, J.-N., & Shen, L. (2009). Rotation-invariant texture classification using circular gabor wavelets. *Optical Engineering*, 48(1), 017001.

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2006). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2), 213–238.

Chapter 5

Neural Network Tomography: Network Replication from Output Surface Geometry

Multilayer perceptron networks whose outputs consist of affine combinations of hidden units using the *tanh* activation function are universal function approximators and are used for regression, typically by reducing the MSE with backpropagation. We present a neural network weight learning algorithm that directly positions the hidden units within input space by numerically analyzing the curvature of the output surface. Our results show that under some sampling requirements, this method can reliably recover the parameters of a neural network used to generate a data set.

5.1 Introduction

Neural networks have been a staple of artificial intelligence since soon after the field began. Early attempts at assigning weights to produce desired output included tuning them by hand and adaptive techniques [Minsky, 1954], such as the backpropagation algorithm, presented in [Werbos, 1974] and popularized in the 1980s. Most subsequent methods of choosing neural network weights are themselves variants of backpropagation, which uses the chain rule to minimize a loss function, usually the mean squared error [Nilsson, 1965; Steinbuch, 1965; Widrow et al., 2005]. Shun-ichi Amari, who has since made significant contributions to this field, came very close to discovering backpropagation during the 1960s [Amari, 1967].

There are two aspects of this approach that are not necessarily desirable. First, every data point in the training set influences every weight to some degree because the training algorithm treats all weights and all data in the same fashion (i.e. by finding the derivatives of the error with respect to every weight, evaluated at every data point). Second, in most applications, the data are treated as being generated by some completely unknown process. If, on the other hand, the data are assumed to be outputs of some parameterized system, novel training methods become possible.

We present a simple regression technique that assigns neural network weights given a data set assumed to be generated by such a parameterized system. Instead of each point influencing every weight in the same way, regions of the input space are ranked by relative usefulness in determining where hidden units should be. This method infers the hidden parameters of a target function by analyzing only its input and output. We describe this technique as *tomographic* in analogy to the computed tomography tools in medical science which infer the internal structures of an intact object by observing the results of probe signals.

Our method begins with the previously stated assumption about the data set: the input-output pairs are the inputs and outputs of a neural network, referred to as the *teacher network*. The only assumptions about the teacher are that it is a two-layer perceptron network with the hyperbolic tangent (*tanh*) activation function in the hidden units and its outputs are affine combinations of the hidden units. The teacher parameters, such as its weights or number of hidden units, are unknown. A *student network* is constructed one hidden unit at a time by iteratively inferring the parameters of the teacher network's hidden units by finding regions of the input space where the curvature of the output is characteristic of a *tanh* function. This works because the affine combinations of the output units preserve curvature. As each new student hidden unit is added, the effect of the corresponding hidden unit in the teacher network is subtracted from the function output in the data set, and the procedure terminates when all of the hidden units have been matched.

The next section contains a history of function approximation and examines the relation of our new technique in terms of this history. Sections 5.3-5.4 contain a detailed description of the tomographic method and provide a few examples of its application. The article concludes with a brief discussion.

5.2 A Brief History of Function Approximation

The history of function approximation is centuries old and can be roughly divided into several different types of approach: methods that approximate a function by construction of a polynomial with the same derivatives as the function, approximation techniques that are based on a trigonometric series, methods that partition the input space into regions which are independently approximated, and methods that use neural network-like parameterizations. Interestingly, these different categories of approximation techniques appeared roughly chronologically.

A side note to the development of the many different function approximation techniques is the development of methods to estimate values for the parameters of the approximations. Perhaps surprisingly, one of the most common methods of parameter estimation today, the method of least squares, was developed at the end of the eighteenth century [Gauss, 1809]. Gauss used least squares to find orbital parameters of a newly discovered celestial body, given only a few observations, and became famous for predicting where this body (the asteroid Ceres) would reappear in the night sky. Despite the instant fame he gained upon this success, the true discoverer of the least squares method is unknown and disputed. Though Gauss claimed to have discovered least squares in 1795, Adrien-Marie Legendre actually published a description of least squares earlier [Legendre, 1805], whereas Gauss did not publish his least squares method until 1809. Gauss's claim is only supported by his word; no clear written indication of his use or discovery of least squares exists before 1805 [Stigler, 1981].

5.2.1 Polynomial-based approximation methods

One of the earliest examples of function approximation is in Newton's Method for root finding, written in his *De analysi per æquationes numero terminorum infinitas* [Newton, 1669]. This method first appeared in its modern form after further development [Raphson, 1690]. The resulting Newton-Raphson method finds the roots of arbitrary functions by successively calculating the zero-crossing of a simple linear approximation of that function identified by its first derivatives.

This idea of function approximation was generalized to higher derivatives and actually explored as a function approximation in itself, not as a tool to reach some other goal, by *Methodus Incrementorum Directa et Inversa* [Taylor, 1715], stating that a function of x may be evaluated at any value of x As "*seriem terminorum numero* *infinitam*," or a series terminating at infinity, each term of which is constructed from incrementally higher derivatives of the function. This idea was further explored by the Scottish mathematician Colin Maclaurin.

5.2.2 Trigonometric methods

An independent approach to function approximation began in the early 19th century, when French mathematician, physicist, and Egyptologist Joseph Fourier published his *Mémoire sur la propagation de la chaleur dans les corps solides* Fourier [1808]. Fourier noted that a function could be represented as an infinite trigonometric series:

$$\phi(y) = a_0 \cos\frac{\pi y}{2} + a_1 \cos 3\frac{\pi y}{2} + a_2 \cos 5\frac{\pi y}{2} + \cdots \ ,$$

and that the coefficients a_i could be found by integrating

$$\text{“ } a_i = \int_{-1}^{+1} \phi(y) \cos(2i+1)\frac{\pi y}{2} dy \ , \text{”}$$

or taking the inner product of the original function with the relevant basis function [Fourier, 1808].

Originally, Fourier decomposed functions into a linear sum of trigonometric terms because the solution to the heat equation in polar coordinates has that form. In his *Théorie analytique de la chaleur*, he exposited the "Development of an arbitrary function in trigonometric series" [Fourier, 1822], and the beginnings of the field of Fourier analysis.

In contrast to Fourier's basis of sines and cosines, which are defined by frequency and phase and therefore have no locality, *wavelet* bases are defined by the addition of position and size. For example, Gabor's logons are designed to constitute a basis, much like Fourier bases but with the addition that each term in the approximating sum only influences the function in a local neighborhood [Gabor, 1946]. This was extended to two dimensions, for example by Daugman, who showed how visual information (functions of two variables) can be represented in such a basis [Daugman, 1988], and by Daubechies, who introduced orthogonal bases of wavelets [Daubechies, 1988].

Representation of arbitrary functions by polynomials or by trigonometric series, differs from our tomographic technique for function approximation in that the basis functions of neural networks are inherently adaptive because of their input weights. Functions are approximated by polynomials or Fourier series by finding coefficients (corresponding to network output weights) with inner products of the function to be approximated with the exhaustive set of basis functions (or an exhaustive sequence of derivatives). It is worth noting that the lack of adaptability of the bases functions in these two types of approaches typically leads to the use of many more terms in an approximation.

5.2.3 Piece-wise methods

The non-global properties of basis elements used in wavelet representations have a corresponding polynomial analog in piece-wise polynomial interpolation/approximation techniques developed first by Paul de Casteljau and Pierre Bézier, who were motivated by practical automotive engineering concerns. Casteljau's breakthrough was to augment the set of points through which a curve is to pass with a set of control points which defined the tangents [Farin, 2002]. The same curve parameterization was independently derived in terms of cylindrical surface intersections by Bézier, with whose name it became associated due to his publishing before Casteljau. The work of Casteljau and Bézier lead directly to the field of spline interpolation, which is applied in computer graphics, for example, as non-uniform rational B-splines (NURBS). In the case of interpolation, a variant of function approximation, reduction of complexity (from a large set of points to a functional parametrization) is not the goal, but rather the creation of a smooth and continuously differentiable function which evaluates to a defined set of values.

This idea of creating functions from real-world data became known in the field of statistics as regression. Jerome Friedman presented a new method, multivariate adaptive regression splines (MARS), in 1991 with several similarities to our new tomographic technique, and to existing function approximation methods [Friedman, 1991]. Hearkening back to older techniques, the approximation consists of a weighted sum of simple functions of the input variables, though these simple basis functions can be multidimensional:

$$f(x_1, x_2, x_3, \ldots) = a_0 + \sum_{K_m=1} f_i(x_i)$$

$$+ \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{K_m=3} f_{ijk}(x_i, x_j, x_k) + \cdots \ ,$$

where $K_m = d$ indicates the set of d indices for which a basis function is defined and each $f_m()$ is a weighted basis function: $f_m() = a_m B()$. Each univariate basis function is a smoothed piecewise linear function that is zero on the left, and linear on the right (or vice-versa), using a cubic term as the segment that smoothes between the two, and allows for continuity and continuous first derivatives:

$$C(x|s=+1,t_-,t,t_+) =$$

$$\begin{cases} 0 & x \leq t_- \\ p_+(x-t_-)^2 + r_+(x-t_-)^3 & t_- < x < t_+ \\ x-t & x \geq t_+ \ , \end{cases}$$

where $s = +1$ indicates that the function grows as x increases (i.e. to the right), and

$$C(x|s=-1,t_-,t,t_+) =$$

$$\begin{cases} -(x-t) & x \leq t_- \\ p_-(x-t_-)^2 + r_-(x-t_-)^3 & t_- < x < t_+ \\ 0 & x \geq t_+ \ , \end{cases}$$

where $s = -1$ indicates that the function grows as x decreases. $p_{+/-}$ and $r_{+/-}$ are chosen to ensure continuity and continuous derivatives. Multivariate basis functions are simply the products of the univariate basis functions.

This model is an extension of an earlier, non-smooth model which uses the simpler basis functions:

$$C(x|s=+1,t) = \begin{cases} 0 & x \leq t \\ (x-t) & x > t \ , \end{cases}$$

and $C(x|s=-1,t)$ defined analogously.

The weights a_i are fitted in rounds, each of which increases the dimensionality of the basis functions to be added. Each round is followed by the elimination of terms which most degraded or least improved the goodness-of-fit in that round. This stepwise approach is not new to statistics, but Friedman innovated a combination of spline knots and smoothing provided by the cubic basis functions that yields continuous and continuously differentiable models fit by dividing up the input space [Friedman, 1991]. Interestingly, the weighted sum of simple basis functions is similar to the weighted sum of hidden units used in neural network regression (described in 5.2.4); the t parameter(s) of the spline knots set the position of the basis function within the input space, analogous to neural network input weights, and the weighting term a_i corresponds to the output weight (with a_0 as the output bias weight).

5.2.4 Neural Network methods

Neural networks have been an active topic of research since the beginnings of Artificial Intelligence. The first researchers attempted to reproduce the capabilities of the brain by crudely simulating its functional components, neurons, resulting in the so-called "perceptron," a single-neuron computational model that showed promise as a classifier [Rosenblatt, 1961], but progress was slow until the incorporation of a differentiable sigmoid activation function and the invention of the backpropagation algorithm [Werbos, 1994]. Since then, neural network classifiers with squared-error loss have been proved to be Bayes-optimal classifiers [Wan, 1990; Ruck et al., 1990] and neural networks with a single hidden layer of neurons with sigmoid activation functions and linearly weighted outputs have been proved to be a universal function approximator, provided that a sufficient number of hidden units are used [Hornik et al., 1989].

A single output of a neural network regression model can be formulated as

$$\hat{y}_t = v_0 + \sum_{h=1}^{H} v_h f\left(u_{0,h} + \sum_{n=1}^{N} u_{n,h} x_{t,n}\right) ,$$

where u and v are the weights to the hidden units and weights to the outputs, respectively, and the activation function $f()$ is a sigmoid activation function, such as the logistic function or the hyperbolic tangent.

The most common training method is to reduce the value of a loss function, which is nearly always mean squared error loss:

$$\epsilon(u, v) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 ,$$

where T is the number of input-output training pairs.

The backpropagation algorithm, invented by Paul Werbos in 1974 [Hecht-Nielsen, 1989], is an efficient method of evaluating the gradient of a loss function with respect to the weights of a neural network. It is widely claimed that gradient descent with backpropagation can perform poorly because of local minima on the error surface. Despite this frequent claim, we have not been able to find any published proof that sub-optimal minima exist, although some evidence has been discovered [Fukumizu & Amari, 2000].

Although local minima may or may not exist, from a practical perspective, gradient descent can be slow in nearly flat regions of the error surface or ineffectually oscillate in slightly downward-sloping "canyons" in the error surface. A traditional approach to improving convergence in these situations is to use a "momentum" factor, α. The weight adjustment used is equal to $1 - \alpha$ times the weight change suggested by the gradient, plus α times the weight change used in the previous iteration (typically, α is close to 1), resulting in a series of changes to the weight vector that can only gradually change its direction and that avoids oscillation. Though momentum is a practical solution for slow convergence, it is by no means a safeguard against local minima that might be mistaken for a global minimum; determining an optimal momentum factor to overcome local minima is not necessarily less difficult than determining a step size that would have overcome the local minimum without using momentum.

Another recent attempt to avoid the potential problem of local minima is to consider the weights as a point on a manifold embedded in Fisher information space and adjust the weights in the direction of maximum improvement [Amari, 1998]. The question of local minima is obviated completely by our tomographic technique which does not perform a gradient descent on the error surface.

An alternative to neural networks with sigmoid activation functions is the *radial basis function* network, in which the hidden units are assigned a position within input space and, when given input, return some measure of the proximity of that input to that hidden unit (typically, using the normal distribution density function). The (linear) output weights can be determined through ordinary least squares [Haykin, 2009]. One drawback to this type of regression method is that, to ensure complete coverage of the input space, typically every input training example (or a significant fraction) is used to create a radial basis function, even if the underlying concept being learned is not complex [Haykin, 2009]. In contrast, our algorithm iteratively increases model complexity until no error remains.

5.2.5 GMDH-type methods

In 1968, Ivakhnenko introduced the so-called "Group Method of Data Handling" (GMDH) [Ivakhnenko, 1971; Galushkin, 2007]. This technique was first used to learn an arbitrary-order polynomial approximation of the input/output characteristics of a control theory “plant” (i.e. a black-box function) given a data set consisting of N pairs of input vectors and output values.

The GMDH learning algorithm contains several ideas that have been explored throughout the history of function approximation. GMDH models are network models with layers of hidden units in which each processing unit $U_{h,i}$ (hidden layer h, unit index i) accepts as inputs a small subset of the previous layer's outputs $z_{h-1,j}$ (where j indexes the units of the previous layer) and computes its function $f_{h,i}(z_a, z_b)$, analogously to the activation function of a neural network [Oh & Pedrycz, 2002].

Each of the functions $f()$ is in the form of a linear combination of simple, possibly nonlinear, functions of the inputs to $f()$. For example, the polynomial GMDH described in [Ivakhnenko, 1971] uses, for two input variables x_1 and x_2,

$$f(x_1, x_2) = \sum_{i=0}^{5} a_i e_i(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$$

$$+a_3 x_1^2 + a_4 x_2^2 + a_5 x_1 x_2,$$

where the a_i are the six input weights to learn for this unit. In this case, the components of $f()$ are the functions

$$\begin{array}{ll} e_0(x_1, x_2) = 1 & \\ e_1(x_1, x_2) = x_1 & e_2(x_1, x_2) = x_2 \\ e_3(x_1, x_2) = x_1^2 & e_4(x_1, x_2) = x_2^2 \\ e_5(x_1, x_2) = x_1 x_2. & \end{array}$$

This is in contrast to neural networks, which capture nonlinear effects not by including nonlinear terms, but by applying a nonlinear activation function (e.g. tanh()) to a sum of linear terms $f(x_1, x_2, ...) = \tanh(a_0 + a_1 x_1 + a_2 x_2 + ...)$. This approach of GMDH, in which each unit consists of a linear combination of functions, has an advantage over neural networks, because parameters can be learned using a matrix pseudo-inverse, thus learning the optimal parameters efficiently, and at once.

The GMDH model learning algorithm adds successive hidden layers to the network until the predictive accuracy of the model ceases to improve. Each layer learning phase uses three steps:

1. constructing a large number of new hidden units by selecting many small subsets of the outputs of the previous layer (typically an exhaustive set of subsets with small cardinality), applying each new hidden unit's nonlinear transformation functions, then learning the weights to linearly combine the transformed inputs into the unit's output,

2. selecting the best performing newly learned units (i.e. the most predictive sets of inputs) using a hold-out set and some performance metric, typically squared error,

3. checking for termination (non-decreasing error) [Oh & Pedrycz, 2002].

For example, in using the two-term units above to estimate a function y of d variables, $x_1..x_d$, the first step learns all $N(N-1)/2)$ sets of 6 weights to minimize the squared error $y - f(x_i, x_j)$ where $f(x_i, x_j) = a_0 + a_1x_i + a_2x_j + a_3x_i^2 + a_4x_j^2 + a_5x_ix_j$ (i.e. the coefficients of a quadratic equation are learned for every pair of inputs to predict the output).

The second step evaluates the new functions on the hold-out set data, eliminating those $f(x_i, x_j)$ with unacceptably high error, either by thresholding, or pruning a predetermine number of them.

In the third step, the new unit with the smallest summed error is compared to the smallest summed error in the previous layer. If the error is still decreasing, and further accuracy is desired, then the process is repeated, otherwise the unit with smallest error is considered the output of the network.

There are a number of attractive features of GMDH-type algorithms. Because each unit is a linear combination of fixed functions of that unit's inputs, the optimal weights can be calculated directly and because each unit has a small number of inputs (typically much smaller than the number of samples), these calculations are usually well-conditioned. The units whose parameters *are* estimated from ill-conditioned systems are usually eliminated in the unit selection phase because they poorly predict the outputs of the hold-out set. Interestingly, the GMDH learning algorithm explicitly induces structure in its models by rejecting connections (units) in phase 2, unlike neural networks in which successive layers are usually completely connected. And finally, there is no restriction on the nonlinear functions that transform the inputs to a unit [Oh & Pedrycz, 2002], whereas the nonlinear transformations within neural networks are typically all the same function. This could be useful, for example, for incorporating prior knowledge.

Some of the drawbacks of GMDH are consequences of the fact that it is essentially a greedy algorithm; at each step, some inputs and/or combinations of inputs are irretrievably lost because they are not linearly predictive of the function output value, even though they may have been helpful to later layers of the network. Another restriction is that because units usually consist of a small number of inputs, the set of learnable unit outputs is restricted, though this restriction can potentially be overcome with additional layers.

5.3 Neural Network Tomography

Our neural network tomography procedure for replicating networks, outlined in Procedure 1, successively adds hidden units to an initially empty (with no hidden units) *student network* by estimating the weights of the *teacher network* hidden units (whose weights are concealed), effectively canceling the teacher hidden units' contributions to the difference surface with each new student hidden unit. Inferring the weights of a student hidden unit occurs in two phases: initialization of the hidden unit weights (Procedure 1, lines 7-12), followed by fine-tuning of these weights to closely match a teacher hidden unit (Procedure 1, lines 13-15).

Our method, like other regression techniques, learns a mapping by some target function, $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$, given only the input-output pairs $\mathbf{y}_t = f(\mathbf{x}_t)$ for T training samples, where $t = 1 \ldots T$. Since a two-layer neural network with a *tanh* activation function in the hidden units and linear output units has been shown to be a universal approximator [Hornik et al., 1989], let us consider the target function as having been approximated by a neural network, the teacher network, with an unknown number of H hidden units (see Figure 5.1). Such a function can be replicated perfectly by a neural network, the student network, with $\hat{H} = H$ hidden units parameterized by any element of a non-abelian Lie group transformation of the teacher network weights [Sussmann, 1992; Chen et al., 1993]. However, traditional gradient descent by minimizing a mean squared error (MSE) objective function (e.g. backpropagation) requires choosing $\hat{H}$ prior to learning or adaptively adding hidden units during learning, where neither approach is immune to slow convergence.

