DEFENSE TECHNICAL INFORMATION CENTER

Information for the Defense Community

DTIC® has determined on 1/10/2009 that this Technical Document has the Distribution Statement checked below. The current distribution for this document can be found in the DTIC® Technical Report Database.

☒ **DISTRIBUTION STATEMENT A.** Approved for public release; distribution is unlimited. per DARPA

☐ **© COPYRIGHTED**; U.S. Government or Federal Rights License. All other rights and uses except those permitted by copyright law are reserved by the copyright owner.

☐ **DISTRIBUTION STATEMENT B.** Distribution authorized to U.S. Government agencies only (fill in reason) (date of determination). Other requests for this document shall be referred to (insert controlling DoD office)

☐ **DISTRIBUTION STATEMENT C.** Distribution authorized to U.S. Government Agencies and their contractors (fill in reason) (date of determination). Other requests for this document shall be referred to (insert controlling DoD office)

☐ **DISTRIBUTION STATEMENT D.** Distribution authorized to the Department of Defense and U.S. DoD contractors only (fill in reason) (date of determination). Other requests shall be referred to (insert controlling DoD office).

☐ **DISTRIBUTION STATEMENT E.** Distribution authorized to DoD Components only (fill in reason) (date of determination). Other requests shall be referred to (insert controlling DoD office).

☐ **DISTRIBUTION STATEMENT F.** Further dissemination only as directed by (inserting controlling DoD office) (date of determination) or higher DoD authority.

Distribution Statement F is also used when a document does not contain a distribution statement and no distribution statement can be determined.

☐ **DISTRIBUTION STATEMENT X.** Distribution authorized to U.S. Government Agencies and private individuals or enterprises eligible to obtain export-controlled technical data in accordance with DoDD 5230.25; (date of determination). DoD Controlling Office is (insert controlling DoD office).

Final Report

Sponsor: Defense Advanced Research Projects Agency
Sponsor ID: HR0011-04-1-0037
Account: 6896353
Expiration Date: 8/31/2009
Title: Learning On-line From a Few Examples

Learning on-line from a few examples

Dr. Tomaso Poggio
Center for Biological & Computational Learning
McGovern Institute
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

Learning from the Brain with Applications to Scene Understanding and to Speech Synthesis and Understanding

Executive Summary

The problem of learning represents a gateway to understanding intelligence in brains and machines and to making intelligent machines that learn from experience. Since our brain represents the best known example of a learning machine, we have developed algorithm directly based on the architecture of the cortex. In the process we have demonstrated systems in applications of interest to IPTO. In particular, we have demonstrated

- a system for *scene understanding in street environments*
- *videorealistic synthesis of a speaking agent.*

In the more recent phases of the project, we have explored architectures for learning and for visual recognition process that have a deeply hierarchical organization and incorporate attentional feedbacks and top-down priors. We have also extended our demonstration to video and the recognition of actions. In particular, we have developed

- A system, derived form our model of visual cortex, for the *automatic, quantitative phenotyping* of mouse behavior *from videos.*
- A mathematical framework – a *theory of hierarchical kernel machines* -- to characterize the properties and limitations of deep learning networks with a hierarchical architecture similar to the cortex.

A Brief History of the Projects

We started the project with the plan of a) extending supervised learning algorithms to the fundamental ability to learn from just a few examples, and (b) to apply existing learning techniques and their extensions to an important application domain for the DoD, eg scene understanding tasks in street images, using a database – that we created -- of images of streets in Cambridge, with a variety of objects, from buildings to stores, people, traffic lights, cars, trucks, buses. The system we proposed had the goal of describing a street scene by identifying the key objects in it and ultimately, by understanding video.

In addition, we proposed to develop learning techniques for multimodal human-computer interfaces (computer graphics and speech synthesis) and in particular *videorealistic synthesis of a speaking agent.* From short video segments, we had developed a technique for learning how a person speaks, and then generated a synthetic video of the person's face speaking an arbitrary segment of somebody else's speech, even in another language. We had planned to extend the system to synthesize 3D videos and to *synthesize the voice of a person* by learning from a very small speech corpus.

We have worked on extending supervised learning algorithms for the last four years with a number of achievements on the theoretical side. The most recent one concern the development of a mathematical framework for hierarchical learning systems, inspired by the architecture of the visual cortex.

We achieved most of our goals on the multimodal human-computer interfaces project at the end of the first two years (see Appendices).

Our main project – of scene and video understanding using a small number of training images – reached most of the planned results after the first three years. Afterwards, it successfully explored new architectures for learning and recognition directly derived from cortical architectures.

Background and Goals

Developing systems that can truly learn from experience, mostly by themselves, in an incremental way, would ultimately be relevant for many DARPA projects. Our applications -- *scene understanding and monitoring in street environments* and *videorealistic synthesis of a speaking agent* -- should have a direct impact on the technology of surveillance in general and on electronic disinformation techniques.

In particular, we worked on *Scene understanding and monitoring tasks in street environments.* We collected a database of images of streets in Cambridge, with a variety of objects, from buildings to stores, people, traffic lights, cars, trucks, buses. The system we proposed should describe a street scene by identifying the key objects in it. Ultimately, such a system may understand video and report anomalous events. The system must be able to learn from a relatively just a few example images of each object.

We also planned to work on multimodal human-computer interfaces (computer graphics and speech synthesis) and in particular *videorealistic synthesis of a speaking agent.* From short video segments, we had already developed a technique for learning how a person speaks, and then generated a synthetic video of the persons face speaking an arbitrary segment of somebody else speech, even in another language. We planned to extend the system to synthesize 3D videos and especially the voice of a person by learning from a small speech corpus.

Main Accomplishments Over the Course of the Project

- Following a model of the ventral stream of visual cortex, we developed a novel set of features for visual recognition. These features outperformed state of the art features on several datasets used in computer vision benchmarks.

- A system for object recognition in street scenes was built on top of these features. The system can reliably identify several different object types such as cars, bikes, pedestrians, sky, road, buildings and trees. While the performance is not fully satisfactory, the system outperforms state of the art systems that we implemented for comparison.

- A novel learning algorithm was developed for learning from few examples. The algorithm is a variant of gentleBoost and was especially designed to avoid overfitting on small datasets. The algorithm selects relevant features and provides a considerable speed-up for our object recognition system. It was shown to outperform existing boosting algorithms not only on a variety of vision datasets but also on various genomic datasets, where learning from few examples is also a concern.