We present an alternative objective function to minimize instead of MSE: mean squared curvature (MSC), not to be confused with Gaussian or mean curvature. We define the *curvature* at input sample $\mathbf{x}_t$ to be a measure of the deviation of the surface

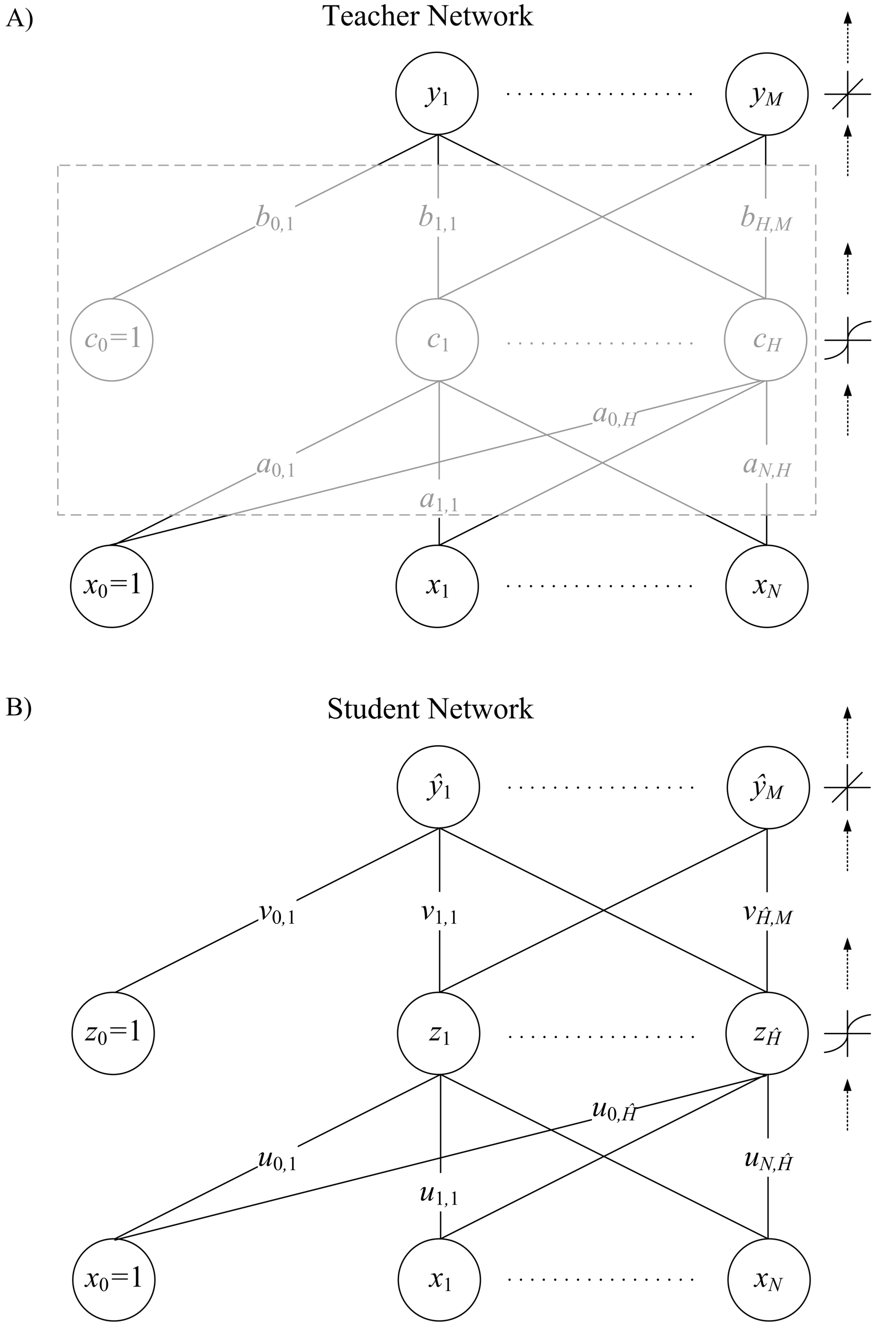


Figure 5.1: The teacher and student networks.
A) The teacher network with H hidden units to be matched iteratively by the student network. The overlaid translucent box illustrates that the teacher hidden units and their weights are concealed from the student network. **B)** The student network, after all H hidden units of the teacher have been learned successfully.

Table 5.1: Neural network tomography notation.

N	$=$	input dimensionality
M	$=$	output dimensionality
H	$=$	number of teacher hidden units
$\hat{H}$	$=$	number of student hidden units
$a_{n,h}$	$=$	teacher network input weight from input unit n to hidden unit h
$b_{h,m}$	$=$	teacher network output weight from hidden unit h to output unit m
$u_{n,h}$	$=$	student network input weight from input unit n to hidden unit h
$v_{h,m}$	$=$	student network output weight from hidden unit h to output unit m
$x_{t,n}$	$=$	the training data input
$y_{t,m}$	$=$	the teacher network output
	$=$	$b_{0,m}+ \sum_{h=1}^{H} b_{h,m} \tanh\left(a_{0,h} + \sum_{n=1}^{N} x_{t,n}a_{n,h}\right)$
$\hat{y}_{t,m}$	$=$	the student network output
	$=$	$v_{0,m}+ \sum_{h=1}^{\hat{H}} v_{h,m} \tanh\left(u_{0,h} + \sum_{n=1}^{N} x_{t,n}u_{n,h}\right)$

in a small neighborhood about $\mathbf{x}_t$ from a planar approximation of the surface in that neighborhood. This technique allows for the weights of each teacher network hidden unit to be reliably and sequentially learned by the student network. Since properties of neural networks with one output generalize to multiple outputs [Sussmann, 1992; Kůrková & Kainen, 1994], we only consider the case where $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\mathbf{x}_t \in \mathbf{X}$ is a compact subset of $\mathbb{R}^2$, and $\mathbf{y} \in \mathbf{Y}$ is a compact subset of $\mathbb{R}$.

Our tomographic procedure begins by adding one zero hidden unit (input and output weights are zero) to the student network. The difference surface (initially identical to the teacher output surface) between the teacher and student networks is defined as

$$\mathbf{d} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - v_1 \tanh(\mathbf{X}\mathbf{u}), \tag{5.1}$$

and in expanded form for teacher $f : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$\begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_T \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} - v_1 \tanh \left(u_{0,1}\mathbf{J}_{T,1} \right.$$

$$\left. +u_{1,1} \begin{pmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{T,1} \end{pmatrix} + u_{2,1} \begin{pmatrix} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{T,2} \end{pmatrix} \right) ,$$

where $\mathbf{J}_{T,1} \in \mathbb{R}^T$ is a column vector of ones. The MSC of this difference surface is gradually reduced as hidden units are added to the student network, causing its output to converge to the teacher network output.

Procedure 1 Neural Network Tomography Outline.

```
X ← training inputs
K ← number of nearest neighbors
repeat
    add a zero hidden unit to the student
    for all $\mathbf{x}_t \in \mathbf{X}$                                              Sec. 5.3.1
        calculate the difference surface curvature, $\kappa(\mathbf{x}_t)$, using $K$ nearest neighbors of $\mathbf{x}$   Equation 5.3
    for all $\mathbf{x}_+ \in \mathbf{X}$ whose $\kappa$ is positive                      Sec. 5.3.2.1
        for all $\mathbf{x}_- \in \mathbf{X}$ whose $\kappa$ is negative
            evaluate the similarity score, $\phi$, using $\mathbf{x}_+$ and $\mathbf{x}_-$         Eqs. 5.4,5.8
            if $\phi$ is the largest so far then
                $(\mathbf{x}'_+, \mathbf{x}'_-) \leftarrow (\mathbf{x}_+, \mathbf{x}_-)$
    set the new student hidden unit input weights using $(\mathbf{x}'_+, \mathbf{x}'_-)$     Eqs. 5.5-5.7
    while the MSC is decreasing                                         Sec. 5.3.2.2
        adjust all the student hidden unit weights to try and minimize the MSC
        recalculate the difference surface MSC                          Equation 5.9
until the difference surface is close to constant                       Sec. 5.3.3
set the student output bias to minimize error                           Equation 5.10
```

5.3.1 Estimated Signed Curvature

Let the function $knn(t, k)$ index the k^{th} nearest neighbor to the input sample $\mathbf{x}_t$ using a Euclidean distance metric. The curvature of the network difference surface can be calculated with the least squares estimate (LSE) of a hyperplane about $\mathbf{x}_t$ and its K

nearest neighbors:

$$\mathbf{p} \quad = \quad \mathbf{X}_t^{\dagger}\mathbf{d}_t = \begin{pmatrix} \mathbf{x}_t^{\intercal} \\ \mathbf{x}_{knn(t,1)}^{\intercal} \\ \vdots \\ \mathbf{x}_{knn(t,K)}^{\intercal} \end{pmatrix}^{\dagger} \mathbf{d}_t \ , \tag{5.2}$$

and in expanded form for teacher $f : \mathbb{R}^2 \to \mathbb{R}$,

$$\begin{pmatrix} p_0 \\ p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} 1 & x_{t,1} & x_{t,2} \\ 1 & x_{knn(t,1),1} & x_{knn(t,1),2} \\ \vdots & \vdots & \vdots \\ 1 & x_{knn(t,K),1} & x_{knn(t,K),2} \end{pmatrix}^{\dagger}$$

$$\bullet \begin{pmatrix} y_t - \hat{y}_t \\ y_{knn(t,1)} - \hat{y}_{knn(t,1)} \\ \vdots \\ y_{knn(t,K)} - \hat{y}_{knn(t,K)} \end{pmatrix} ,$$

where † denotes the Moore-Penrose pseudoinverse, $\mathbf{x}_{knn(t,k)}$ is the vector of inputs (including the input bias) for the k^{th} nearest neighbor to the t^{th} training sample, and $\mathbf{X}_t$ and $\mathbf{e}_t$ are the neighborhood subset about $\mathbf{x}_t$ of input samples and difference surface outputs, respectively. The curvature of the difference surface is estimated as the product of the sign of the LSE residual with the mean squared error of the hyperplane linear fit:

$$\kappa(\mathbf{x}_t) = -\mathrm{sgn}\left(y_t - \mathbf{x}_t^{\intercal}\mathbf{p}\right) \frac{1}{K} \sum_{k=1}^{K} \left(y_{knn(t,k)} - \mathbf{x}_t^{\intercal}\mathbf{p}\right)^2 \ . \tag{5.3}$$

This measure of curvature emphasizes the regions of the difference surface which are most influenced by an underlying hidden unit. Thus, each point is assigned a curvature magnitude and sign.

5.3.2 Hidden Unit Parameterization

Each hidden unit of the teacher network is parameterized by $N + 1 + M$ weights, namely the input weights, the input bias, and the output weights. The goal of our

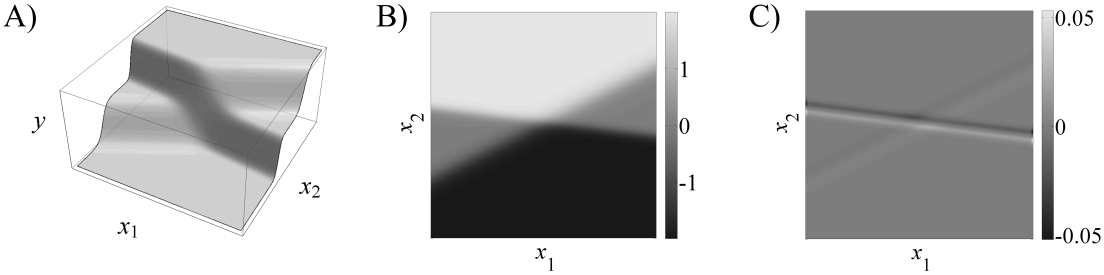


Figure 5.2: Estimated signed curvature.
A 3D (**A**) and 2D (**B**) illustration of a teacher network input and output with 2 hidden units to be replicated by the student network. **C)** The calculated MSC at each point in the difference surface. Note that the regions of maximum curvature correspond to the steeper of the two teacher hidden units.

approach is to estimate each of the teacher network hidden units sequentially, beginning with the hidden unit contributing most to the MSC. Minimizing the MSE, in many cases, will fail to achieve this goal (see Figure 5.6). Alternatively, minimizing the MSC of the difference surface will reliably uncover the weights of one hidden unit at a time.

5.3.2.1 Optimization Initialization

The MSC objective function is prone to local minima (see Figure 5.4B) introduced in many cases by matching the student hidden unit weights with those of at least one of the teacher hidden units. Therefore, we must first initialize the optimization procedure with input weights and an input bias close to the MSC local minimum associated with the most prominently contributing hidden unit. One way to accomplish this is to use a comparison kernel to evaluate the similarity of the difference surface curvature and a hypothetical *tanh* hidden unit ramp. We define the kernel, $\psi()$, as a function of a pair of training sample inputs (one at which the curvature is positive, $\mathbf{x}_+$, and the other negative, $\mathbf{x}_-$) in the form of the second derivative of the hidden layer transfer function, *tanh*:

$$\psi\left(t, \mathbf{x}_+, \mathbf{x}_-\right) = \tanh\left(\mathbf{x}_t^\intercal \mathbf{u}'\right) \operatorname{sech}^2\left(\mathbf{x}_t^\intercal \mathbf{u}'\right) \ , \tag{5.4}$$

where $\mathbf{u}'$ is a vector of weights that parameterize a line within input space bisecting the line segment between the pair of training sample inputs $\mathbf{x}_+$ and $\mathbf{x}_-$. These weights are:

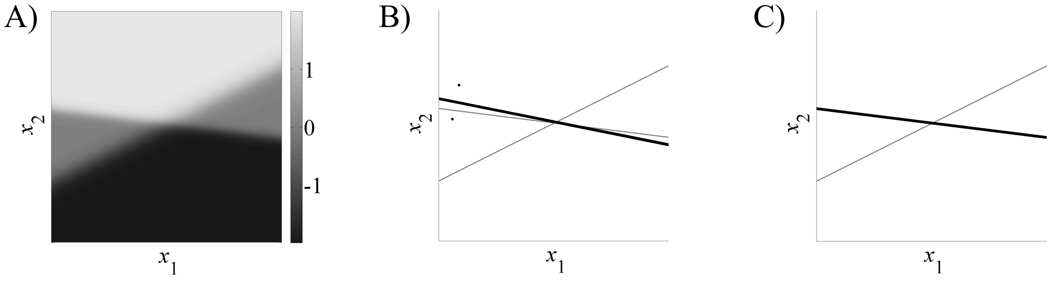


Figure 5.3: Hidden unit alignment.
A) The teacher network introduced in Figure 5.2A. **B)** The teacher hidden unit positions within input space, indicated by thin lines where $tanh\left(\mathbf{x}_t^\intercal \mathbf{a}_h\right) = 0$, overlaid by the initialized student hidden unit, indicated by a thick line where , and the two training samples, $\mathbf{x}_+, \mathbf{x}_-$, used for initialization. **C)** The teacher hidden unit positions overlaid by the optimized student hidden unit, indicated by a thick line where $tanh\left(\mathbf{x}_t^\intercal \hat{\mathbf{u}}\right) = 0$.

$$
\begin{aligned}
u'_{1,1} &= \frac{x_{+,1} - x_{-,1}}{\|\mathbf{x}_+ - \mathbf{x}_-\|_2^2} && (5.5)\\
u'_{2,1} &= \frac{x_{+,2} - x_{-,2}}{\|\mathbf{x}_+ - \mathbf{x}_-\|_2^2} && (5.6)\\
u'_{0,1} &= \frac{u'_{1,1}}{2}\left(x_{+,1} + x_{-,1}\right) + \frac{u'_{2,1}}{2}\left(x_{+,2} + x_{-,2}\right)\ . && (5.7)
\end{aligned}
$$

A normalized dot product similarity score, $\phi()$, is then calculated between the kernel and the difference surface curvature over all training samples:

$$
\phi\left(\mathbf{x}_+, \mathbf{x}_-\right) = \frac{\sum_{t=1}^{T} \kappa\left(\mathbf{x}_t\right) \psi\left(t, \mathbf{x}_+, \mathbf{x}_-\right)}{\sum_{t=1}^{T} \left|\psi\left(t, \mathbf{x}_+, \mathbf{x}_-\right)\right|}\ . \qquad (5.8)
$$

The vector of weights, $\mathbf{u}'$, that maximizes the similarity score over any pair of training sample inputs is used to initialize an optimization procedure to minimize the MSC objective function. The search for the pair of training samples that maximize the similarity score begins with samples at which the largest magnitude of curvature was calculated (Equation 5.3).

5.3.2.2 Optimization

The optimization procedure, initialized within the region of convergence to a local minimum of the MSC objective function, can now be invoked to fully parameterize the student hidden unit, including the output weight:

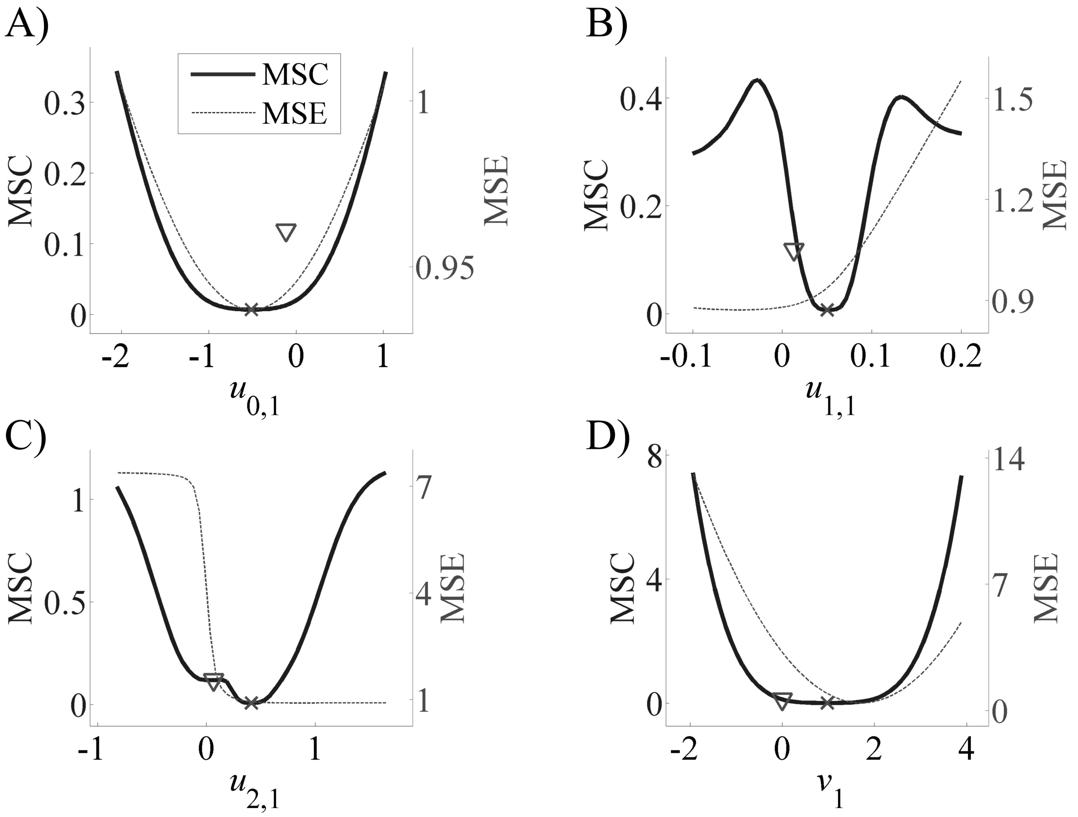


Figure 5.4: Optimization with the MSC objective function.
A) MSC and MSE objective functions along the $u_{0,1}$ axis. The inverted triangle and the cross mark the values of $u_{0,1}$ prior to and after optimization, respectively. Similarly, **B)**, **C)**, and **D)** depict the objective functions along the $u_{1,1}$, $u_{2,1}$, and v_1 weight axes.

$$\{\hat{\mathbf{u}}, \hat{v}_1\} = argmin_{\{\mathbf{u},v_1\}\subset\mathbb{R}^4} \frac{1}{T}\sum_{t=1}^{T} \kappa'\left(\mathbf{x}_t\right)^2 \ , \tag{5.9}$$

where $\kappa'\left(\mathbf{x}_t\right)$ is the difference surface curvature at training sample input $\mathbf{x}_t$ for a student network with the candidate parameterization of the hidden unit being optimized.

This optimization problem can be solved with any off-the-shelf optimization package, such as an unconstrained line search, since the initialization step (see Figure 5.3B) already places the weights close to a minimum.

5.3.3 Iterative Hidden Unit Matching

The parameters found to minimize the MSC objective function above reliably recover one set of weights of a teacher hidden unit, and, when included in the student

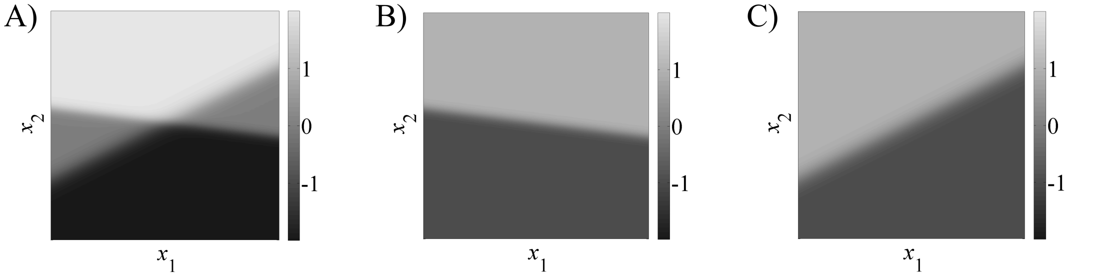


Figure 5.5: Cancellation of a teacher hidden unit by the student network.
A) The teacher network introduced in Figure 5.2. **B)** The first student hidden unit matched to the teacher network. **C)** The difference surface between the teacher and student networks after one hidden unit is matched.

network, effectively cancel the contribution of this hidden unit to the error between the teacher and student network. This technique can be repeated to find the next most prominently contributing hidden unit by reevaluating the difference surface after adding the hidden unit to the student network and, again, parameterizing a new zero student hidden unit.