- We completed the development of hierarchical feedforward architecture for object recognition based on the anatomy and the physiology of the visual cortex, and showed that the resulting performance on several databases of complex images is as good as or better than the best available computer vision systems (*2007 PAMI paper*).

- We have developed the notion of "audio flow", which is inspired by the notion of "optical flow" from computer vision, and which models the shifting which occurs in the formants during speech. Audio flow defines the correspondence from one vocal tract filter to another.

- We have successfully created a morphable model of the vocal tract filter space, in which any filter is viewed as a "morphed" combination of prototype filters extracted from a small 20 second corpus. In the morphable model we define audio flow between 60-80 prototype filters. Any novel vocal tract filter is modeled as a morph between those 60-80 prototype filters

- We have also successfully created a vector space for the excitation signal, in which any pitch period in the excitation signal is viewed as a linear combination of prototype pitch periods extracted from a small 10 sec corpus.

- We developed the feedforward path of a new architecture for object recognition based on the anatomy and the physiology of the visual cortex.

- We showed that the resulting performance on complex imagery outperforms state-of-the-art vision systems.

- We also showed for the first time that a neurobiological model of cortex does as well as humans and better than state-of-the-art computer vision systems on a challenging, natural image recognition task (*2007 PNAS paper*)..

- We have collected a database of images of streets in Cambridge, with a variety of objects, from buildings to stores, people, traffic lights, cars, trucks, buses and completed a first version of a system capable of scene understanding in such a domain (street images). See http://cbcl.mit.edu/software-datasets/streetscenes/

- We have obtained preliminary results of speech synthesis from very short training sequences. Separately, we improved a system for learning how a person speaks, and for then generating a synthetic video of the persons face speaking an arbitrary segment of somebody else speech, even in another language.

- We investigated the use of morphable models for audio synthesis, which enabled us to synthesize a voice from a very small audio corpus (< 1 minute). A crucial component in achieving this goal was to develop a representation of speech that is smooth and which can accurately reconstruct speech. If the representation is smooth, it may then be placed in a morphable model framework, and we can then morph segments of speech to produce new realistic utterances from very small amounts of data.

- We have made significant progress in this regard; in particular, we have developed a novel representation called Max-Gabor analysis, which produces a smooth representation of speech. This analysis is inspired mainly by the work of Shamma and colleagues, who have developed a two-stage auditory model based on psycho-acoustical and neurophysiological findings in the early and central stages of the auditory pathway. Also this representation borrows from the work of Riesenhuber and Poggio, who developed a model of object recognition in visual cortex by embedding a MAX operator in a hierarchical neural model.

- Max-Gabor analysis works by analyzing small spectro-temporal patches P of a two-dimensional magnitude spectrogram S(f,t), and representing each patch by its **locally dominant** spectro-temporal periods T(f,t) and orientations Theta(f,t). The method also estimates local patch amplitudes A(f,t) and phases Phi(f,t) as well. Since the local patches

are Gabor-like, Max-Gabor analysis operates by performing a two-dimensional Gabor-like analysis of the spectrogram, retaining only the parameters of the 2D-Gabor filter with **maximal amplitude** response within the local region. Hence we call our technique a Max-Gabor analysis of spectrograms. Given the estimated local periods T(i,j), orientations Theta(i,j),amplitudes A(i,j), and phases Phi(i,j), the spectrogram S(f,t) can be reconstructed by synthesizing individual local 2D Gabors Gij(f,t) for each patch, and overlap-adding them together.

- We analyzed and re-synthesized several test utterances of different speakers uttering the phrase "Hi Jane". Our results may be seen on the web at http://cuneus.ai.mit.edu:8000/research/ maxgabor. In general, we have found that the Max-Gabor parameters are smooth, meaningful, and capable of reconstructing the original spectrogram. Our future goal is to see if it is possible to morph the Max-Gabor parameters to generate novel speech.

- We have extended the model of the ventral stream to incorporate neuroscience data on backprojections and control of attention and eye movements in collaboration with Bob Desimone (McGovern Institute and BCS). Preliminary results show that this extended model can predict human eye movements in top-down tasks better than other standard models of saliency.

- We are developing – with neuroscience details -- an extension of the model to the dorsal stream for the recognition of actions.

- We have used the system above to phenotype mice behavior – developing a vision system that could be developed into a useful tool for biologists. We have a prototype system that we will test in several labs at MIT and the Broad Institute working with mutant mice as models of mental and neurological diseases.

Main Publications Related to the Project

Bouvrie, J., L. Rosasco, G. Shakhnarovich, and S. Smale, "Notes on the Shannon Entropy of the Neural Response", *CBCL-281, MIT-CSAIL-TR-2009-049*, Massachusetts Institute of Technology, Cambridge, MA, October 9, 2009

Chikkerur, S., T. Serre, and T. Poggio, "A Bayesian inference theory of attention: neuroscience and algorithms" *MIT-CSAIL-TR-2009-047/CBCL-280*, Massachusetts Institute of Technology, Cambridge, MA, October 3, 2009.

Chikkerur, S., T. Serre, and T. Poggio, "Attentive processing improves object recognition" *MIT-CSAIL-TR-2009-046 /CBCL-279*, Massachusetts Institute of Technology, Cambridge, MA, October 2, 2009.

Chikkerur, S., C. Tan, T. Serre, and T. Poggio, "An integrated model of visual attention using shape-based features" *MIT-CSAIL-TR-2009-029 / CBCL-278*, Massachusetts Institute of Technology, Cambridge, MA, June 20, 2009.

De Mol, C., E. De Vito and L. Rosasco. "Elastic-Net Regularization in Learning Theory", to be published in the *Journal of Complexity*, available online January 30, 2009.

Lo Gerfo L., Rosasco L., Odone F., De Vito E. and Verri, A. "Spectral Algorithms for Supervised Learning", *Neural Computation*. 2008 20: 1873-1897.

Rosasco, L., S. Mosci, M. Santoro, A. Verri, and S. Villa, "Iterative Projection Methods for Structured Sparsity Regularization", *MIT-CSAIL-TR-2009-50 / CBCL-282*, Massachusetts Institute of Technology, Cambridge, MA, October 14, 2009.