After each teacher hidden unit is sequentially matched and canceled, the MSC approaches zero and the difference surface is essentially flat; the stop condition is met. The final parameter to learn in the student network is the output bias, which is calculated as the mean of the difference surface over all training samples:

$$v_0 = \frac{1}{T}\sum_{t=1}^{T} d_t \ . \tag{5.10}$$

5.4 Results

Duplicating the weights of a teacher network requires a new approach, such as our neural network tomography procedure, because minimizing the mean squared error does not guarantee that the resulting student hidden unit weights will be the same as the teacher's. Consider the simple case, a target function approximated by a teacher network with two hidden units and an insufficiently complex student network (i.e. too few hidden units) attempting to learn the target function with only a single hidden unit (see Figure 5.6). Minimizing the MSE will attempt to average across both hidden units of the teacher network. Whereas our approach, minimizing the MSC, will not only uncover

Figure 5.6: Comparison between objective functions.

A) A teacher network output with 2 hidden units to be replicated by the student network. **B)** A student network output learned by minimizing the MSE with one hidden unit. **C)** A student network output learned by minimizing the MSC with one hidden unit. MSE and MSC are the mean square and the mean squared curvature, respectively, of the difference surface between the student and teacher networks.

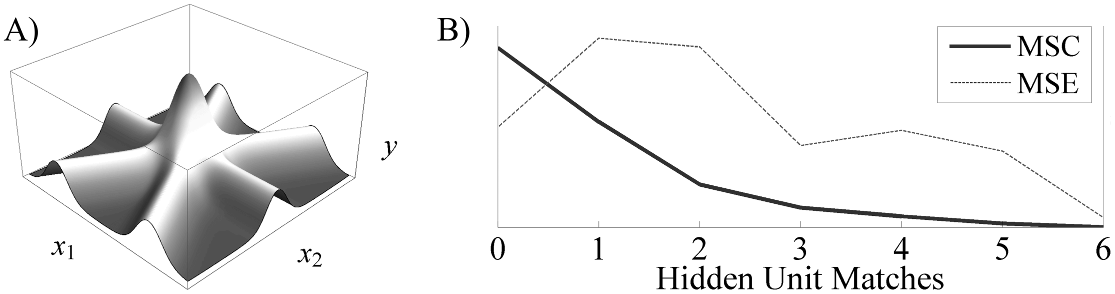


Figure 5.7: MSE progression as hidden units are matched.
A) A teacher network with six hidden units resulting in a radially symmetric output surface. **B)** The steadily decreasing MSC, juxtaposed with the MSE, as the student network replicates each of the hidden units in the teacher network. The dependent axes are normalized for clarity. Note that the MSE does not monotonically decrease as hidden units are matched in the student network, and therefore MSE is surprisingly not very well suited as an objective function for learning.

the true input weights to one of the teacher hidden units, but will also correctly match the output weight.

The radially symmetric teacher network (see Figure 5.7A) is a particularly useful example for contrasting our approach with MSE minimization. Neural network tomography attempts to minimize the MSC objective function with each appended student hidden unit. In doing so, the MSE is not monotonically reduced (see Figure 5.7B). Following a gradient descent algorithm to minimize the MSE objective function will not produce the desired result of matching the student hidden units one at a time.

This technique is also capable of matching the teacher hidden units in more complex networks. Even when presented with a radially symmetric teacher output surface or a randomly weighted teacher network with ten hidden units, the student is able to reliably match the hidden units one at a time (see Figure 5.8).

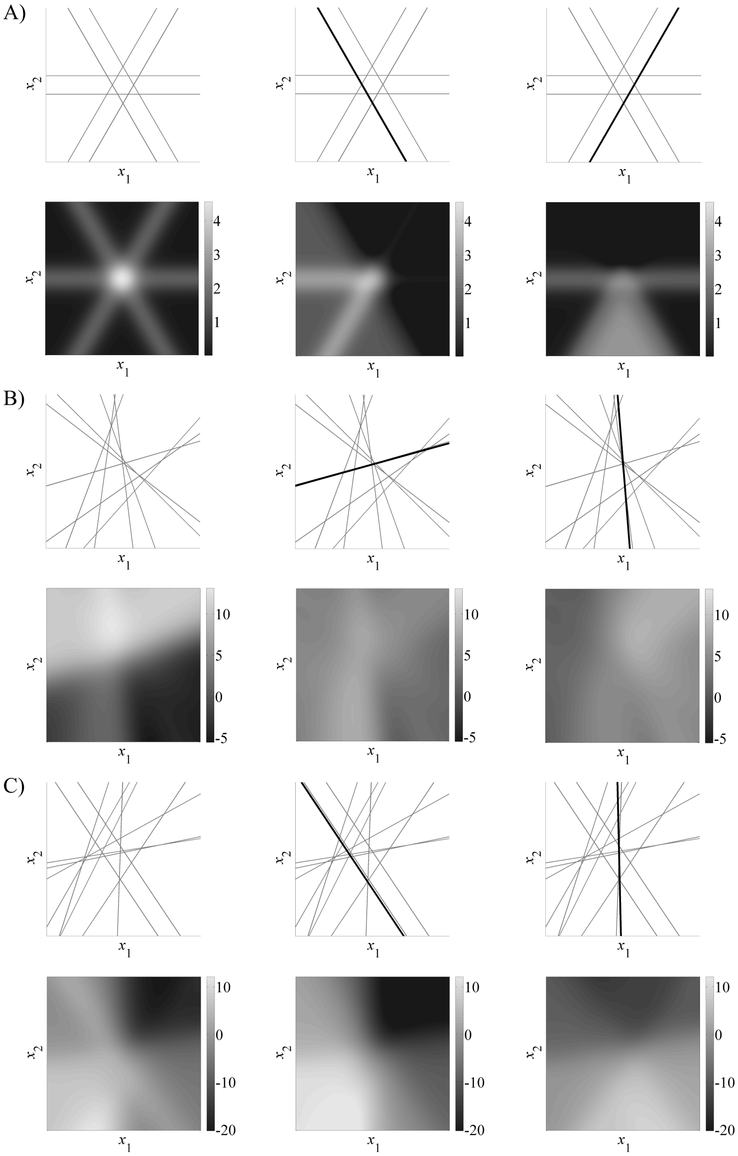


Figure 5.8: First and second hidden unit matches for various teacher networks.
For each of three teacher networks, there are six figures displaying the progress of the student matching hidden units. **A)** The teacher network with six hidden units used in 5.7A. The first row shows the position in input space of the teacher hidden units followed with overlays of the first and second student hidden unit matches. The second row shows the difference surface between the teacher and student networks prior to any matches, after one match, and after two matches. Observe the steadily decreasing total curvature as student hidden unit matches cancel the contribution of teacher hidden units to the error. Similarly **B)** and **C)** show matching results for teacher networks with ten randomly weighted hidden units.

5.4.1 A Note on Sampling Requirements

In order to achieve a reasonable performance with our method, we acknowledge that there are minimum sampling requirements. This method, like backpropagation or any other regression technique, cannot overcome undersampled training data. Therefore, a minimum sampling density is necessary for the student to reliably recover the teacher hidden units. Specifically, each hidden unit in the teacher must be sufficiently sampled by the training data on either side of the $tanh\left(\mathbf{x}_t^\intercal \mathbf{a}_h\right) = 0$ line (i.e. the training data domain must significantly intersect the inflection lines of all teacher hidden units), although we suspect our algorithm might be improved to perform successfully with a lower sampling density than what is currently required.

5.5 Discussion

We have presented a neural network weight-finding technique vastly different from all others of which we are aware. Instead of using every data point to influence every weight, the hidden units of the network that generated the data set are sequentially inferred by analyzing the curvature of that network's output surface. Matching hidden unit parameters in this fashion allows the student network to replicate the teacher network iteratively.

This tomographic method has several advantages. It is straightforward to apply, and requires little of the tuning required with backpropagation (e.g. learning rates or learning rate schedules, and momentum factors). Our technique has the potential to be more desirable in some situations than gradient-descent-based approaches, as its performance is not susceptible to the occasionally counterproductive features of the MSE surface. In addition, this technique is compatible with active-learning techniques, for instance by focusing queries in regions of high-curvature.

As far as we are aware, our algorithm is the first neural network weight-finding procedure that directly matches the underlying weights of a hidden network (the teacher) by analyzing its input-output behavior, a potentially useful method for reverse-engineering of models, or in situations for which only some of the teacher's hidden units are sought.

One drawback to our tomographic method is that it is not as universally applicable as backpropagation is to training neural networks. Backpropagation affords the use of any differentiable loss function. In addition, neural network tomography does not, as of yet, have proved bounds on the sampling density required to be guaranteed the ability to distinguish between similarly aligned hidden units, and we have not established that its performance is gracefully degraded as the sampling density is lowered. It is our expectation that these issues will soon be resolved, and our belief that neural network tomography has great potential.

Chapter 5, in full, is a modified reprint of the material as it appears in Neural Networks. Minnett, Rupert C.J.; Smith, Andrew T.; Lennon Jr., William C.; Hecht-Nielsen, Robert, Elsevier, 2011. The dissertation author was the primary investigator and author of this material.

5.6 Chapter References

Amari, S. (1967). A theory of adaptive pattern classifiers. *IEEE Trans. on Electronic Computers*, 16(3), 299 –307.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251–276.

Chen, A. M., Lu, H., & Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation*, 5(6), 910–927.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7), 909–996.

Daugman, J. G. (1988). Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(7), 1169–1179.

Farin, G. (2002). *Curves and surfaces for CAGD: a practical guide.* San Francisco: Morgan Kaufmann Publishers Inc.

Fourier, J. B. J. (1808). *Mémoire sur la propagation de la chaleur dans les corps solides.* Paris: Bulletin de la Sociétė Philomatique.

Fourier, J. B. J. (1822). *Théorie analytique de la chaleur.* Paris: Firmin Didot.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–67.

Fukumizu, K. & Amari, S. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptions. *Neural Networks*, 13(3), 317–327.

Gabor, D. (1946). Theory of communication. *J. Inst. Elec. Eng.*, 93(3), 429–457.

Galushkin, A. I. (2007). *Neural Networks Theory.* Berlin; New York: Springer.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium: In Sectionibus Conicis Solem Ambientium*. Hamburg: F. Perthes and I. H. Besser.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson Education, Inc., 3rd edition.

Hecht-Nielsen, R. (1989). *Neurocomputing*. Boston: Addison-Wesley Longman Publishing Co., Inc.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.

Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. *IEEE Trans. on Systems, Man and Cybernetics*, 1(4), 364 –378.

Kůrková, V. & Kainen, P. C. (1994). Functionally equivalent feedforward neural networks. *Neural Computation*, 6(3), 543–558.

Legendre, A. (1805). *Nouvelles méthodes pour la détermination des orbites des cométes*. Paris: Firmin Didot.

Minsky, M. L. (1954). *Neural Nets and the Brain Model Problem*. PhD thesis, Princeton University, Princeton.

Newton, I. (1669). *De analysi per æquationes numero terminorum infinitas*. London.

Nilsson, N. J. (1965). *Learning machines; foundations of trainable pattern-classifying systems*. New York: McGraw-Hill.

Oh, S.-K. & Pedrycz, W. (2002). The design of self-organizing polynomial neural networks. *Inf. Sci. Inf. Comput. Sci.*, 141(3-4), 237–258.

Raphson, J. (1690). *Analysis æquationum universalis*. London.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Ithica: Cornell Aeronautical Laboratory.

Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., & Suter, B. W. (1990). The multilayer perceptron as an approximation to a bayes optimal discriminant function. *IEEE Trans. on Neural Networks*, 1(4), 296–298.

Steinbuch, K. (1965). *Automat und Mensch*. Heidelberg: Springer, 3rd edition.

Stigler, S. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9(3), 465–474.

Sussmann, H. J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4), 589 – 593.

Taylor, B. (1715). *Methodus Incrementorum Directa et Inversa*. London.

Wan, E. A. (1990). Neural network classification: A bayesian interpretation. *IEEE Trans. on Neural Networks*, 1(4), 303–305.

Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, Cambridge.

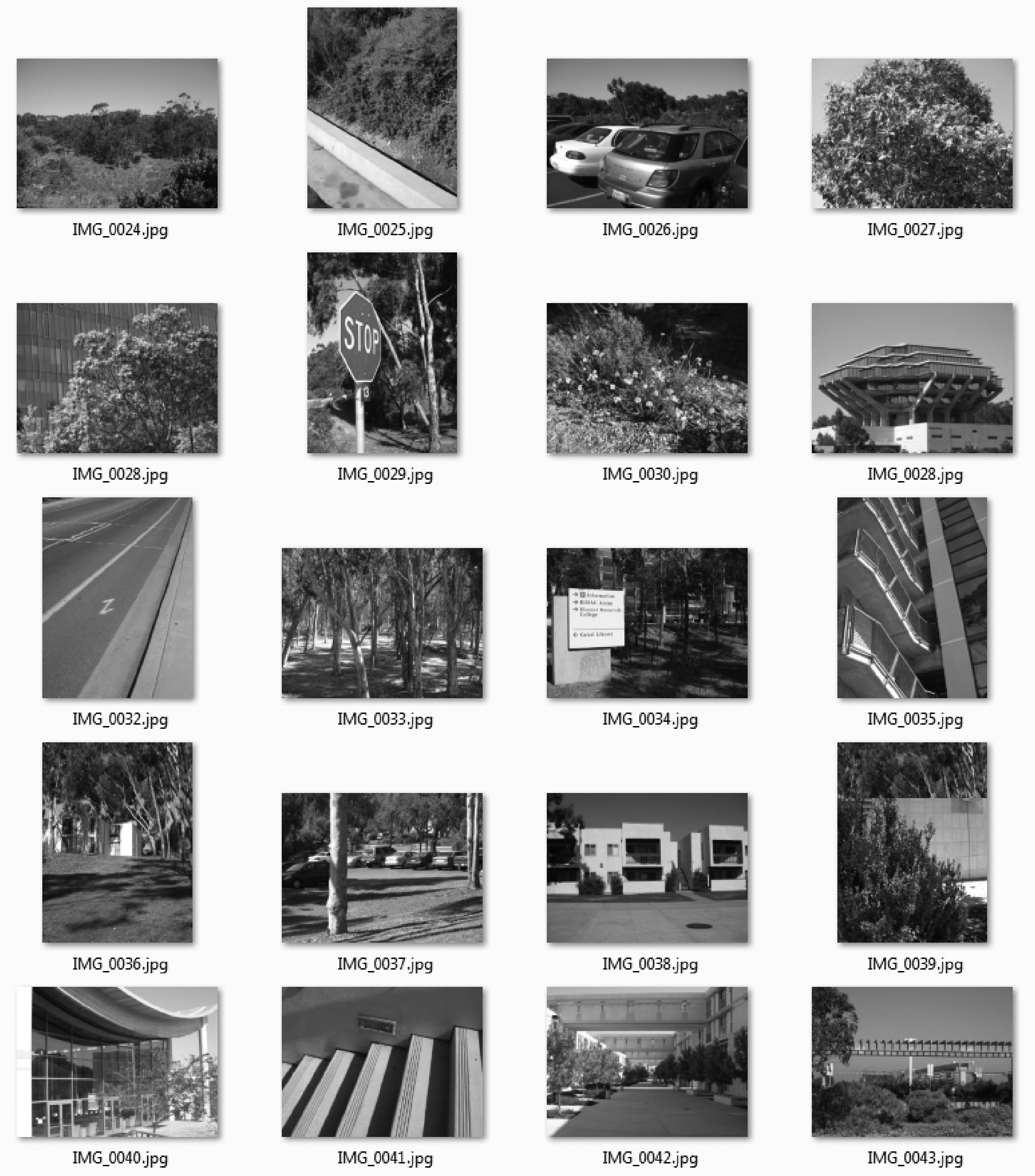
Werbos, P. J. (1994). *The Roots of Backpropagation*. John Wiley & Sons, Inc.

Widrow, B., Hartenstein, R., & Hecht-Nielsen, R. (2005). Eulogy: Karl Steinbuch 1917-2005. *IEEE Computational Intelligence Society*.

Appendix A

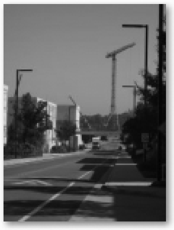
UCSD Natural Scenes Dataset

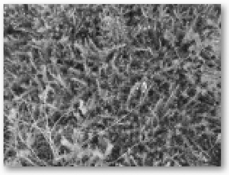
Table A.1: UCSD Natural Scenes Dataset samples 1 - 20.



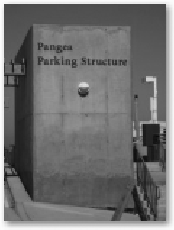
IMG_0024.jpg	IMG_0025.jpg	IMG_0026.jpg	IMG_0027.jpg
IMG_0028.jpg	IMG_0029.jpg	IMG_0030.jpg	IMG_0028.jpg
IMG_0032.jpg	IMG_0033.jpg	IMG_0034.jpg	IMG_0035.jpg
IMG_0036.jpg	IMG_0037.jpg	IMG_0038.jpg	IMG_0039.jpg
IMG_0040.jpg	IMG_0041.jpg	IMG_0042.jpg	IMG_0043.jpg

Table A.2: UCSD Natural Scenes Dataset samples 21 - 40.


IMG_0044.jpg


IMG_0045.jpg

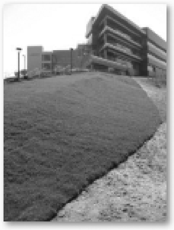
IMG_0046.jpg



IMG_0047.jpg

IMG_0048.jpg

IMG_0049.jpg


IMG_0050.jpg

IMG_0051.jpg

IMG_0052.jpg

IMG_0053.jpg

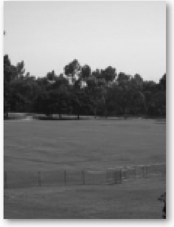
IMG_0054.jpg

IMG_0055.jpg

IMG_0056.jpg

IMG_0057.jpg

IMG_0058.jpg


IMG_0059.jpg

IMG_0060.jpg

IMG_0061.jpg

IMG_0062.jpg

IMG_0063.jpg

Table A.3: UCSD Natural Scenes Dataset samples 41 - 60.

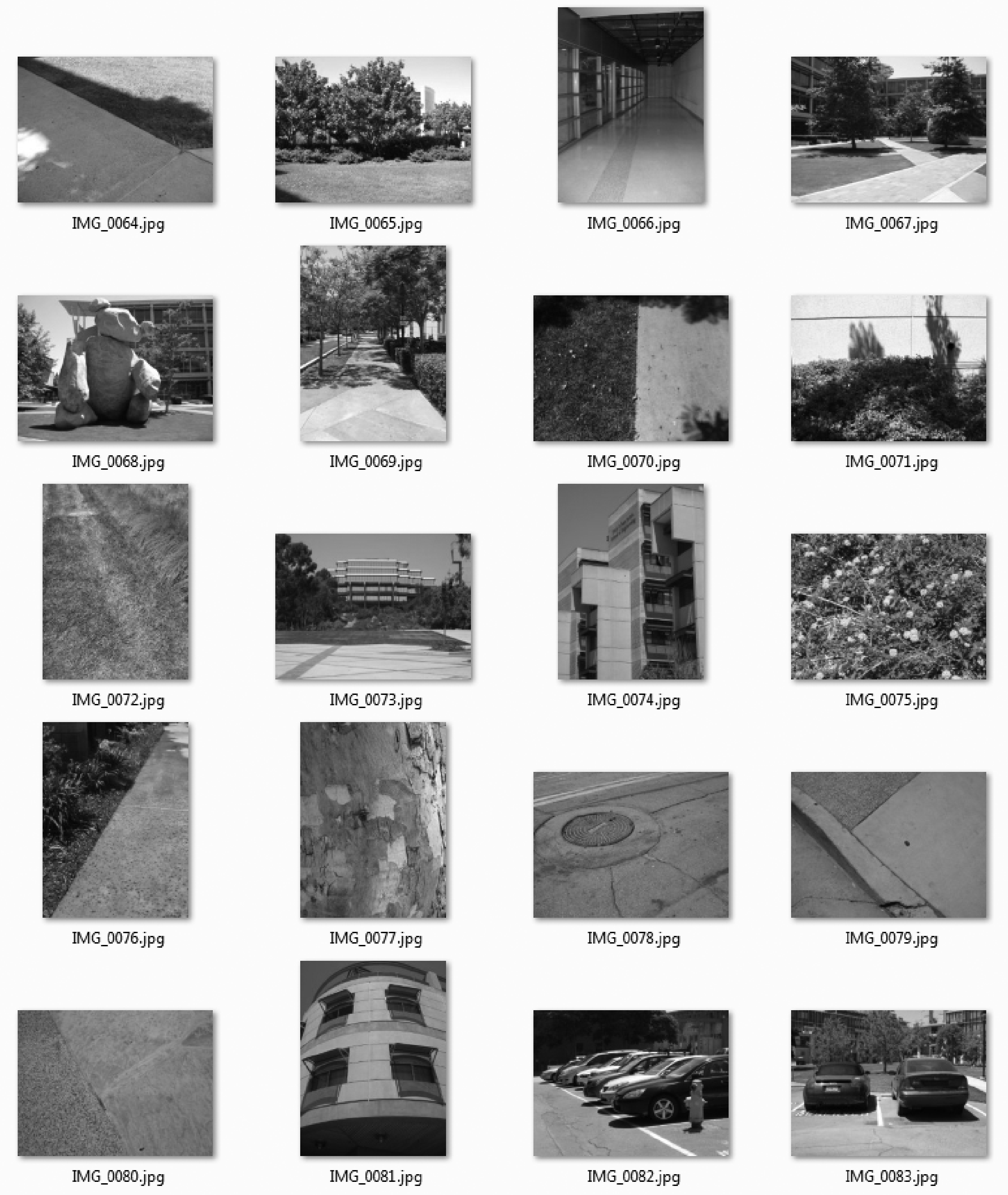


Table A.4: UCSD Natural Scenes Dataset samples 61 - 80.