Smale, S., L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio, "Mathematics of the Neural Response", *Foundations of Computational Mathematics*, June 2009 (online)

Terashima, Y. "Scene Classification with a Biologically Inspired Method", *CBCL paper #277/CSAIL Technical Report#2009-020, CBCL-277* Massachusetts Institute of Technology, Cambridge, MA, May 10, 2009.

Barla, A., Mosci, S., Rosasco, L. and Verri, A. "A method for robust variable selection with significance assessments" *16th European Symposium on Artificial Neural Networks.*

Bileschi, S.M. Object Detection at Multiple Scales Improves Accuracy, *ICPR 2008.*

Bileschi, S.M. A Multi-Scale Generalization of the HoG and HMAX Image Descriptors for Object Detection, *CBCL paper #271/CSAIL Technical Report #2008-019,* Massachusetts Institute of Technology, Cambridge, MA, April 9, 2008.

Bouvrie, J., T. Ezzat, and T. Poggio. "Localized Spectro-Temporal Cepstral Analysis of Speech", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, 2008.

Caponnetto, A., T. Poggio and S. Smale On a model of visual cortex: learning invariance and selectivity from image sequences, *CBCL paper#272/CSAIL Technical Report #2008-030* , Massachusetts Institute of Technology, Cambridge, MA, April 4, 2008

De Mol, C., E. De Vito and L. Rosasco. "Elastic-Net Regularization in Learning Theory", *CBCL paper #273/ CSAIL Technical Report #TR-2008-046* , Massachusetts Institute of Technology, Cambridge, MA, July 24, 2008.

De Vito, E., S. Pereverzyev, and L. Rosasco. "Adaptive Kernel Methods Using the Balancing Principle", *CBCL paper #275/CSAIL Technical Report#TR-2008-062*Massachusetts Institute of Technology, Cambridge, MA, October 16, 2008.

Ezzat, T. and T. Poggio. "Discriminative Word-Spotting Using Ordered Spectro-Temporal Patch Features", SAPA Workshop, Interspeech, Brisbane, Australia, 2008

Geiger, G., C. Cattaneo, R. Galli, U. Pozzoli, M. Lorusso, A. Facoetti, and M. Molteni. "Wide and diffuse perceptual modes characterize dyslexics in vision and audition", Perception Vol. 37, Issue 11, Pages 1745 – 1764

Kouh, M. and T. Poggio. "A Canonical Neural Circuit for Cortical Nonlinear Operations" *Neural Computation*, June 2008, Vol. 20, No. 6, Pages 1427-1451

LeCun, Y., D.G. Lowe, J. Malik, J. Mutch, P. Perona, and T. Poggio. Object Recognition, Computer Vision, and the Caltech 101: A Response to Pinto et al., *PLoS Computational Biology,* Posted Online March 2008.

Lo Gerfo L., Rosasco L., Odone F., De Vito E. and Verri, A. "Spectral Algorithms for Supervised Learning", *Neural Computation*. 2008 20: 1873-1897.

Meyers, E., and L. Wolf. Using Biologically Inspired Visual Features for Face Processing. *International Journal of Computer Vision*, Vol 76, No. 1, 93-104, 2008

Meyers, E.M., D. J. Freedman, G. Kreiman, E.K. Miller, and T. Poggio. "Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex". *Journal of Neurophysiology* Vol. 100: 1407-1419, June 18, 2008.

Mosci, S.; A. Barla; A. Verri and L. Rosasco. " Finding Structured Gene Signatures". *Proc. of IEEE BIBM*, 2008, Philadelphia, PA. USA.

Rosasco, L., M. Belkin, and E. De Vito. "A Note on Perturbation Results for Learning Empirical Operators ", *CBCL paper #274/ CSAIL Technical Report #TR-2008-052* , Massachusetts Institute of Technology, Cambridge, MA, August 19, 2008.

S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. "Mathematics of the Neural Response", *CBCL Paper #276/MIT CSAIL Technical Report #TR2008-070,* Massachusetts Institute of Technology, Cambridge, MA, November, 2008

Bileschi, S. Wolf, L. "Image representations beyond histograms of gradients: The role of Gestalt descriptors", *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, 2007 - CVPR '07*, June 2007, Pages 1-8, Digital Object Identifier 10.1109/CVPR.2007.383122

Cadieu, C., M. Kouh, A. Pasupathy, C. Connor, M. Riesenhuber, and T. Poggio. A Model of V4 Shape Selectivity and Invariance, *Journal of Neurophysiology,* Vol. 98, 1733-1750, June, 2007.

Ezzat, T., J. Bouvrie, and T. Poggio. Spectro-Temporal Analysis of Speech Using 2-D Gabor Filters, *Interspeech*, Belgium 2007.

Ezzat, T., J. Bouvrie, and T. Poggio. AM-FM Demodulation of Spectrograms using 2-D Max-Gabor Analysis, *ICASSP*, Hawaii, 2007.

Heisele, B., T. Serre and T. Poggio. A Component-based Framework for Face Detection and Identification, *International Journal of Computer Vision*, 74(2), pp. 167-181, 2007.

Jhuang H., T. Serre, L. Wolf and T. Poggio. A Biologically Inspired System for Action Recognition, In: *Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV)*, 2007.

Serre, T., A. Oliva and T. Poggio. A Feedforward Architecture Accounts for Rapid Categorization, *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 104, No. 15, 6424-6429, 2007.

Serre, T., L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio. Object Recognition with Cortex-like Mechanisms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 3, 411-426, 2007.

Rifkin, R., J. Bouvrie, K. Schutte, S. Chikkerur, M. Kouh, T. Ezzat and T. Poggio. Phonetic Classification Using Hierarchical, Feed-forward, Spectro-temporal Patch-based Architectures, *CBCL Paper #266/AI Technical Report #2007-019*, Massachusetts Institute of Technology, Cambridge, MA, March, 2007.

Masquelier, T., T. Serre, S. Thorpe and T. Poggio. Learning complex cell invariance from natural videos: A plausibility proof, *CBCL Paper #269/AI Technical Report #2007-060,* Massachusetts Institute of Technology, Cambridge, MA, December, 2007.

Poggio, T. "Neuroscience: New Insights for AI?". In: *Web Intelligence Meets Brain Informatics*, First WICI International Workshop, WImBI 2006,Beijing, China, December 2006.