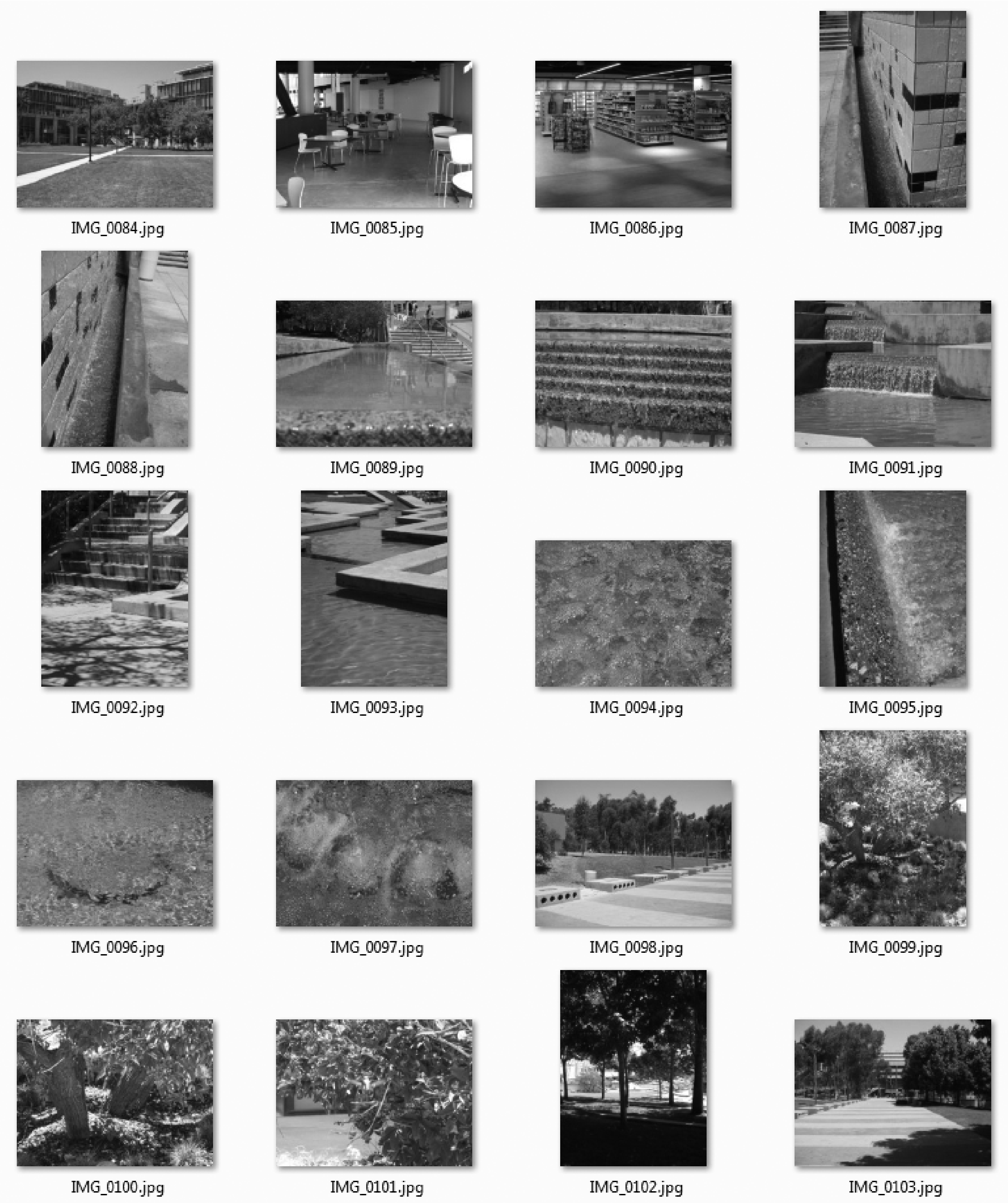
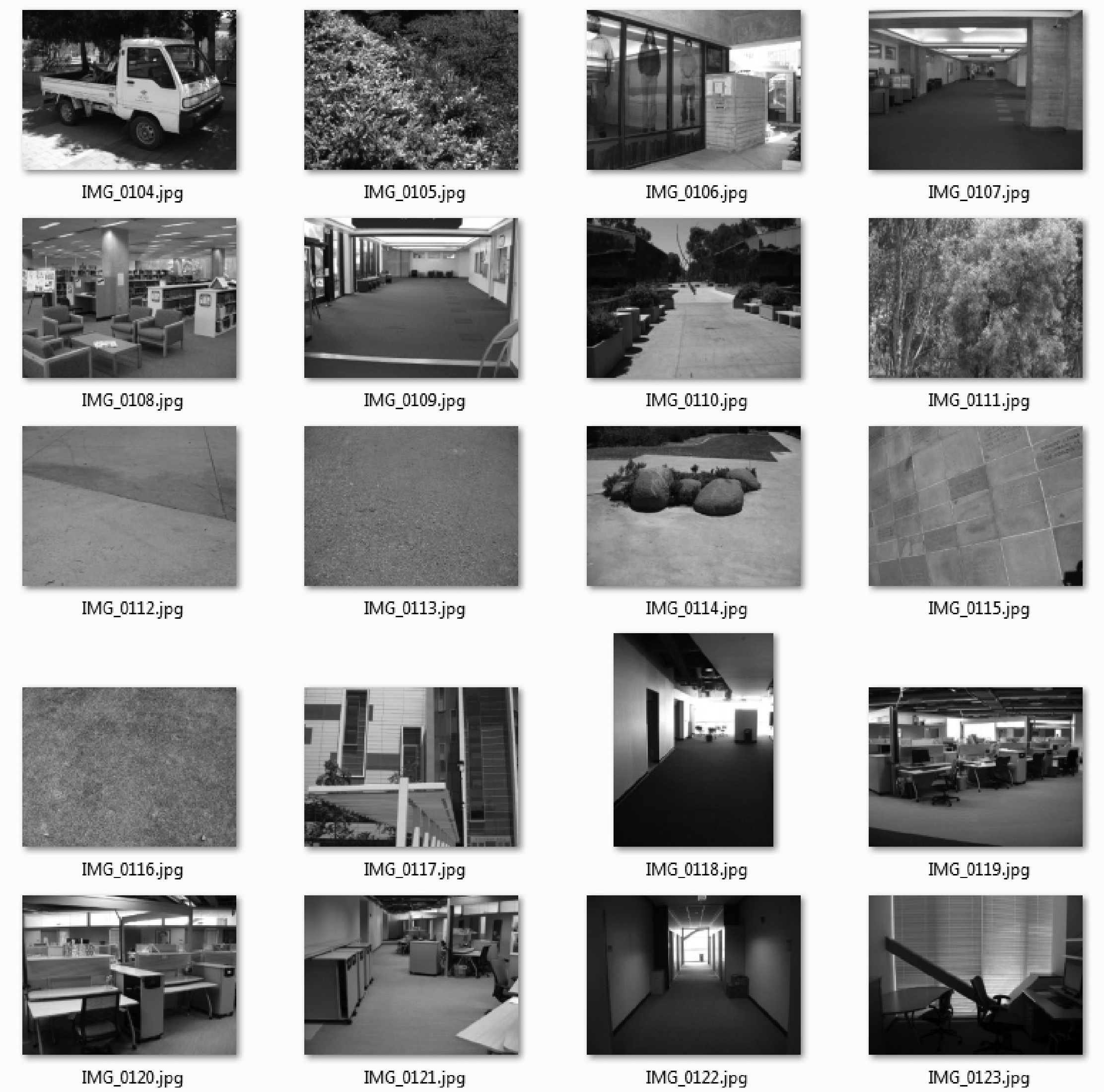


Table A.5: UCSD Natural Scenes Dataset samples 81 - 100.



Appendix B

UIUCTex Dataset

Table B.1: UIUCTex Dataset texture class 1 (bark 1) samples.

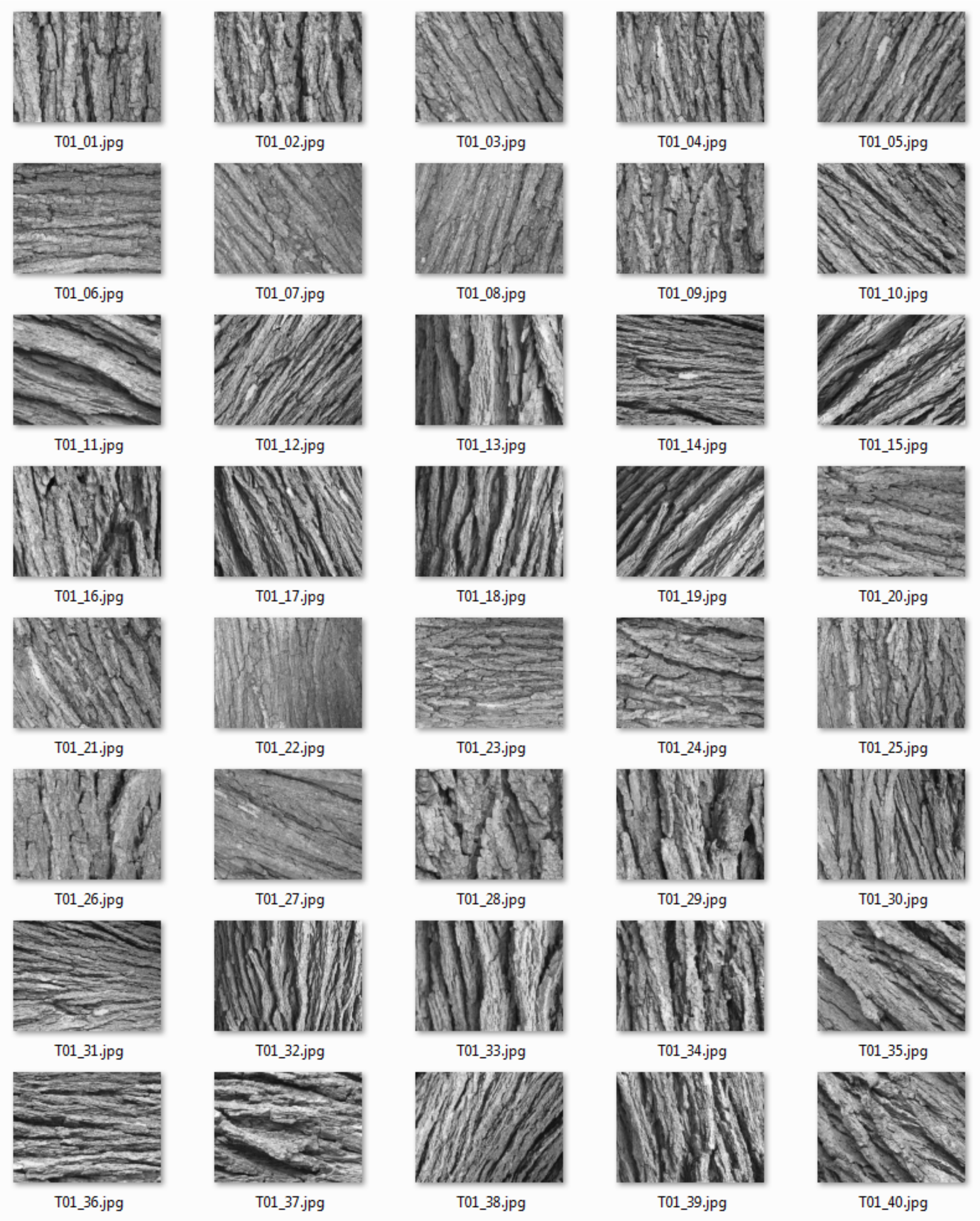


Table B.2: UIUCTex Dataset texture class 2 (bark 2) samples.

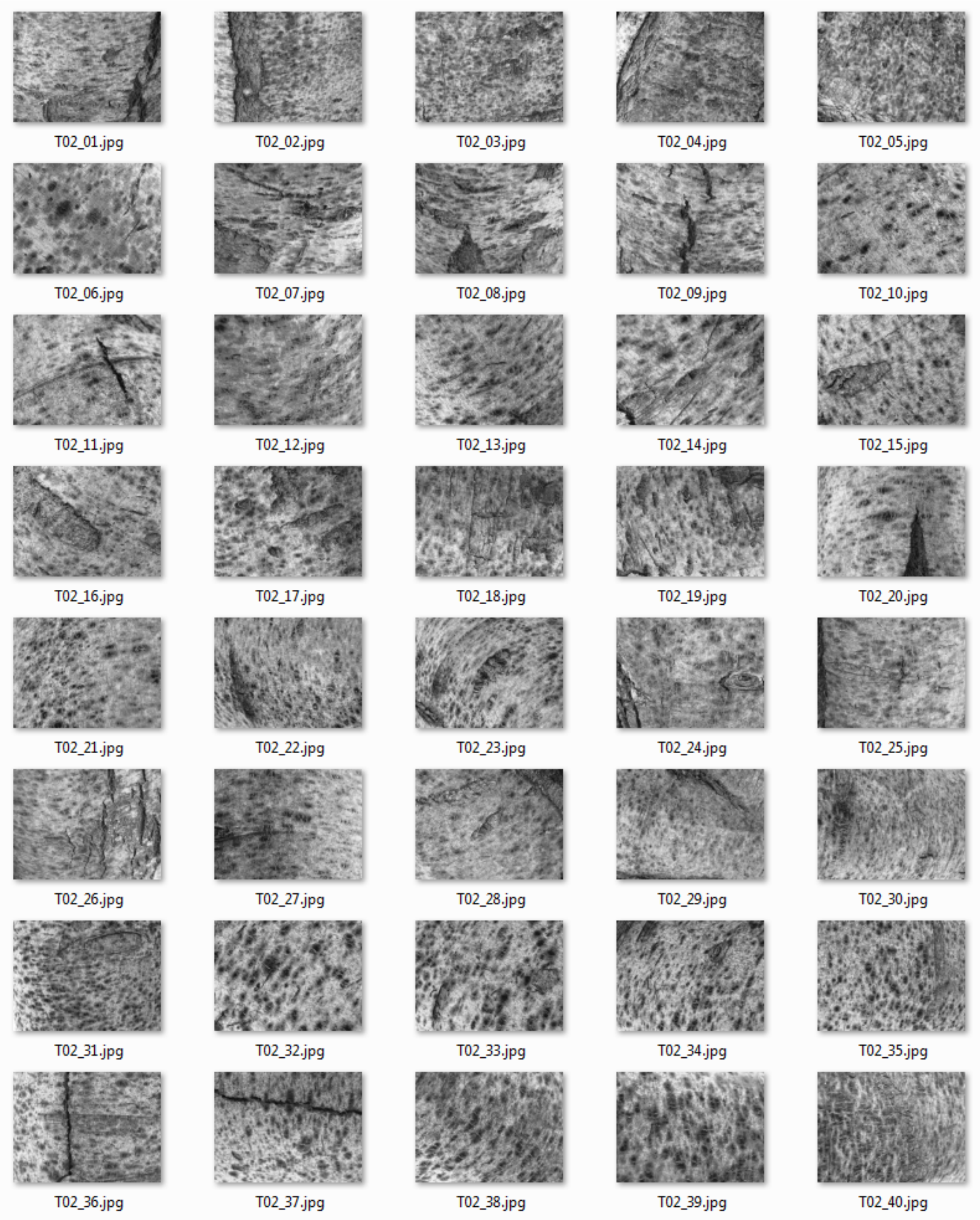


Table B.3: UIUCTex Dataset texture class 3 (bark 3) samples.

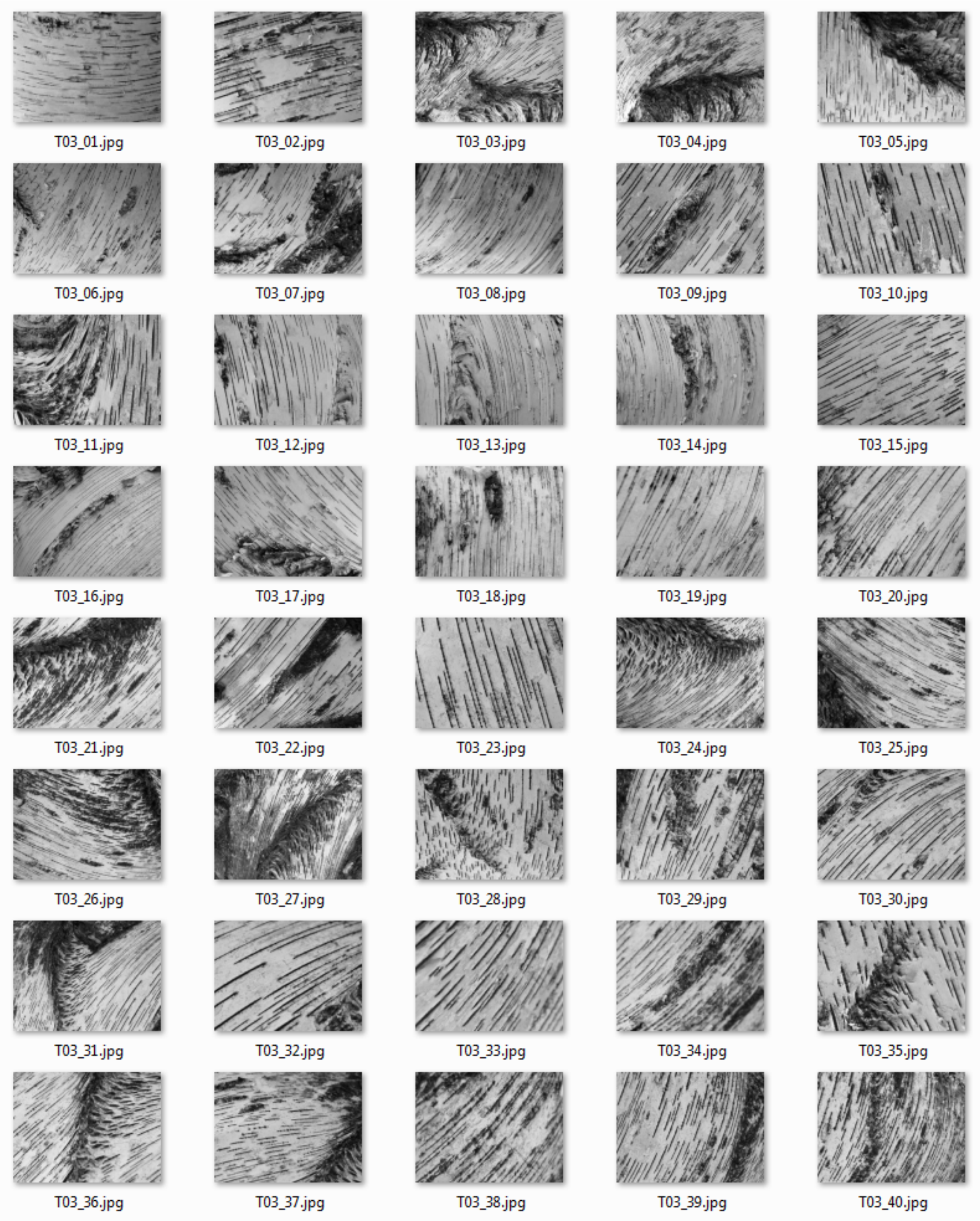


Table B.4: UIUCTex Dataset texture class 4 (wood 1) samples.

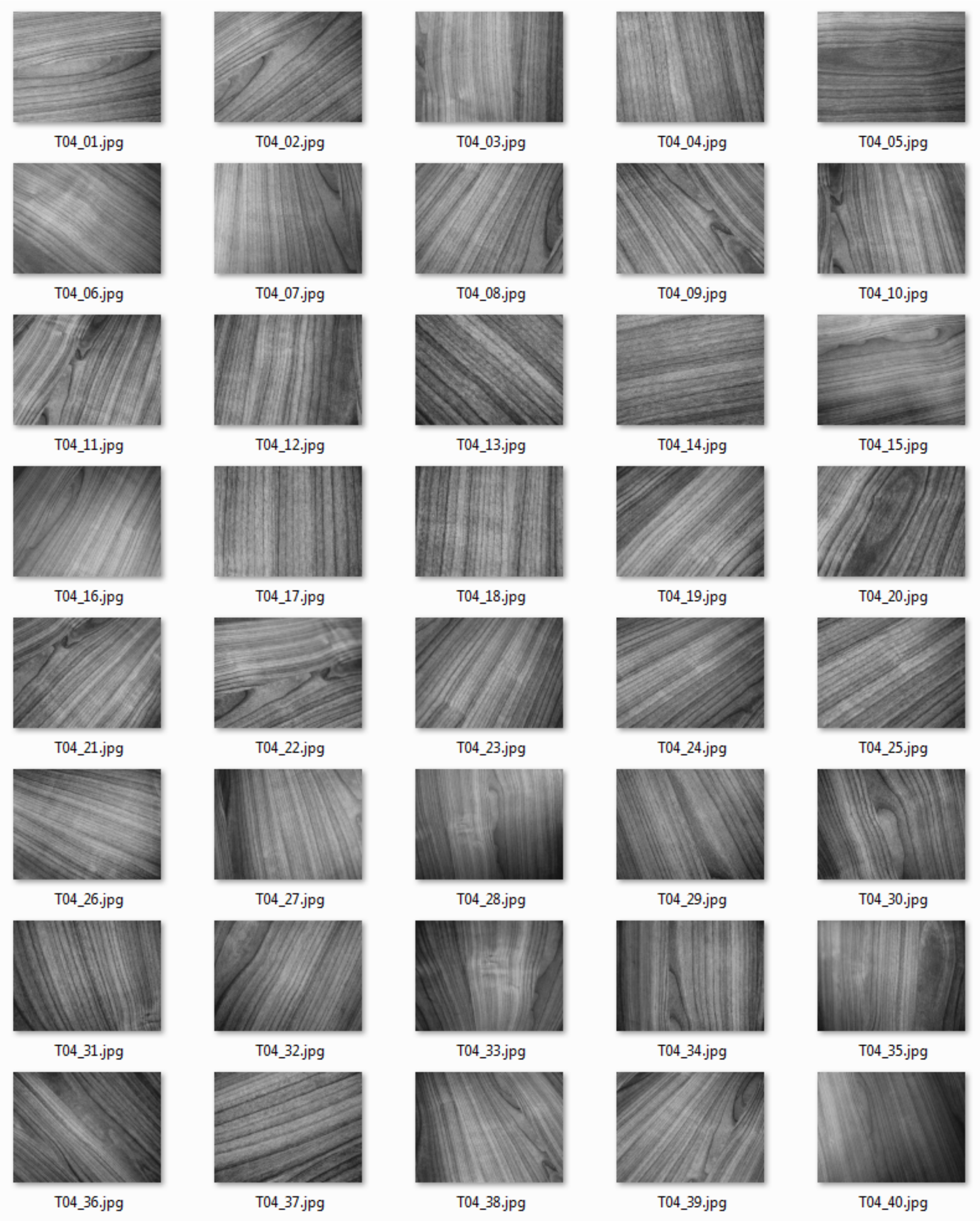


Table B.5: UIUCTex Dataset texture class 5 (wood 2) samples.

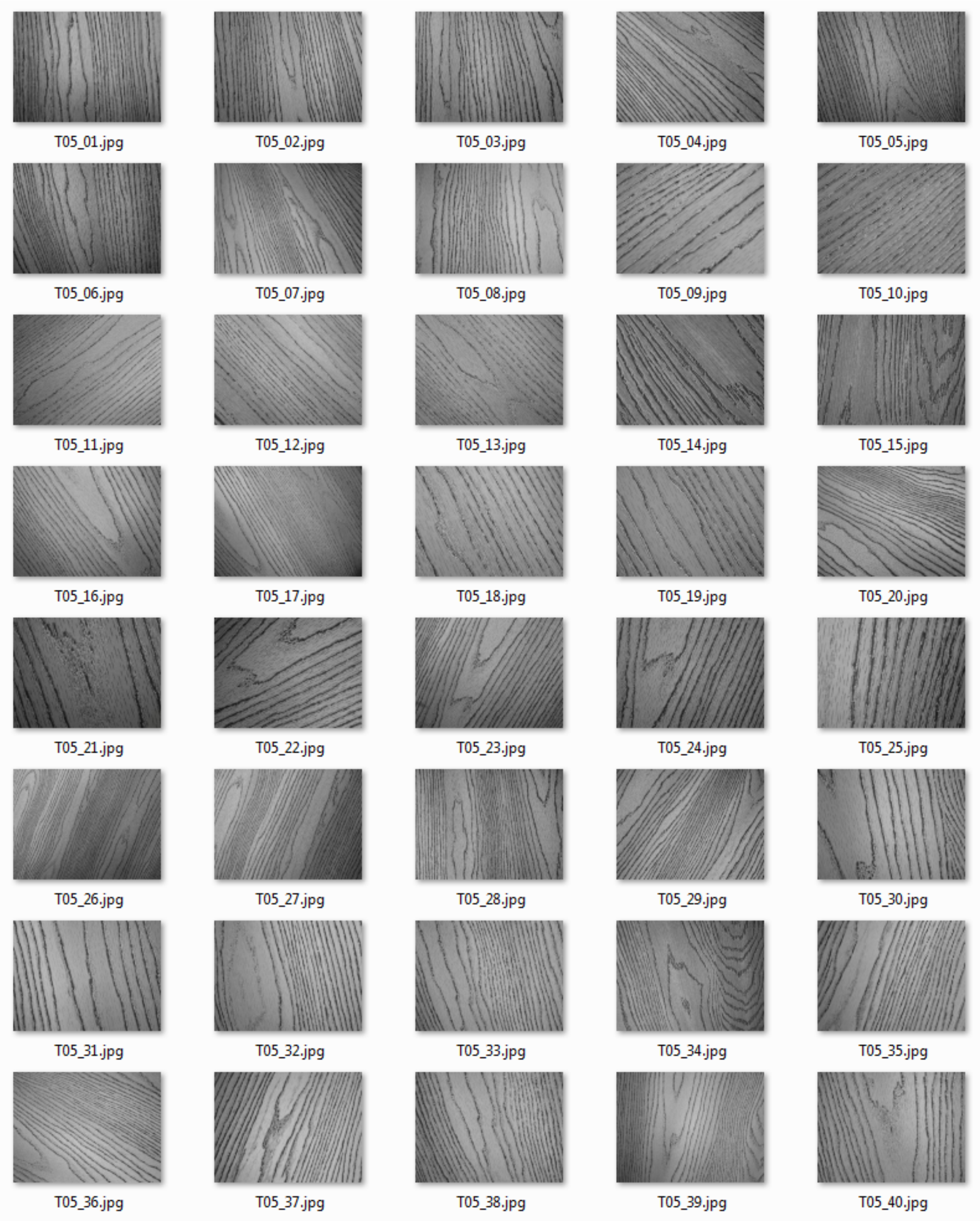


Table B.6: UIUCTex Dataset texture class 6 (wood 3) samples.

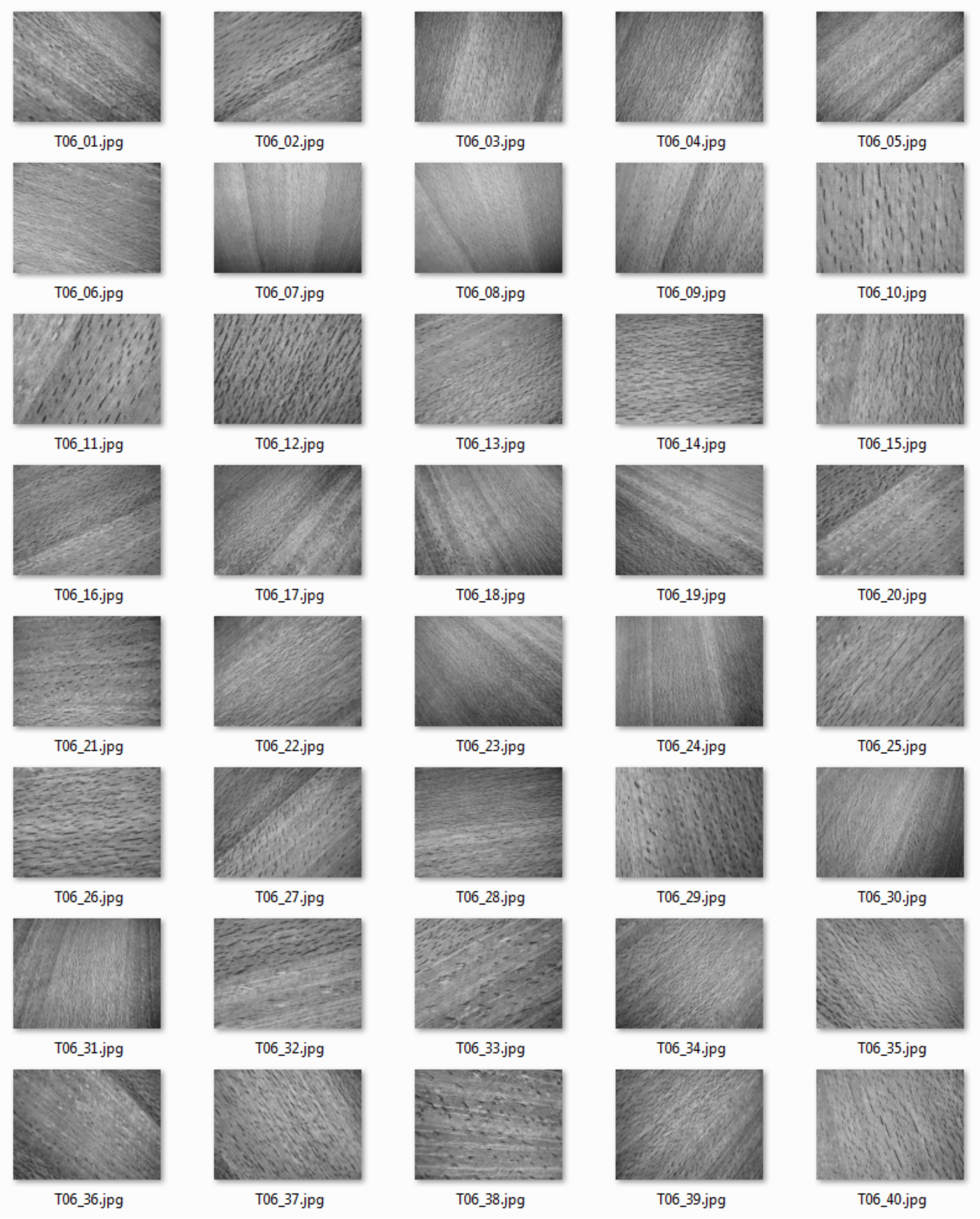


Table B.7: UIUCTex Dataset texture class 7 (water) samples.

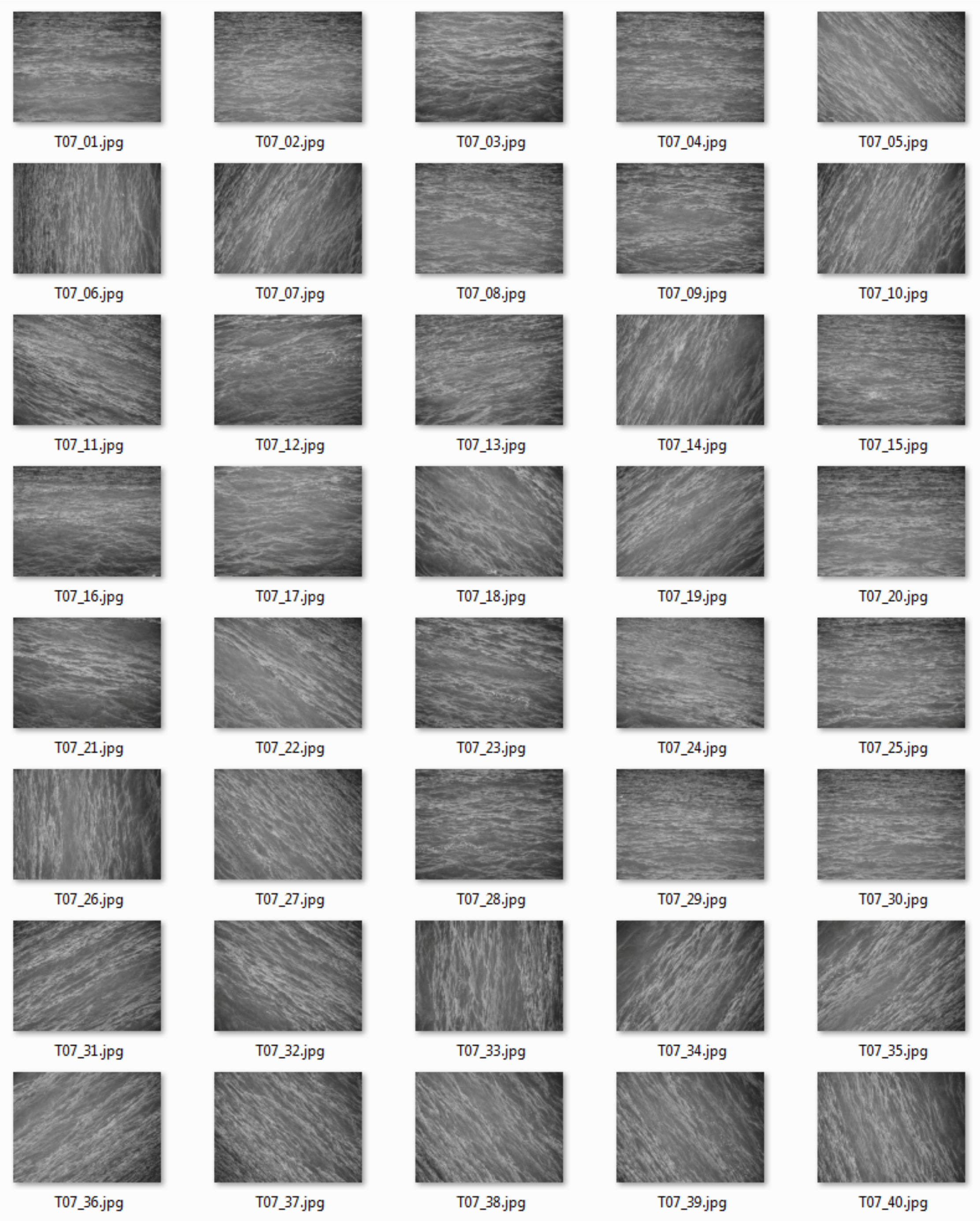


Table B.8: UIUCTex Dataset texture class 8 (granite) samples.

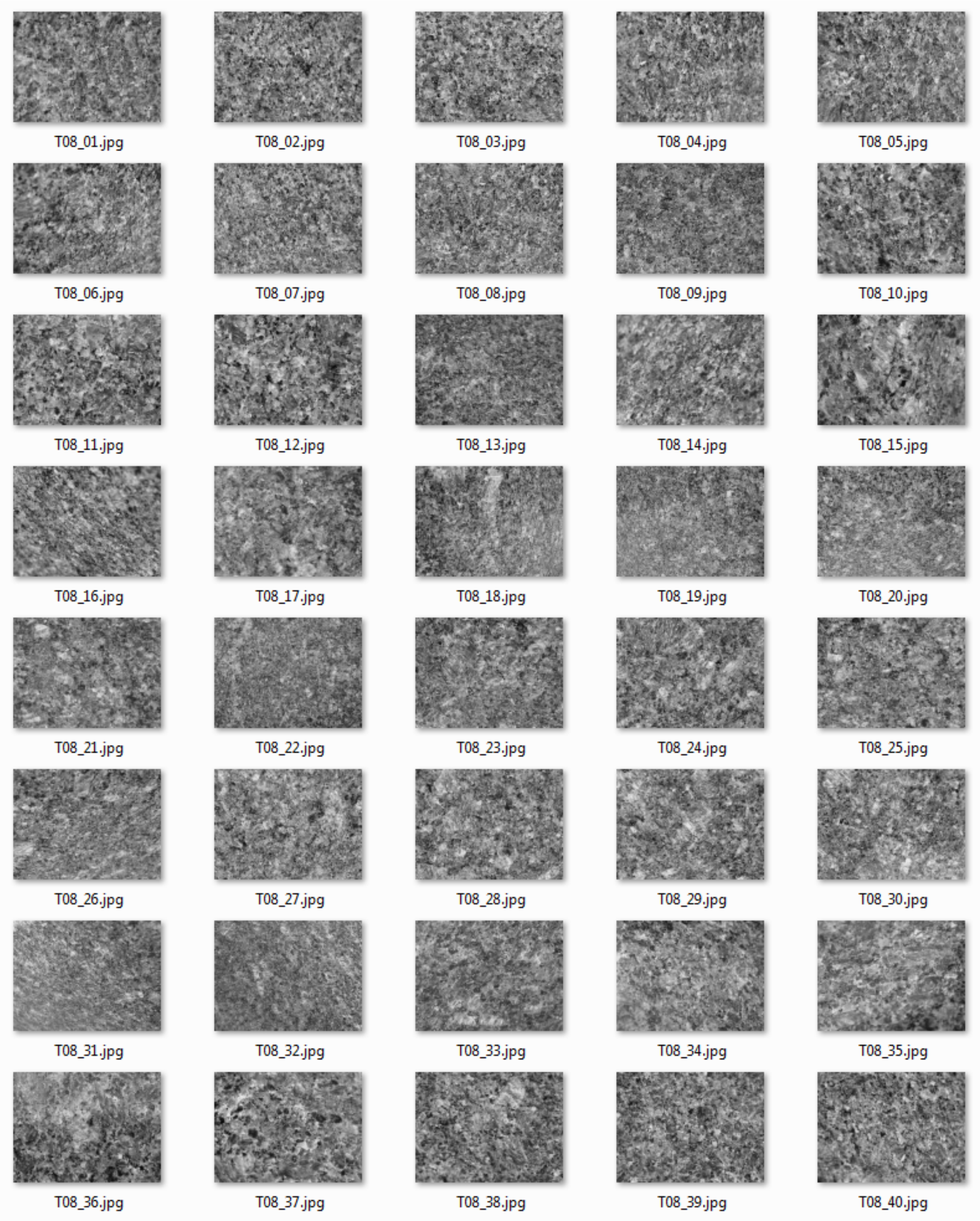


Table B.9: UIUCTex Dataset texture class 9 (marble) samples.