Poggio. T. How the Brain Might Work: The Role of Information and Learning in Understanding and Replicating Intelligence. In: *Information: Science and Technology for the New Century*, Editors: G. Jacovitt, A. Pettorossi, R. Consolo and V. Senni, Lateran University Press, Quaderni Sefir, 7, pp. 45-61, 2007.

Rifkin, R., K. Schutte, D. Saad, J. Bouvrie, and J. Glass. Noise Robust Phonetic Classification with Linear Regularized Least Squares and Second-Order Features, *ICASSP*, Honolulu, 2007.

Rifkin, R.,. and R.A. Lippert. Notes on Regularized Least-Squares, *CBCL Paper #268/AI Technical Report #2007-019*, Massachusetts Institute of Technology, Cambridge, MA, May, 2007.

Rifkin, R., J. Bouvrie, K. Schutte, S. Chikkerur, M. Kouh, T. Ezzat and T. Poggio. Phonetic Classification Using Hierarchical, Feed-forward, Spectro-temporal Patch-based Architectures, *CBCL Paper #267/AI Technical Report #2007-019*, Massachusetts Institute of Technology, Cambridge, MA, March, 2007.

Serre, T., A. Oliva and T. Poggio. A Feedforward Architecture Accounts for Rapid Categorization, *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 104, No. 15, 6424-6429, 2007.

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

Serre, T., L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio. Object Recognition with Cortex-like Mechanisms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 3, 411-426, 2007.

Smale, S., T. Poggio, A. Caponnetto, and J. Bouvrie. Derived Distance: towards a mathematical theory of visual cortex, *CBCL Paper*, Massachusetts Institute of Technology, Cambridge, MA, November, 2007.

Wolf, L., H.Jhuang and T.Hazan. Modeling Appearances with Low-Rank SVM, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

Zoccolan, D., M. Kouh, T. Poggio, and James J. DiCarlo . Trade-off between object selectivity and tolerance in monkey inferotemporal cortex, *Journal of Neuroscience,* Volume 27(45), pp.12292–12307, November, 2007.

Poggio. T. How the Brain Might Work: The Role of Information and Learning in Understanding and Replicating Intelligence. In: *Information: Science and Technology for the New Century*, Editors: G. Jacovitt, A. Pettorossi, R. Consolo and V. Senni, Lateran University Press, Quaderni Sefir, 7, pp. 45-61, 2007.

Bileschi, Stanley (Ph.D. Thesis, EECS, MIT, 2006): StreetScenes: Towards Scene Understanding in Still Images.

Bileschi, S. and L. Wolf. A Unified System for Object Detection, Texture Recognition and Context Analysis Based on the Standard Model Feature Set. In: British Machine Vision Conference (BMVC), 2006.

Bouvrie, J. and T. Ezzat. An Incremental Algorithm for Signal Reconstruction from Short-time Fourier Transform Magnitude. In: Ninth International Conference on Spoken Language Processing (ICSLP-Pittsburgh, PA), 2006.

Bouvrie, J. and T. Ezzat. An Incremental Algorithm for Signal Reconstruction from Short-time Fourier Transform Magnitude. In: Ninth International Conference on Spoken Language Processing (ICSLP-Pittsburgh, PA), 2006.

Caponnetto, A. and A. Rakhlin. Stability Properties of Empirical Risk Minimization over Donsker Classes, *Journal of Machine Learning Research*, Vol. 7, 2565-2583, 2006.

Caponnetto, A. and Y. Yao. Adaptation for Regularization Operators in Learning Theory, *CBCL Paper #265/AI Technical Report #063*, Massachusetts Institute of Technology, Cambridge, MA, September, 2006.

Caponnetto, A. Optimal Rates for Regularization Operators in Learning Theory, *CBCL Paper #264/AI Technical Report #062*, Massachusetts Institute of Technology, Cambridge, MA, September, 2006.

Chikkerur, S. and L. Wolf. Empirical Comparison between Hierarchical Fragments Based and Standard Model Based Object Recognition Systems, *CBCL Paper #MMVI-01*, Massachusetts Institute of Technology, Cambridge, MA, March, 2006.

Ezzat, T., J. Bouvrie and T. Poggio. Max-Gabor Analysis and Synthesis of Spectrograms. In: Ninth International Conference on Spoken Language Processing (ICSLP-Pittsburgh, PA), 2006.

Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. Experience dependent sharpening of visual shape selectivity in inferior temporal cortex, *Cerebral Cortex,* 16: 1631-1644, 2006.

Kreiman, G., C.P. Hung, A. Kraskov, R.Q. Quiroga, T. Poggio and J.J. DiCarlo. Object Selectivity of Local Field Potentials and Spikes in the Macaque Inferior Temporal Cortex, *Neuron*, Vol. 49, 433-445, 2006.

Mukherjee, S., P. Niyogi, T. Poggio and R. Rifkin. Learning Theory: Stability is Sufficient for Generalization and Necessary and Sufficient for Consistency of Empirical Risk Minimization, *Advances in Computational Mathematics*, 25, 161-193, 2006.

Poggio, T. Neuroscience: New Insights for AI? , In: *Web Intelligence Meets Brain Informatics*, First WICI International Workshop (WImBI 2006) Bejing, China, December 2006.

Rakhlin, Alexander (Ph.D. Thesis, BCS, MIT, April 2006): Applications of Empirical Processes in Learning Theory: Algorithmic Stability and Generalization Bounds.

Rakhlin, A. and A. Caponnetto. Stability of K-means Clustering. In: Neural Information Processing Systems Conference, 2006.

Serre, Thomas R. (Ph.D. Thesis, BCS, MIT, March 2006): Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines.

Serre, T. Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines, *CBCL Paper #260/AI Technical Report #028*, Massachusetts Institute of Technology, Cambridge, MA, March, 2006.

Wolf, L. and S. Bileschi. A Critical View of Context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

Wolf, L., S. Bileschi and E. Meyers. Perception Strategies in Hierarchical Vision Systems. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

Wolf F., T. Poggio, P. Sinah Human Document Classification Using Bags of Words , *CBCL paper #263/CSAIL Technical Report #2006-054,* Massachusetts Institute of Technology, Cambridge, MA, August 9, 2006.

Yokono, J.J. and T. Poggio. A Multiview Face Identification Model With No Geometric Constraints, *Sony Intelligence Dynamics Laboratories, Inc.* March, 2006.