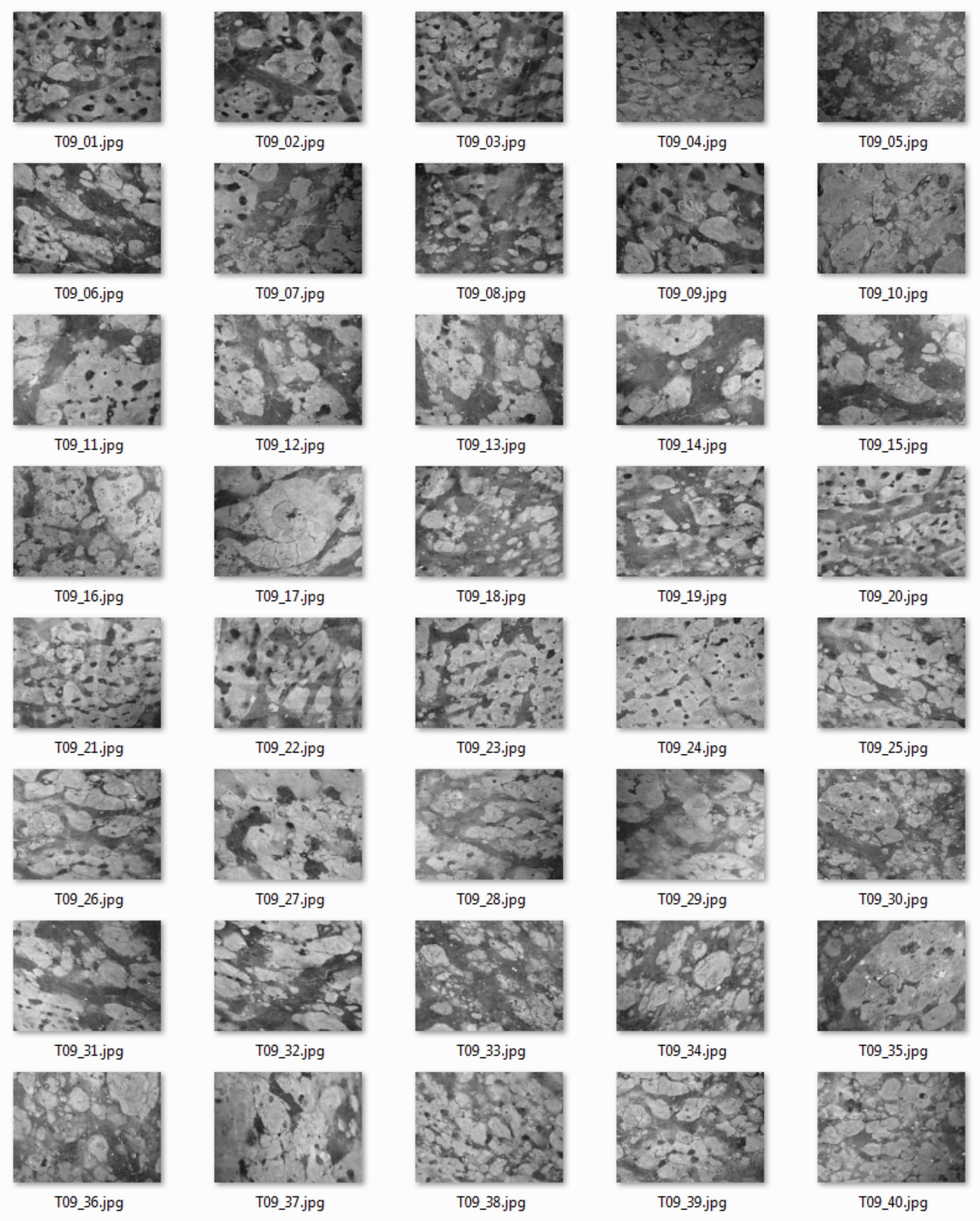


Table B.10: UIUCTex Dataset texture class 10 (floor 1) samples.

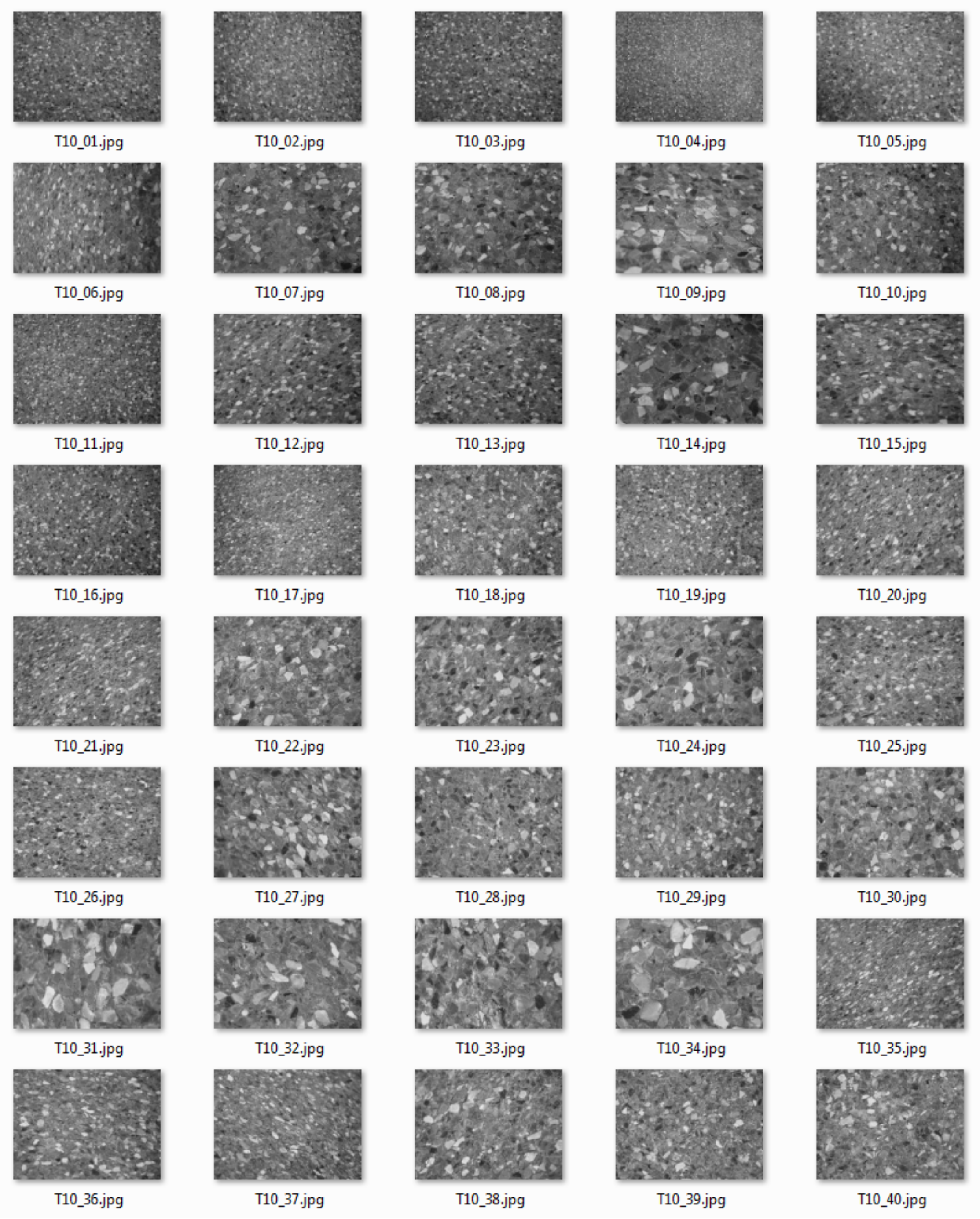


Table B.11: UIUCTex Dataset texture class 11 (floor 2) samples.

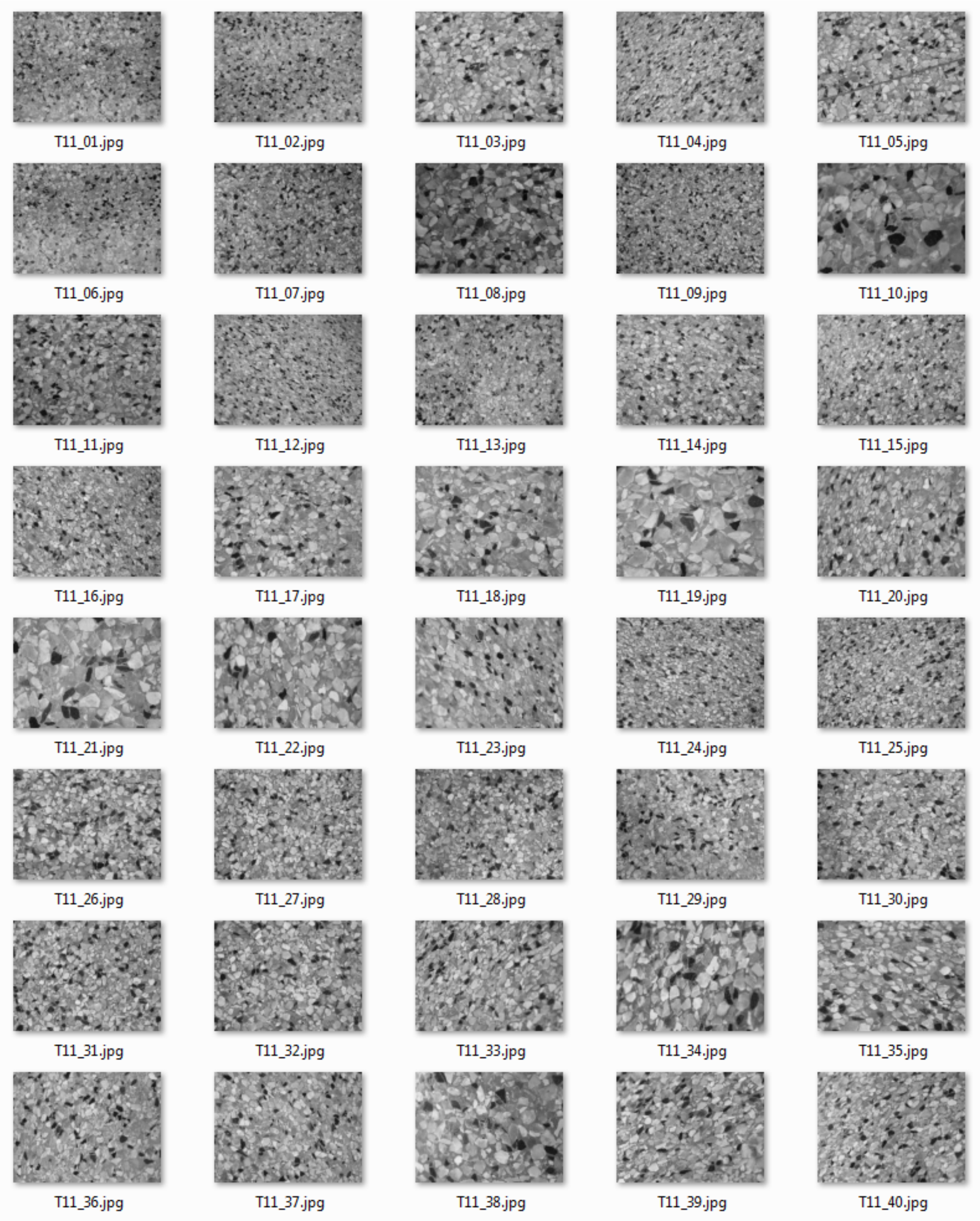


Table B.12: UIUCTex Dataset texture class 12 (pebbles) samples.

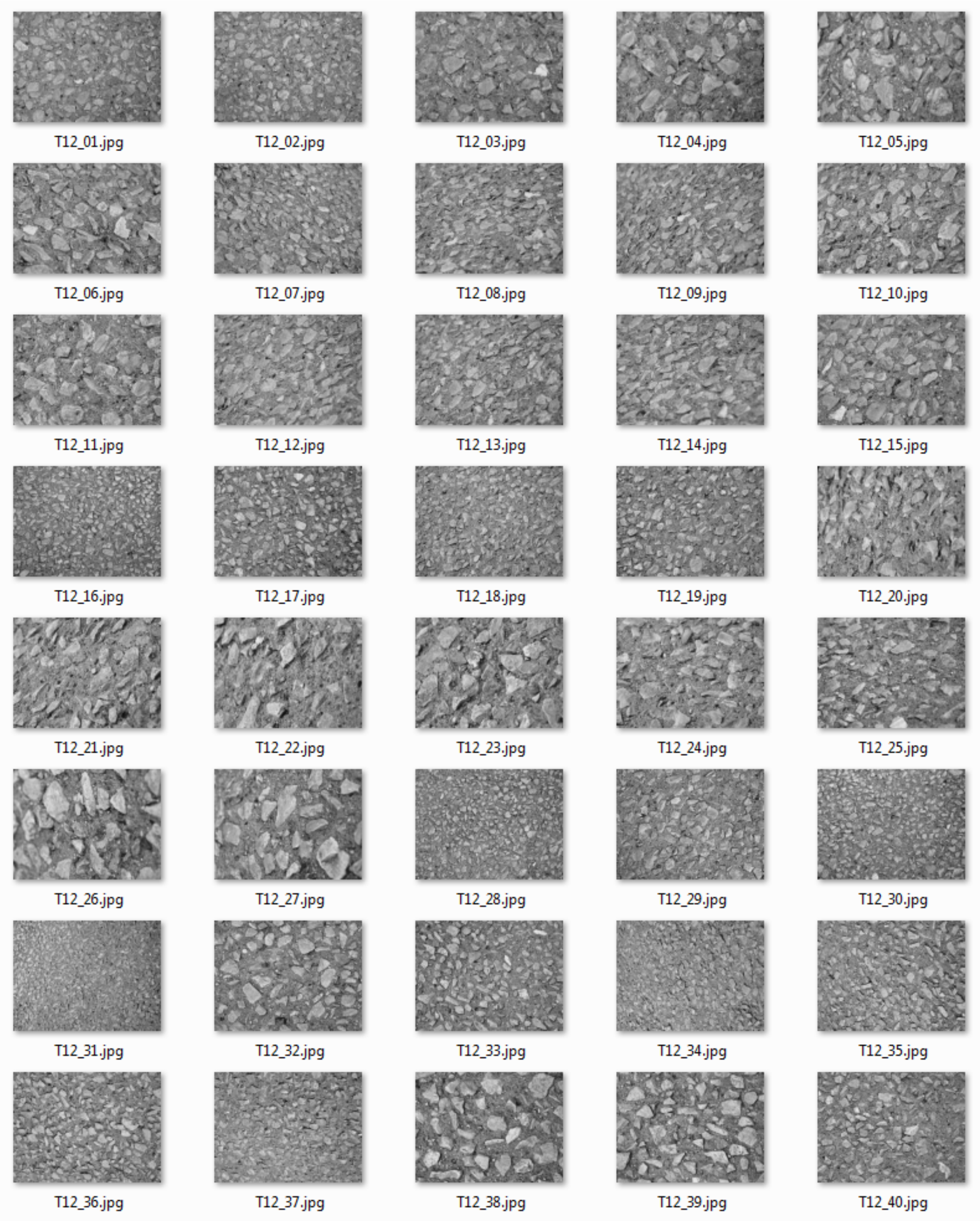


Table B.13: UIUCTex Dataset texture class 13 (wall) samples.

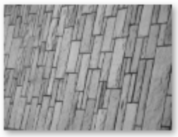
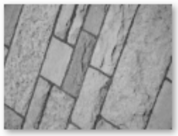
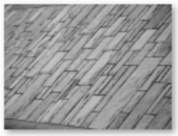
 T13_01.jpg	 T13_02.jpg	 T13_03.jpg	 T13_04.jpg	 T13_05.jpg
 T13_06.jpg	 T13_07.jpg	 T13_08.jpg	 T13_09.jpg	 T13_10.jpg
 T13_11.jpg	 T13_12.jpg	 T13_13.jpg	 T13_14.jpg	 T13_15.jpg
 T13_16.jpg	 T13_17.jpg	 T13_18.jpg	 T13_19.jpg	 T13_20.jpg
 T13_21.jpg	 T13_22.jpg	 T13_23.jpg	 T13_24.jpg	 T13_25.jpg
 T13_26.jpg	 T13_27.jpg	 T13_28.jpg	 T13_29.jpg	 T13_30.jpg
 T13_31.jpg	 T13_32.jpg	 T13_33.jpg	 T13_34.jpg	 T13_35.jpg
 T13_36.jpg	 T13_37.jpg	 T13_38.jpg	 T13_39.jpg	 T13_40.jpg

Table B.14: UIUCTex Dataset texture class 14 (brick 1) samples.

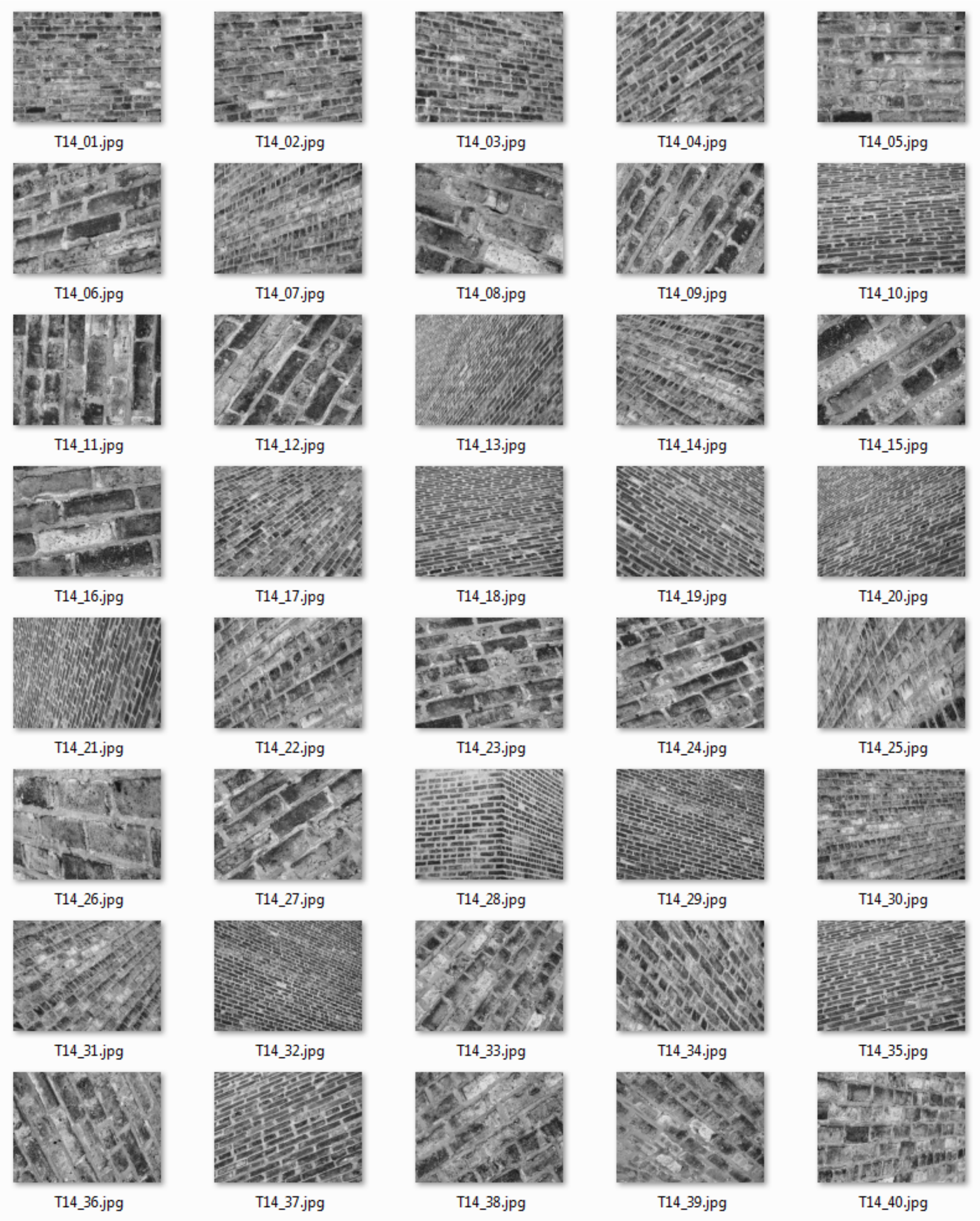
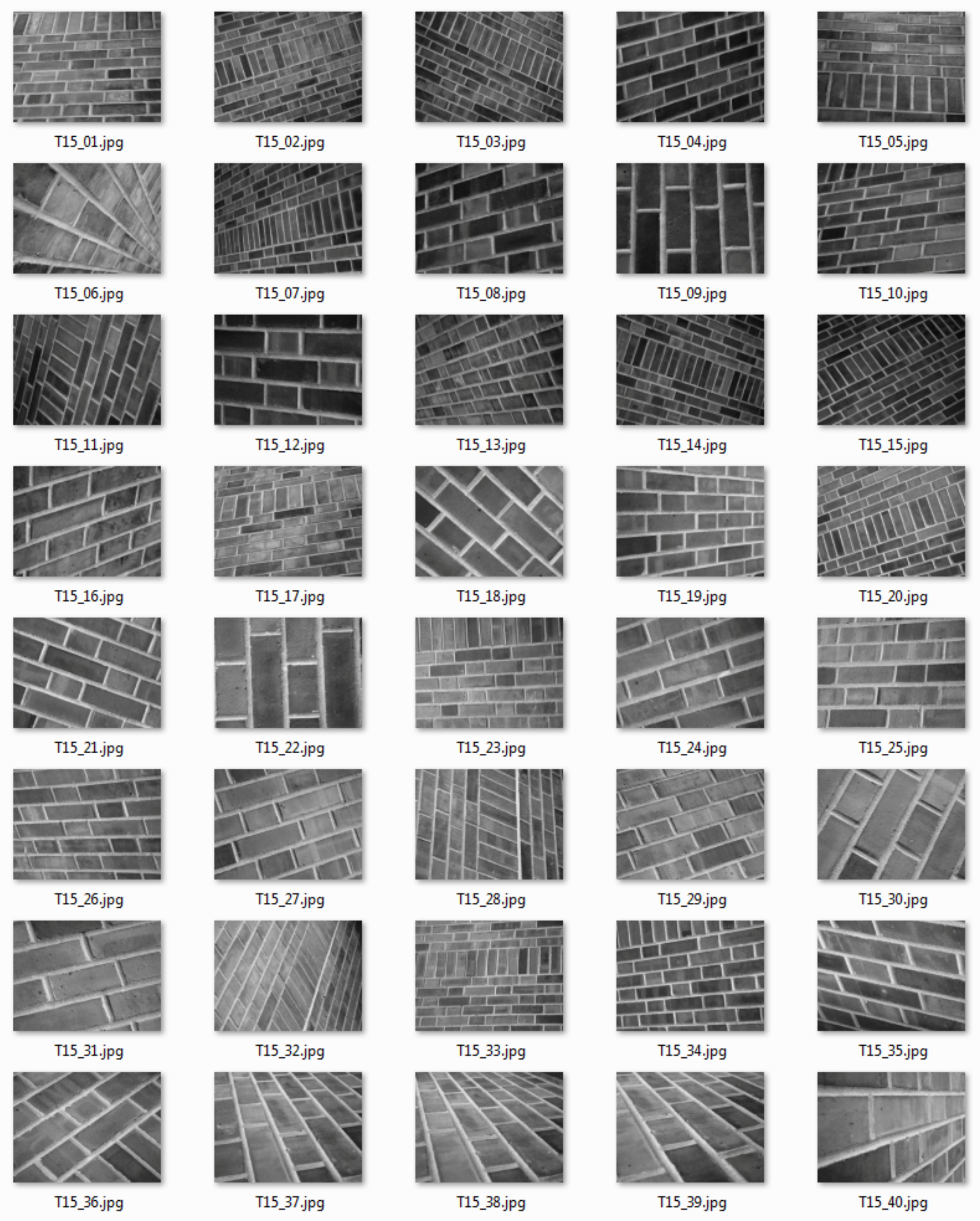


Table B.15: UIUCTex Dataset texture class 15 (brick 2) samples.