Hung, C.P., G. Kreiman, T. Poggio and J.J. DiCarlo. Fast Readout of Object Identity from Macaque Inferior Temporal Cortex, *Science*, Vol. 310, 863-866, 2005.

Serre, T., M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman and T. Poggio. A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex, *CBCL Paper #259/AI Memo #2005-036*, Massachusetts Institute of Technology, Cambridge, MA, October, 2005.

Serre, T., L. Wolf and T. Poggio. Object Recognition with Features Inspired by Visual Cortex. In: Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Press, San Diego, June 2005.

Sigala, R., T. Serre, T. Poggio and M. Giese. Learning Features of Intermediate Complexity for the Recognition of Biological Motion. In: *ICANN 2005*, Warsaw, Poland, 241-246, September 2005.

Skelley, James P. (S.M. Thesis, EECS, MIT, August 2005): Experiments in Expression Recognition.

Wolf, L. and S. Bileschi. Combining Variable Selection with Dimensionality Reduction, *CBCL Paper #247/AI Memo #2005-009*, Massachusetts Institute of Technology, Cambridge, MA, March 2005.

Wolf, L. and I. Martin. Robust Boosting for Learning from Few Examples. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

Wolf, L. and A. Shashua. Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-based Approach, *Journal of Machine Learning Research*, 6, 1855-1887, 2005.

Wu, Jia Jane (S.M. Thesis, EECS, MIT, May 2005): Comparing Visual Features for Morphing Based Recognition.

Yokono, J.J. and T. Poggio. Boosting a Biologically Inspired Local Descriptor for Geometry-free Face and Full Multi-view 3D Object Recognition, *CBCL Paper #254/AI Memo #2005-023*, Massachusetts Institute of Technology, Cambridge, MA, July 2005.

Yokoyama, M. and T. Poggio. A Contour-Based Moving Object Detection and Tracking. In: Proceedings of Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (in conjunction with ICCV 2005), Beijing, China, October 15-16, 2005.

Zoccolan, D., D.D. Cox and J.J. DiCarlo. Multiple Object Response Normalization in Monkey Infero-temporal Cortex, *Journal of Neuroscience*, 25(36), 8150-64, 2005.

Bouvrie, Jacob V. (S.M. Thesis, EECS, MIT, June 2004): Multi-Source Contingency Clustering.

Cadieu, C., M. Kouh, M. Riesenhuber and T. Poggio. Shape Representation in V4: Investigating Position-specific Tuning for Foundary Conformation with the Standard Model of Object Recognition, *CBCL Paper #241/AI Memo #2004-024*, Massachusetts Institute of Technology, Cambridge, MA, November, 2004.

Ezzat, T., G. Geiger and T. Poggio. Trainable Videoreaslistic Speech Animation. In: *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (FGR2004, Seoul, Korea), 57-64, 2004.

Fischer, Robert (S.M. Thesis, Math & Natural Sciences, MIT/Univ. of Applied Science Darmstadt, Germany, October 2004): Automatic Facial Expression Analysis and Emotional Classification.

Heisele, B. and T. Koshizen. Components for Face Recognition. In: *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, 153-158, 2004.

Ivanov, I., B. Heisele and T. Serre. Using Component Features for Face Recognition. In: *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, 421-426, 2004.

Kouh, M. and T. Poggio. A General Mechanisms for Tuning: Gain Control Circuits and Synapses Underlie Tuning of Cortical Neurons, *CBCL Paper #245/AI Memo #2004-031*, Massachusetts Institute of Technology, Cambridge, MA, December 2004.

Kreiman, G. Neural Coding: Computational and Biophysical Perspectives, *Physics of Life Reviews*, 2, 71-102, 2004.

Kreiman, G., C. Hung, T. Poggio and J. DiCarlo. Selectivity of Local Field Potentials in Macaque Inferior Temporal Cortex, *CBCL Paper #240/AI Memo #2004-020*, Massachusetts Institute of Technology, Cambridge, MA, September, 2004.

Leung, Brian (S.M. Thesis, EECS, MIT, May 2004): Component-based Car Detection in Street Scene Images.

Lorusso, M.L., A. Facoetti, S. Pesenti, C. Cattaneo, M. Molteni and G. Geiger. Wider Recognition in Peripheral Vision Common to Different Subtypes of Dyslexia, *Vision Research*, 44, 2413-2424, 2004.

Paysan, Pascal (S.M. Thesis, Computer Science, Fachochschule Esslingen, February 2004): Stereovision-based Vehicle Classification Using Support Vector Machines.

Poggio, T. and E. Bizzi. Generalization in Vision and Motor Control, *Nature*, Vol. 431, 768-774, 2004.

Poggio, T., R. Rifkin, S. Mukherjee and P. Niyogi. General Conditions for Predictivity in Learning Theory, *Nature*, Vol. 428, 419-422, 2004.

Rakhlin, A., S. Mukherjee, and T. Poggio. On Stability and Concentration of Measure, CBCL Paper, Massachusetts Institute of Technology, Cambridge, MA, June 2004.

Rakhlin, A., D. Panchencko and S. Mukherjee. Risk Bounds for Mixture Density Estimation, *CBCL Paper #233/AI Memo #2004-001*, Massachusetts Institute of Technology, Cambridge, MA, January, 2004.

Riesenhuber, M., I. Jarudi, S. Gilad and P. Sinha. Face Processing in Humans is Compatible with A Simple Shape-based Model of Vision, *Proc. R. Soc. Lond. B (Suppl.)*, DOI 10.1098/rsbl.2004.0216, 04BL0061.S1-04BL0061.S3, 2004.

Riesenhuber, M., I. Jarudi, S. Gilad and P. Sinha Face Processing in Humans is Compatible with a Simple-Shape-based Model of Vision, *CBCL Paper #236/AI Memo #2004-006*, Massachusetts Institute of Technology, Cambridge, MA, March, 2004.

Rifkin, R. and A. Klautau. In Defense of One-vs-All Classification, *Journal of Machine Learning Research*, Vol. 5, 101-141, 2004.

Schneider, R. and M. Riesenhuber. On the Difficulty of Feature-based Attentional Modulations in Visual Object Recognition: A Modeling Study, *CBCL Paper #235/AI Memo #2004-004*, Massachusetts Institute of Technology, Cambridge, MA, February, 2004.