T15_01.jpg	T15_02.jpg	T15_03.jpg	T15_04.jpg	T15_05.jpg
T15_06.jpg	T15_07.jpg	T15_08.jpg	T15_09.jpg	T15_10.jpg
T15_11.jpg	T15_12.jpg	T15_13.jpg	T15_14.jpg	T15_15.jpg
T15_16.jpg	T15_17.jpg	T15_18.jpg	T15_19.jpg	T15_20.jpg
T15_21.jpg	T15_22.jpg	T15_23.jpg	T15_24.jpg	T15_25.jpg
T15_26.jpg	T15_27.jpg	T15_28.jpg	T15_29.jpg	T15_30.jpg
T15_31.jpg	T15_32.jpg	T15_33.jpg	T15_34.jpg	T15_35.jpg
T15_36.jpg	T15_37.jpg	T15_38.jpg	T15_39.jpg	T15_40.jpg

Table B.16: UIUCTex Dataset texture class 16 (glass 1) samples.

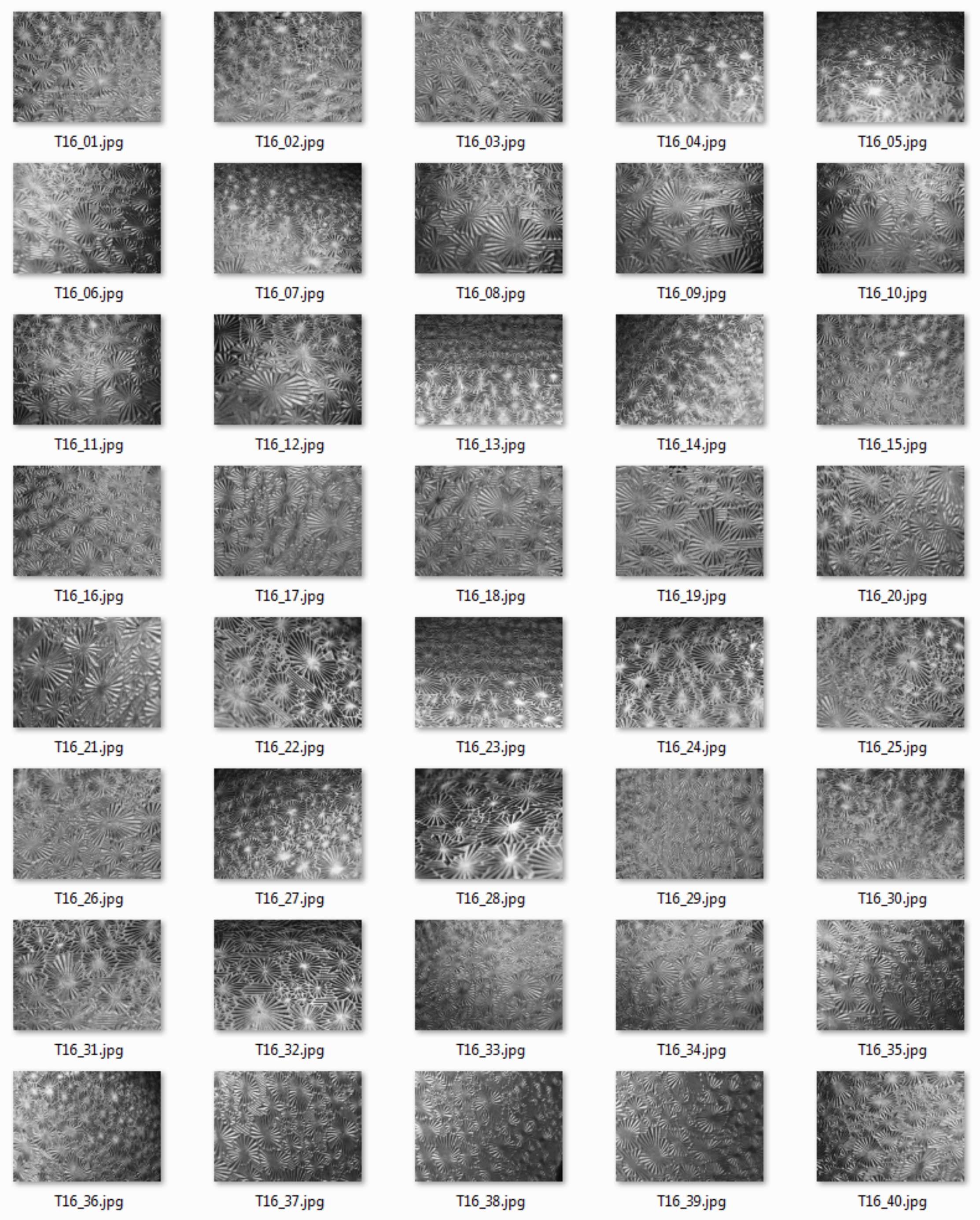


Table B.17: UIUCTex Dataset texture class 17 (glass 2) samples.

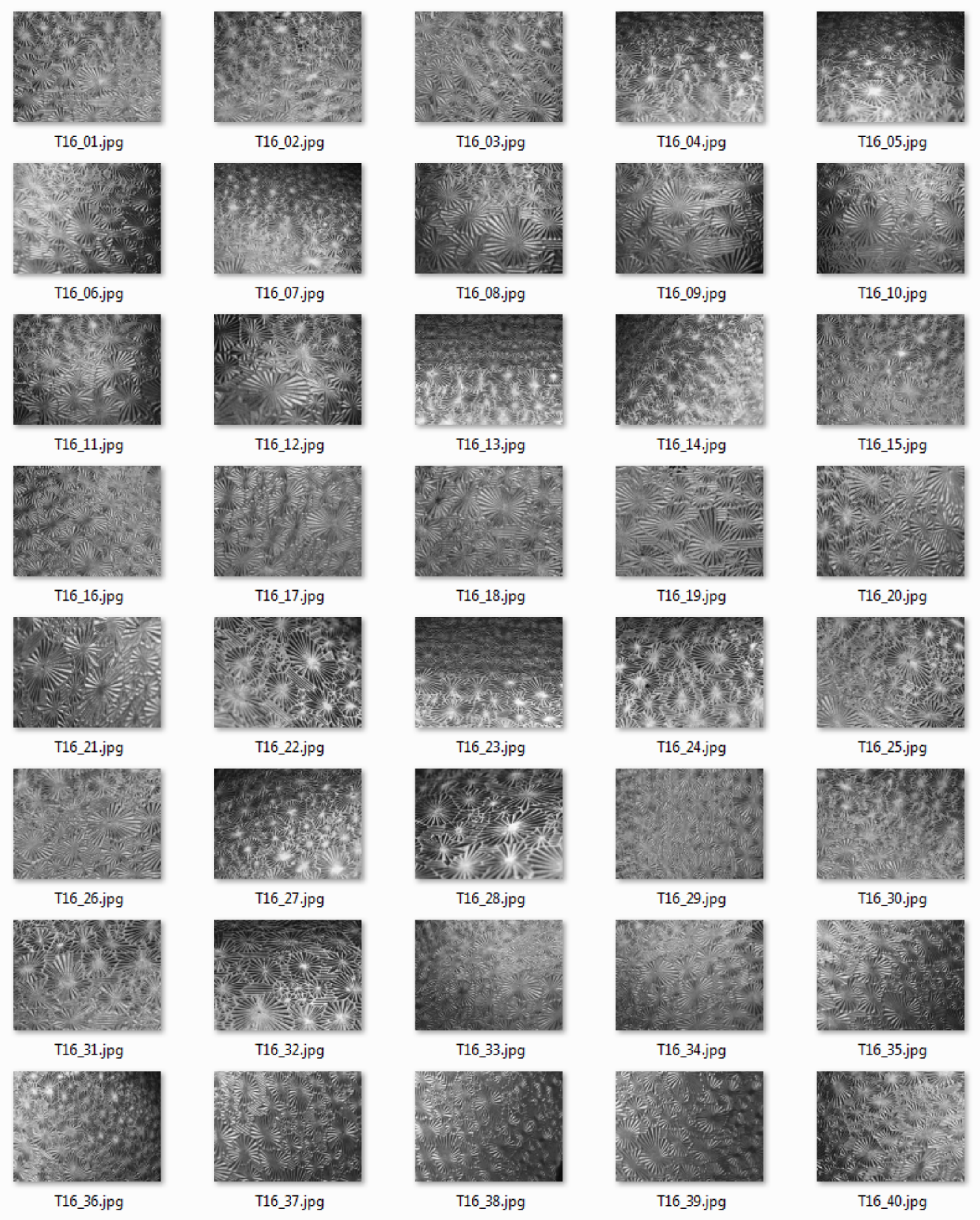


Table B.18: UIUCTex Dataset texture class 18 (carpet 1) samples.

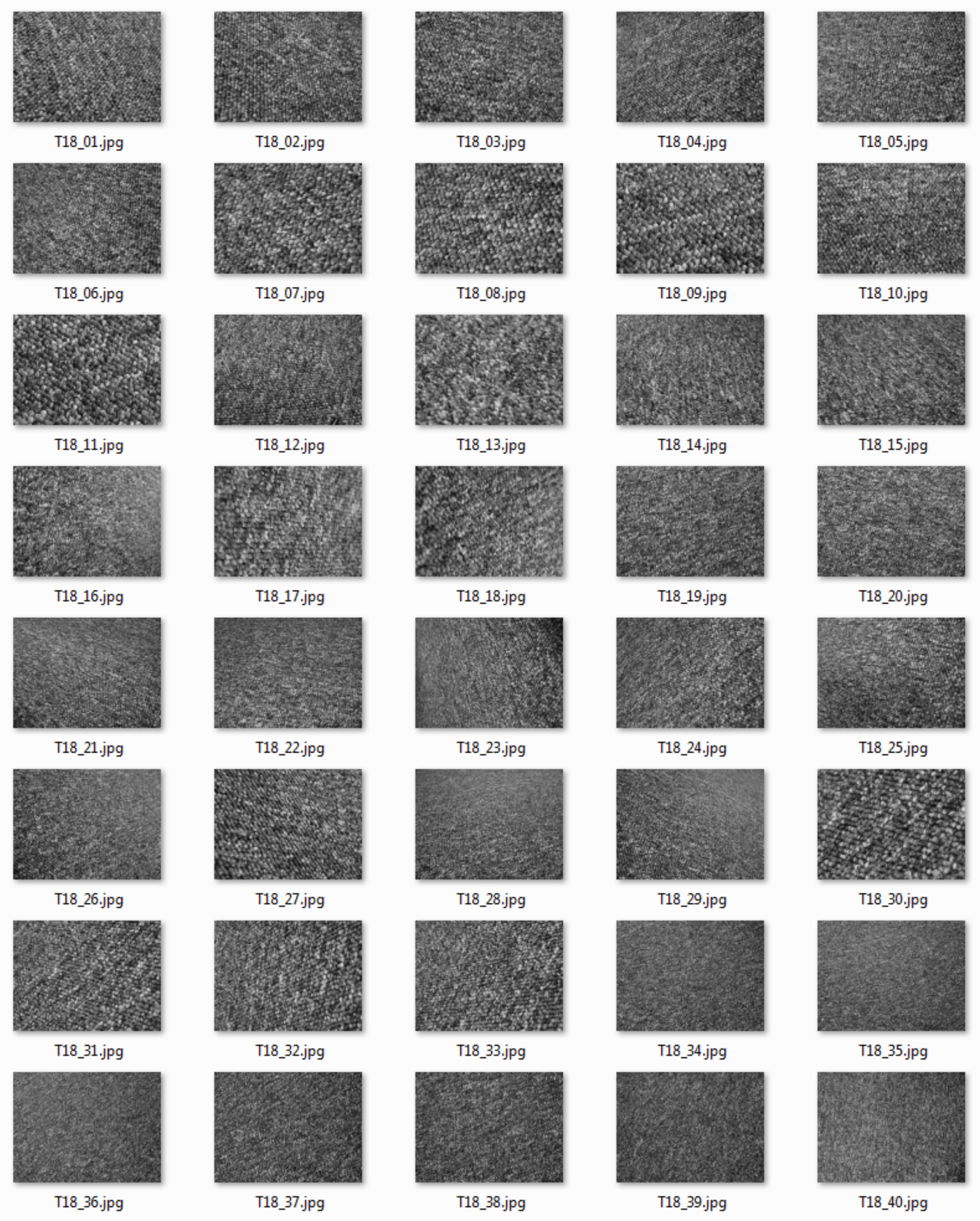


Table B.19: UIUCTex Dataset texture class 19 (carpet 2) samples.

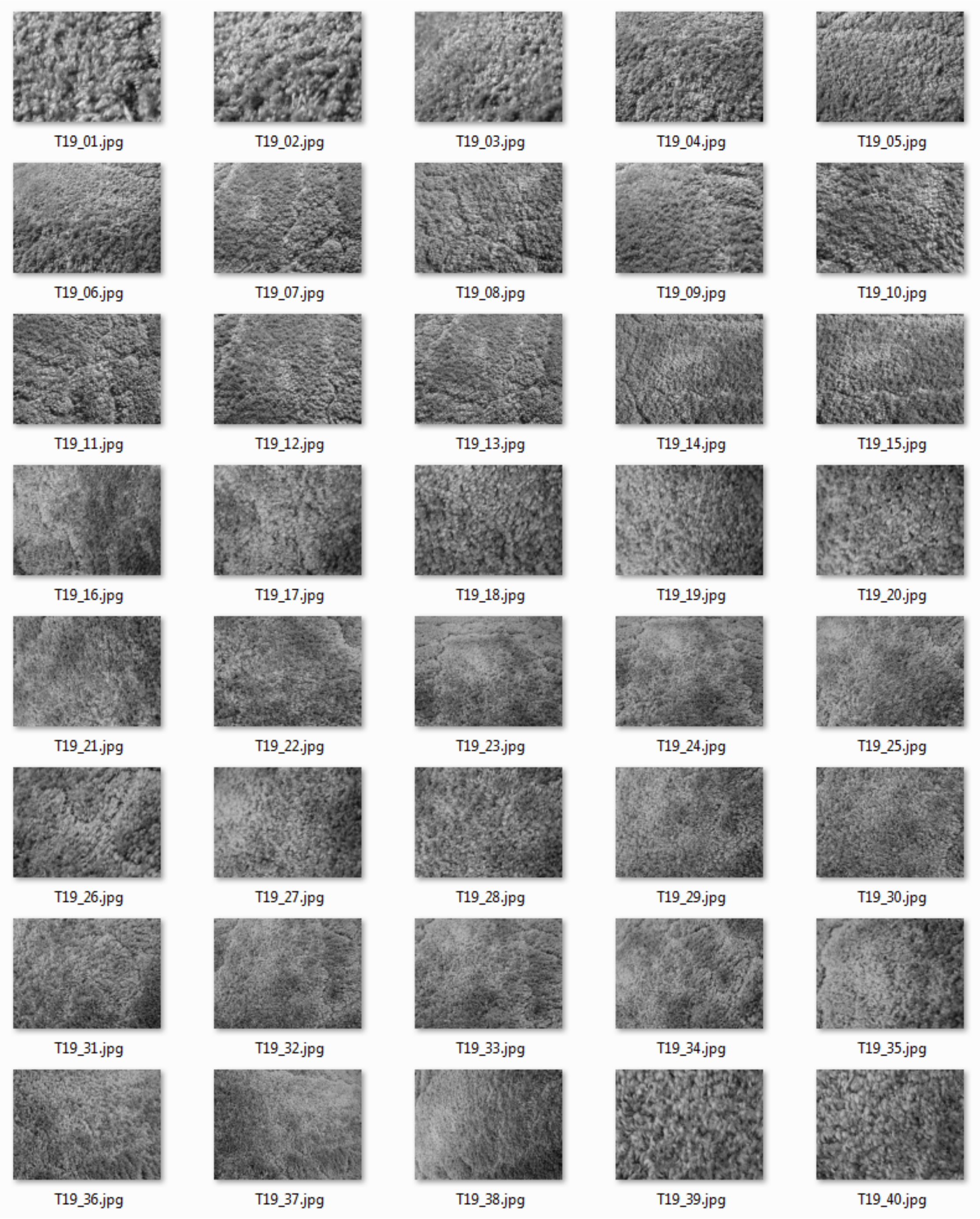
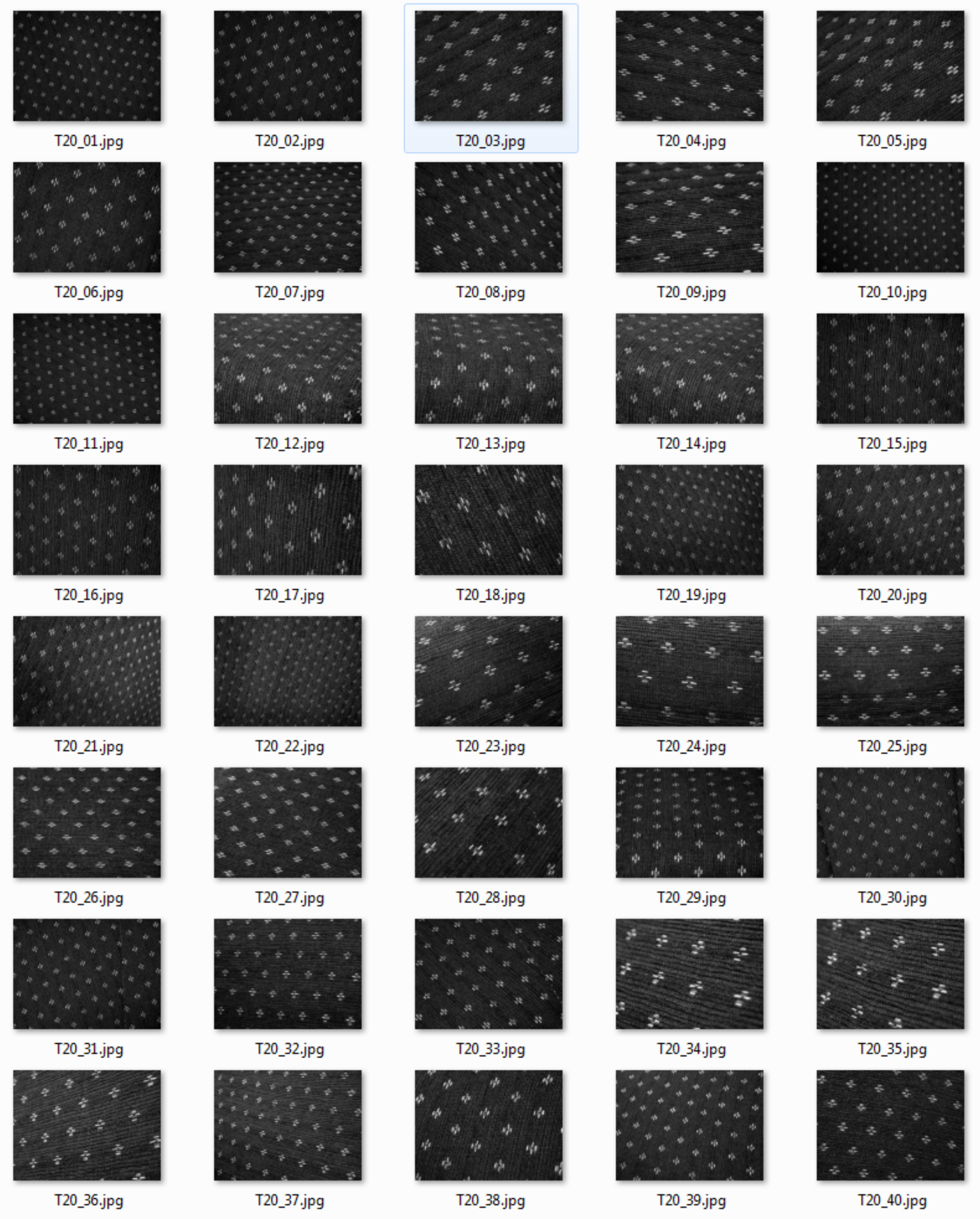


Table B.20: UIUCTex Dataset texture class 20 (upholstery) samples.