Serre, T. and M. Riesenhuber. Realistic Modeling of Simple and Complex Cell Tuning in the HMAX Model, and Implications for Invariant Object Recognition in Cortex, *CBCL Paper #239/AI Memo #2004-017*, Massachusetts Institute of Technology, Cambridge, MA, August, 2004.

Serre, T., L. Wolf and T. Poggio. A New Biologically Motivated Framework for Robust Object Recognition, *CBCL Paper #243/AI Memo #2004-026*, Massachusetts Institute of Technology, Cambridge, MA, November, 2004.

Shimizu, H. and T. Poggio. Direction Estimation of Pedestrian from Multiple Still Images. In: *IEEE Intelligent Vehicles Symposium 2004*, Parma, Italy, June 14-17, 2004.

Weyrauch, B., J. Huang, B. Heisele and V. Blanz. Face Processing in Video. In: *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, IEEE Computer Society Press, Washington, DC, 2004, in press.

Wolf, L. and I. Martin. Regularization through Feature Knock Out, *CBCL Paper #242/AI Memo #2004-025*, Massachusetts Institute of Technology, Cambridge, MA, November, 2004.

Wolf, L., A. Shashua and S. Mukherjee. Selecting Relevant Genes with a Spectral Approach, *CBCL Paper #234/AI Memo #2004-002*, Massachusetts Institute of Technology, Cambridge, MA, January, 2004.

Yeo, Gene W. (Ph.D. Thesis, EECS, MIT, November 2004): Identification, Improved Modeling and Integration of Signals to Predict Constitutive and Alternative Splicing.

Yokono, J.J. and T. Poggio. Oriented Filters for Object Recognition: An Empirical Study. In: *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (FGR2004, Seoul, Korea), 755-760, 2004.

Yokono, J.J. and T. Poggio. Rotation Invariant Object Recognition from One Training Example, *CBCL Paper #238/AI Memo #2004-010*, Massachusetts Institute of Technology, Cambridge, MA, April, 2004.

Yokono, J.J. and T. Poggio. Evaluation of Sets of Oriented and Non-oriented Receptive Fields as Local Descriptors, *CBCL Paper #237/AI Memo #2004-007*, Massachusetts Institute of Technology, Cambridge, MA, March, 2004.

Some of the recent articles on our research

(more information is at http://cbcl.mit.edu/news/index-news.html)

- Tomaso Poggio was awarded the **Okawa Prize**, October, 2009
- **Evolutionary Portrait Art: Tomaso Poggio** by Günter Bachelier: Portraits of the masterminds are transformed...
- **Technology Review** -- By Anne-Marie Corley, A Robot that Navigates Like a Person: A new robot navigates using humanlike visual processing and object detection., Tuesday, June 30, 2009
- **Biomedical Computation Review** -- by Roberta Freidman, PhD, "Reverse Engineering the Brain", Volume 5, Issue 2, pages 10-17, ISSN 1557-3192, Spring 2009
- **LaStampa.it Tecnologia News**, Tomaso Poggio, uno dei padri della neuroscienza e professore al MIT di Boston: "I robot non saranno una minaccia per almeno altri 10 anni. Oggi è Google un potenziale pericolo" (translation: Tomaso Poggio, one of the fathers of the neuroscience and university professor at MIT, Boston: "The robots will not be a threat for at least another 10 years. Today it is Google which is a potential danger"), February 26, 2009
- Masterminds of Artificial Intelligence - Evolutionary Portrait Art by Günter Bachelier, Janaury 2009
- **PC Magazine**: Future Watch: Understanding the Brain, August 2008
- **TERRA ACTUALIDAD - INTERNACIONAL** -- Tecnalia desarrolla en Massachussets la tesis doctoral de una investigadora vasca sobre biologia informatica y neurologia
- **MIT NEWS --** by David Chandler Learning about brains from computers, and vice versa: - Tomaso Poggio.
- **BBC --** This is part of the excellent BBC series entitled "visions of the future". This short clip here shows work performed at CBCL (MIT) about a computational neuroscience model of the ventral stream of the visual cortex. The story here focuses on recent work by Serre, Oliva and Poggio on comparing the performance of the model to human observers during a rapid object categorization task. Visions of the Future - Tomaso Poggio, Thomas Serre and Aude Oliva.
- **SCIENTIFIC AMERICAN --** by Larry Greenemier (February 20, 2008): Visionary Research: Teaching Computers to See Like a Human
- **IEEE Journal: COMPUTING IN SCIENCE & ENGINEERING--** by Pam Frost Gorder (March/April 2008): Computer Vision, Inspired by the Human Brain
- **RAI3.IT - podcast** (December 31, 2007): Interview at RAI3 with Tomaso Poggio - Tomaso Poggio.
- (Pasadena, September 26, 2007): Saliency-based attentional selection with HMAX for object classification: Machine vision demo #2 R. Peters, L. Itti, S. Chikkerur, T. Poggio, J. Harel and C. Koch.
- (Genova, June 14-16, 2007): A Journey through Computation - A. Verri, G. Geiger, F. Girosi, and C. Koch.
- **RAI3.IT - podcast** (June 13, 2007): An Interview by Franco Carlini with Tomaso Poggio at RAI3 - Tomaso Poggio.
- **RAI International** (Salimbeni, May 4, 2007): Tomaso Poggio and His Thinking Machines - Tomaso Poggio.
- **NEWS IN REVIEW - A Look at Today's Ideas and Trends** by Linda Roach, edited by Brian A. Francis (Amer. Acad. of Ophthal. web site; June 2007): When Computer Vision Imitates Life - Thomas Serre and Tomaso Poggio.
- **FORBES** by Robert M. Metcalfe (May 7, 2007): It's All In Your Head; The latest supercomputer is way faster than the human brain. But guess which is smarter?? - Raymond Kurzweil, Tomaso Poggio and Chris Diorio.
- **BRAIN+COGNITIVE SCIENCES NEWS** (Spring 2007): BCS Researchers Find Synergies between Basic Science and Real-World Problem Solving - Pawan Sinha and Tomaso Poggio.
- **IL SOLE 24 ORE - Inchieste: Intelligenza Artificale/Progressi** (April 19, 2007) by Tomaso Poggio: Piu Vicini al Mistero della Visione - Tomaso Poggio.
- **MCGOVERN INSTITUTE NEWS** by Cathryn M. Delude (April 4, 2007): Computer Model Mimics Blink of A Human Eye - Tomaso Poggio, Aude Oliva and Thomas Serre.
- **THE ECONOMIST** - computer vision: a computer can now recognise classes of things as accurately as a person can (April 3, 2007): Easy on the Eyes - Tomaso Poggio and Thomas Serre.