T20_01.jpg	T20_02.jpg	T20_03.jpg	T20_04.jpg	T20_05.jpg
T20_06.jpg	T20_07.jpg	T20_08.jpg	T20_09.jpg	T20_10.jpg
T20_11.jpg	T20_12.jpg	T20_13.jpg	T20_14.jpg	T20_15.jpg
T20_16.jpg	T20_17.jpg	T20_18.jpg	T20_19.jpg	T20_20.jpg
T20_21.jpg	T20_22.jpg	T20_23.jpg	T20_24.jpg	T20_25.jpg
T20_26.jpg	T20_27.jpg	T20_28.jpg	T20_29.jpg	T20_30.jpg
T20_31.jpg	T20_32.jpg	T20_33.jpg	T20_34.jpg	T20_35.jpg
T20_36.jpg	T20_37.jpg	T20_38.jpg	T20_39.jpg	T20_40.jpg

Table B.21: UIUCTex Dataset texture class 21 (wallpaper) samples.

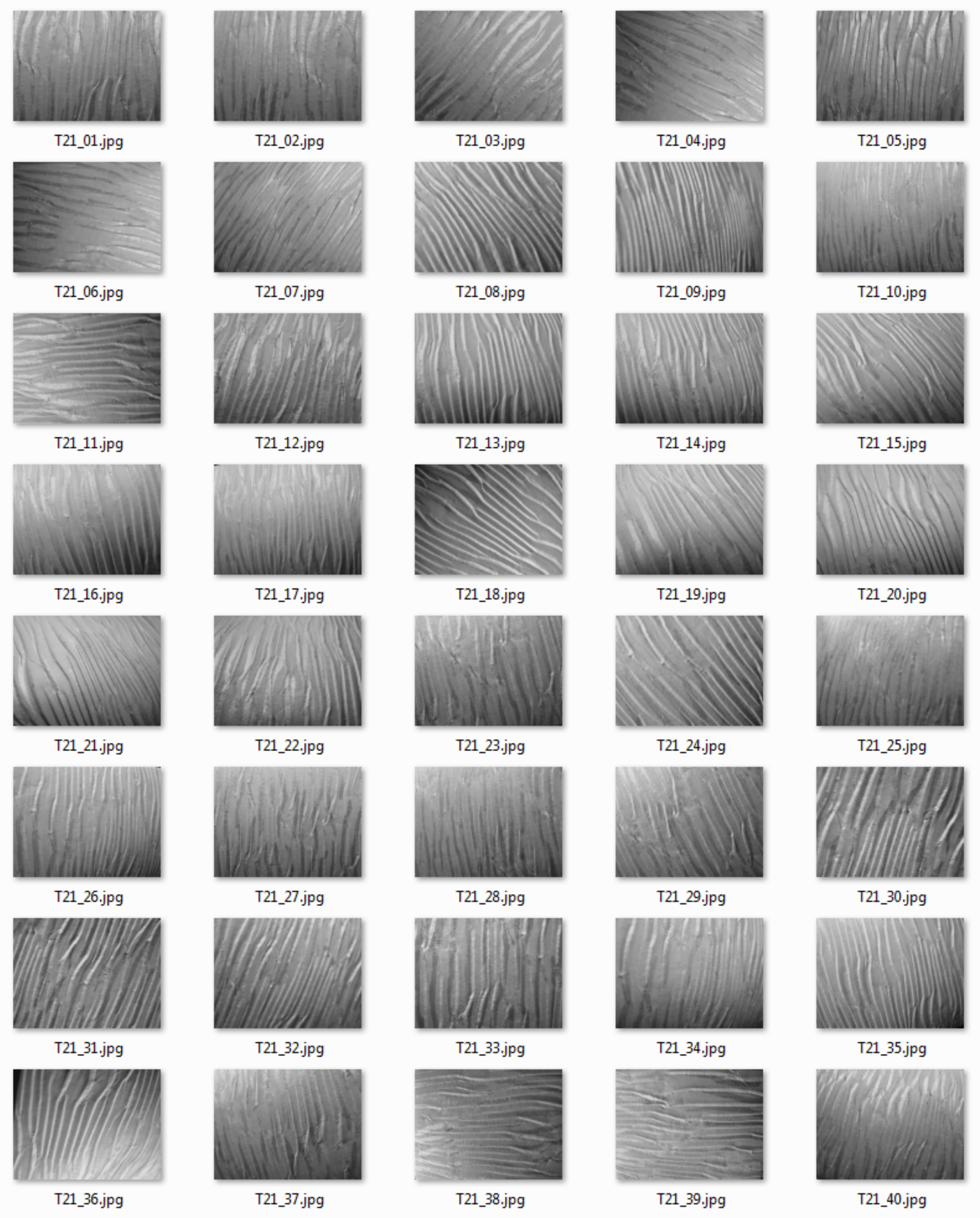


Table B.22: UIUCTex Dataset texture class 22 (fur) samples.

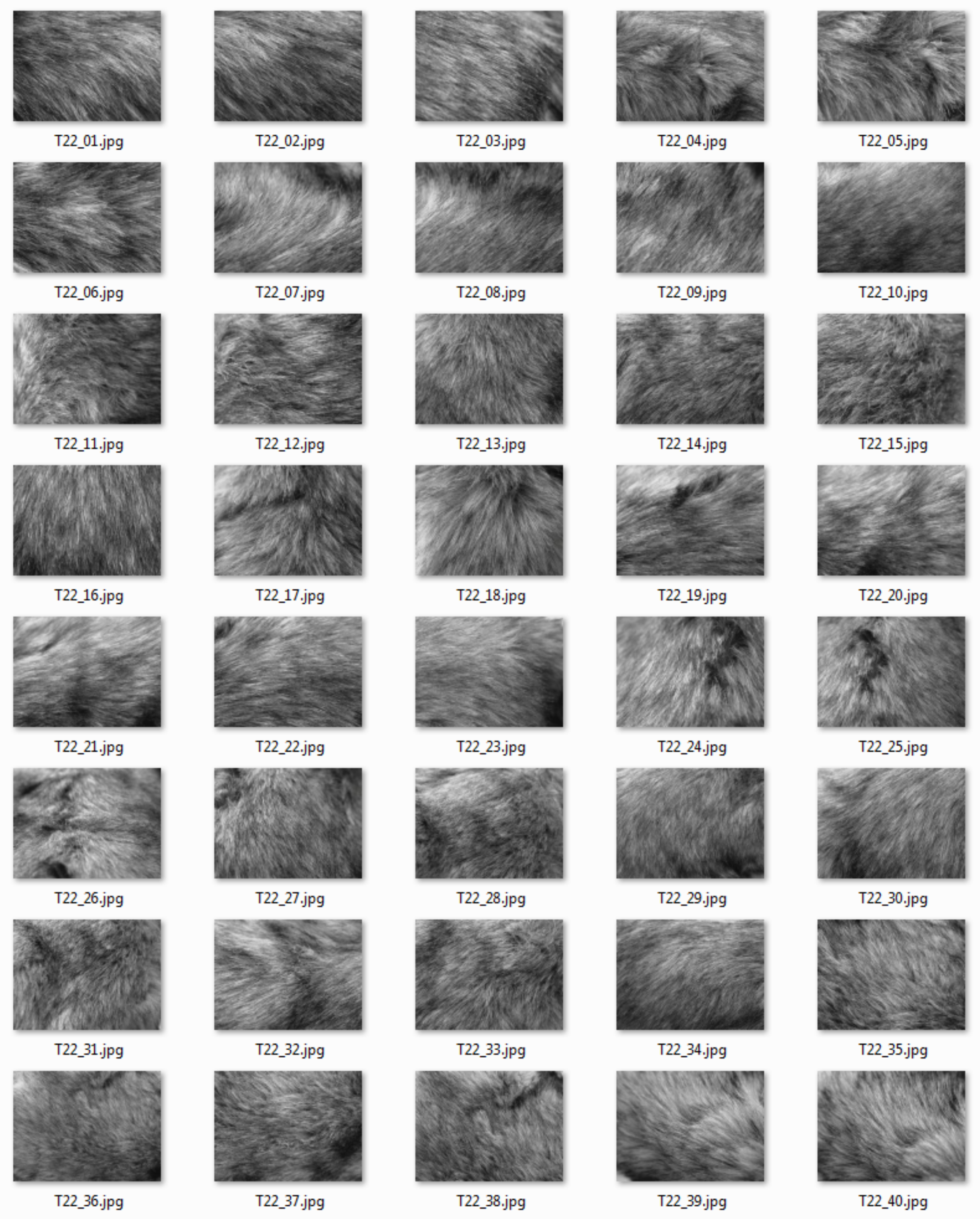


Table B.23: UIUCTex Dataset texture class 23 (knit) samples.

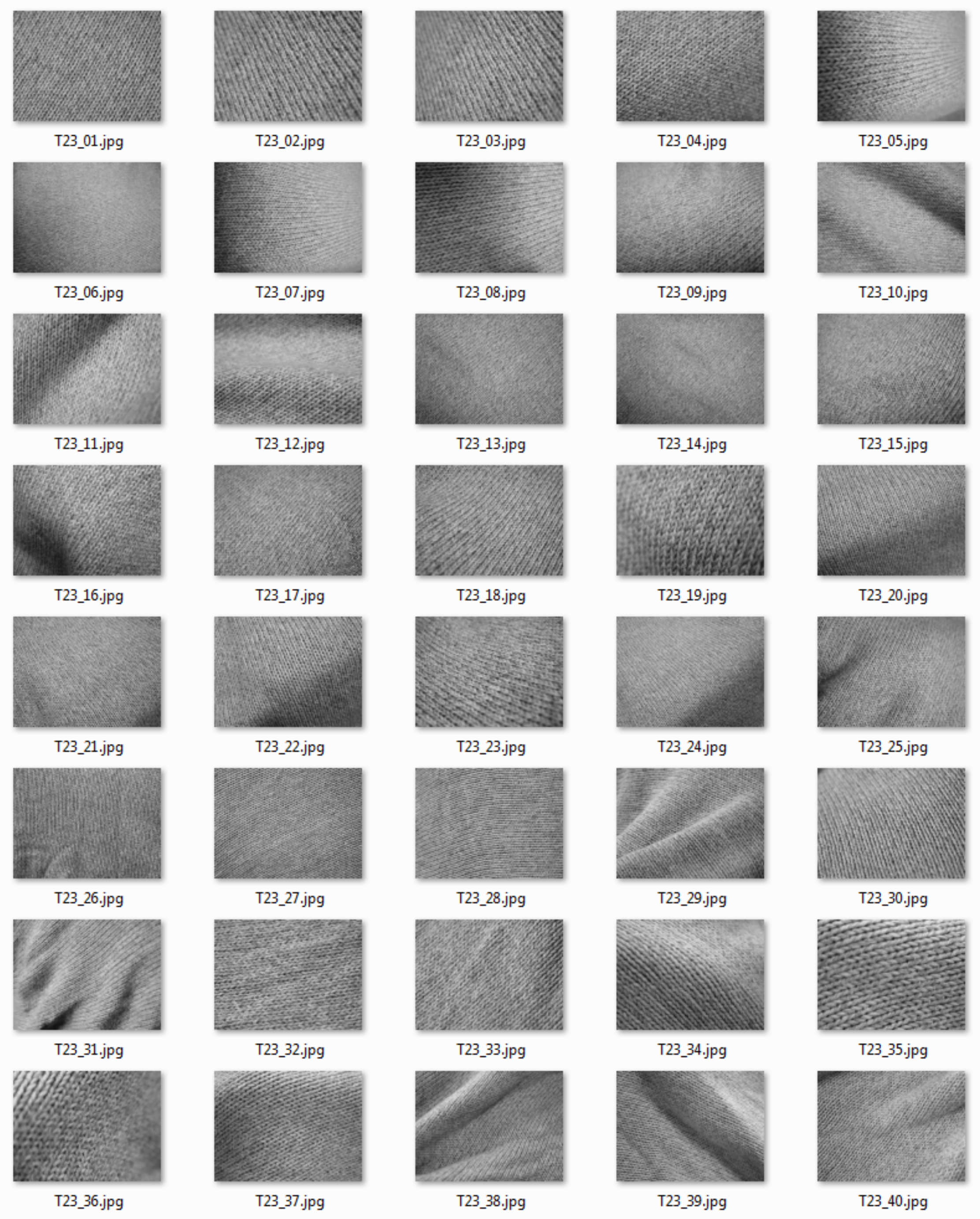


Table B.24: UIUCTex Dataset texture class 24 (corduroy) samples.

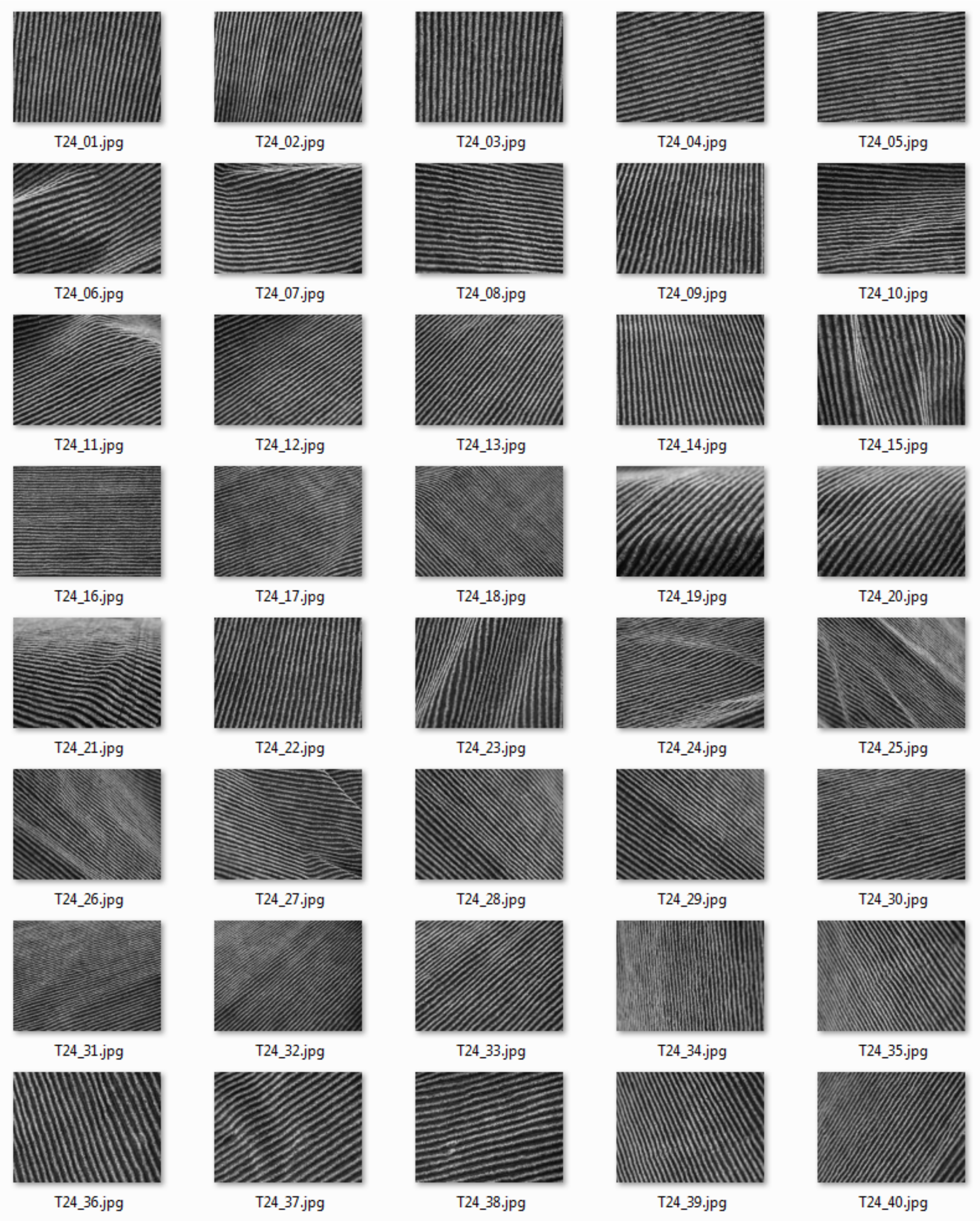
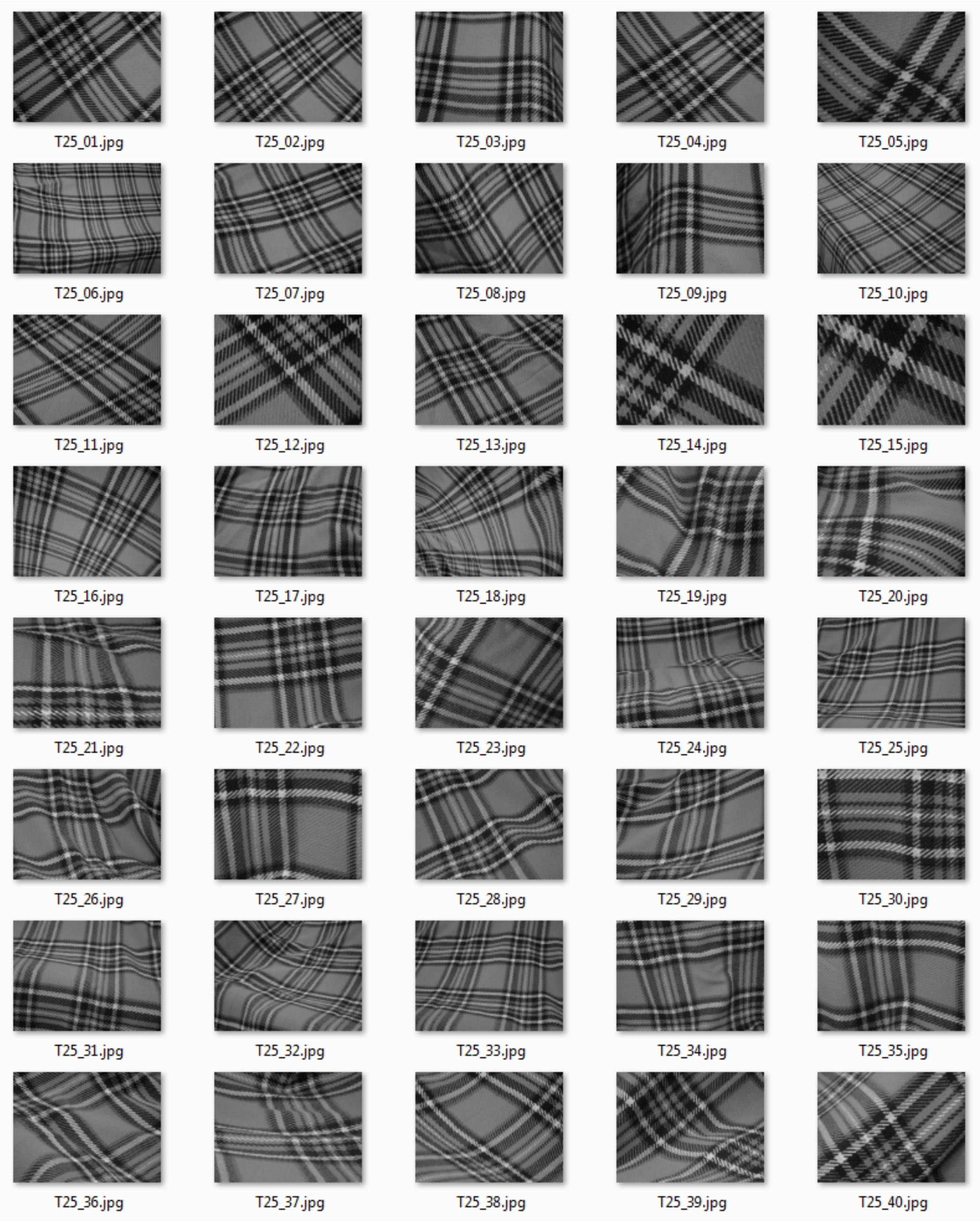


Table B.25: UIUCTex Dataset texture class 25 (plaid) samples.



Appendix C

Human Subject Practice Texture Dataset

Table C.1: Example textures used in the human texture classification data collection instruction pages.