- **NEWSCIENTIST** - NewScientistTech (April 3, 2007): Visual-cortex Simulator Sees Animals as Humans Do - Tomaso Poggio and Thomas Serre.
- **MCGOVERN INSTITUTE NEWS** by Cathryn M. Delude (April, 2007): McGoven Media Coverage of the Poggiolab's 2007 IEEE paper on Street Scene Recognition - Thomas Serre, Stanley Bileschi and Tomaso Poggio.
- **PHYSORG.COM** - Science: Physics: Tech: Nano: Nes (April 2, 2007): First Impressions: Computer Model Behaves Like Humans on Visual Categorization Task - Tomaso Poggio and Thomas Serre.
- **NEUROFUTURE BLOGSPOT** - brain science and the culture of future (March 29, 2007): Computational Vision - Tomaso Poggio and Thomas Serre.
- **NEWSCITECH** - new science and Technology, source for science and technology breakthroughs (February 26, 2007): Computer Model Mimics Neural Processes in Object Recognition - Tomaso Poggio and Thomas Serre.>
- **MIT NEWS OFFICE: TECH TALK** by Cathryn M. Delude (February 27, 2007): Computer Model Mimics Neural Processes in Object Recognition - applications include surveillance, visual search engines, biomedical imaging analysis and robots with realistic vision - Tomaso Poggio, Thomas Serre, Stanley Bileschi, Maximilian Riesenhuber and Lior Wolf.
- **TECHNOLOGY REVIEW** by Duncan Graham-Rowe (February 21, 2007): Biologically Inspired Vision Systems - a computer model of the brain has learned to detect and classify objects - Thomas Serre, Stanley Bileschi and Tomaso Poggio.
- **SLASHDOT** - news for Nerds, Stuff that Matters (February 11, 2007): Recognizing Scenes Like the Brain Does - Tomaso Poggio.
- **MCGOVERN INSTITUTE NEWS** by Cathryn M. Delude (January, 2007): Mimicking How the Brain Recognizes Street Scenes - Tomaso Poggio.
- **LA STAMPA WEB** (September 29, 2006): Capire Come Funziona il Cervello e la Sfida Piu Grande del XXI Secolo - Tomaso Poggio.
- **TECHNOLOGY REVIEW** by Fred Hapgood (July 11, 2006): Reverse-Engineering the Brain - at MIT, neuroscience and artificial intelligence are beginning to intersect - Earl Miller, Jim DiCarlo and Tomaso Poggio.
- **MCGOVERN INSTITUTE NEWS** by Cathryn M. Delude (February 21, 2006): New Approach Bridges the Gap between Neuronal Activity and Human Brain Imaging - Tomaso Poggio and James DiCarlo.
- **MIT NEWS OFFICE: TECH TALK** by Cathryn M. Delude (November 3, 2005): Neuroscientists Break Code on Sight - Tomaso Poggio and James DiCarlo.
- **NEWSCIENTIST.COM** by Anna Gosline (June 22, 2005): Why Your Brain has a 'Jennifer Aniston Cell' - Itzhak Fried, Rodrigo Quiroga, Christof Koch and Gabriel Kreiman.
- **YAHOO!NEWS** by Malcolm Ritter (June 22, 2005): Brain Cells 'Recognize' Famous People - Itzhak Fried, Rodrigo Quiroga, Christof Koch and Gabriel Kreiman.
- **CALTECH MEDIA** by Dan Page (June 22, 2005): Single-Cell Recognition: A Halle Berry Brain Cell - Itzhak Fried, Rodrigo Quiroga, Christof Koch and Gabriel Kreiman.
- **DISCOVER** by John Hogan (June 2005): Can a Single Brain Cell Think? - Itzhak Fried, Rodrigo Quiroga, Christof Koch and Gabriel Kreiman.
- **IL SOLE 24 ORE** by Rosanna Mameli (December 31, 2004): La Formula dell' Apprendimento - Tomaso Poggio.
- **SALK INSTITUTE FOR BIOLOGICAL STUDIES,** La Jolla, CA (September 27, 2004): Remembering Francis Crick.
- **INCONTRI DELLA LOGGIA**, Levanto, Italy (July 16, 2004): "Geni e Memi" - Incontro Dibattito: Scienza e Futuro della Nostra Società – Tomaso Poggio.
- **MIT NEWS OFFICE: TECH TALK**, Source: Picower Center for Learning and Memory (April 2004): "Experimental Evidence for an Old Theory" - See: Brain Circuitry Findings Could Shape Computer Design - Guosong Liu (and T. Poggio).
- **MIT NEWS OFFICE: TECH TALK** by Elizabeth Thomson (April 1, 2004): MIT Team Reports New Insights in Visual Recognition - Pawan Sinha, David Cox and Ethan Meyers.
- **NEWS AND VIEWS** by Carlo Tomasi (March 25, 2004): Past Performance and Future Results - Tomaso Poggio.

- **YAHOO! FINANCE**, Source: The McGovern Institute for Brain Research (March 25, 2004): McGovern Institute's Tomaso Poggio Offers New Paradigm for Understanding Learning - Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi.
- **IL SECOLO XIX** by Gilda Ferrari (February 14, 2004): Lo Scienziato Del MIT - Poggio: L'IIT, Un Sogno Americano - Tomaso Poggio.

Appendices: details on some of the accomplishments

1. Model of the ventral stream of visual cortex for object recognition

We have been using the quantitative model of visual cortex for object recognition tasks. The model achieves outstanding performance on a detection task involving a variety of object categories and simultaneously enables learning from only a few training examples. The resulting system outperforms state-of-the-art systems over a variety of object image data sets from different groups. The model detects both the large amorphous objects (trees, sky, buildings and road), and the rigid objects (cars, pedestrians, bikes). On both of these tasks the system outperforms other systems on a very large and challenging data set we collected as a benchmark. Moreover, the features used by the model are excellent features for learning from only few training examples. Table 1 summarizes the comparisons we performed between the model and other state-of-the-art computer vision systems on datasets from several groups (including our own StreetScene Database).

The model fits well neuroscience data.

This new set of features is indeed qualitatively and quantitatively consistent with several properties of cells in V1, V2, V4, and IT, PFC as well as several fMRI and psychophysical data. For instance, the model predicts, at the C1 and C2 levels respectively, the max-like behavior of a subclass of complex cells in V1 and V4. It also agrees with other data in V4 (Reynolds *et al.*, 1999) about the response of neurons to combinations of simple two-bar stimuli (within the receptive field of the S2 units) and some of the C2 units in the model show a tuning for boundary conformations (Pasupathy & Connor, 2001) which is consistent with recordings from V4 (Serre *et al.*, 2005). Read-out from C2b units in the model predicted (Serre *et al.*, 2005) recent read-out experiments in IT (Hung *et al.*, 2005), showing very similar selectivity and invariance for the same set of stimuli.

The model mimics human performance on a challenging detection task.

The new set of features, when used to classify between animal and non-animal images performed at the level of human observers (with rapid presentations). The model was shown to predict the pattern of performance of human observers on different animal subcategories (see Fig. 1). Additionally we found that both the model and human observers tend to produce similar responses (both correct and incorrect). The overall image-by-image correlation between the model and human observers is high (specifically 0.71, 0.84, 0.71 and 0.60 for heads, close-body, medium-body and far-body respectively, with p value $p < 0.01$). Finally we found that surprisingly the model and human observers exhibit a similar robustness to image orientation (90° rotation and inversion).

The model's implementation has been improved significantly.

In the last year many improvements to the model occurred. These include more effective representation at the higher level of the model, the addition of image descriptors that capture important gestalt properties and a significant speedup (see slide #). In the last year the model improved significantly in accuracy while achieving a considerable drop in run time.

		Welzmann				CalTech			MIT-CBCL	
		Fac	Cow	Lea	Car	Fac	Air	Mot	Fac	Car
[Serre et al, 2005]	**Model [Serre et al, 2005]**			**97.0**	**99.7**	**98.2**	**96.7**	**98.0**	**95.9**	**95.1**
	Constellation [Weber et al, 2000, Fergus et al, 2003]			84.0	84.8	96.4	94.0	95.0		
	Component-based [Heisele et al, 2002]								90.4	
	Component-based [Leung, 2004]									75.4
[Chikkerur & Wolf, 2006]	**Model [Serre et al, 2005]**	**100.0**	**92.0**	**97.9**		**94.5**		**96.5**		
	Fragments [Epshtein & Ullman, 2005]	98.0	78.7	87.4		66.8		52.6		
	Single template SVM	100.0	77.3	71.6		62.2		65.6		
		MIT-CBCL Street Scene Database								
		Bik	Ped	Car	Bui	Tre	Roa	Sky		
[Bileschi & Wolf, 2005]	**Model [Serre et al, 2005]**	**87.8 84.1**	**81.7 88.8**	**89.6 92.9**	**80.3**	**90.8**	**88.9**	**94.7**		
	Component-based [Torralba et al, 2004]	68.5	79.8	69.9						
	Part-based [Leibe et al, 2004]	80.9	85.2	85.9						
	Single template SVM	67.8	70.0	85.0						
	Blobworld [Carson et al, 1999]				66.1	85.8	73.1	68.2		
	Texton [Renninger & Malik, 2002]				69.7	70.4	58.1	65.1		
	Histogram of edges				63.3	63.7	73.3	68.3		

Table 1: Summary of the comparisons performed between the model and other computer vision systems. For all comparisons, all systems were trained and tested on the same sets.

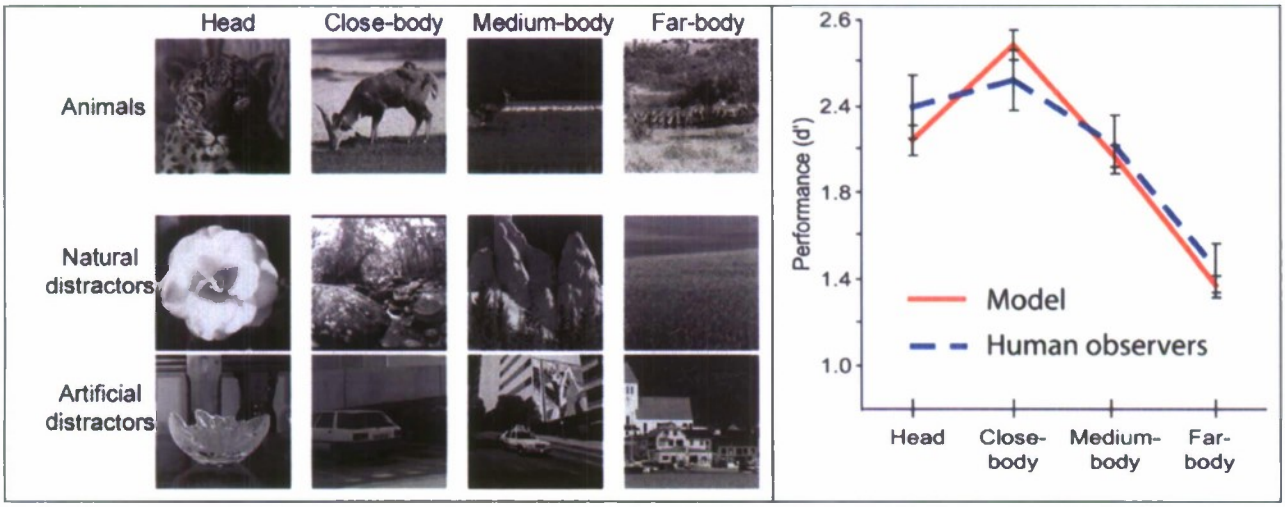

Figure 1: Comparison between the model and human observers on a rapid animal vs. non-animal categorization task. (left) The stimulus dataset, *i.e.*, four animal subcategories with matching distractors. (right) Performance of the model and human observers. The error measure reported is the d' which is a sensitivity measure that combines both the hit and false-alarm rates of each observer into one standardized score.

2. Comparison with physiological observations

The quantitative implementation of the model allows for direct comparisons between the responses of units in the model and electrophysiological recordings from neurons in the visual cortex. Here we illustrate this approach by directly comparing the model against recordings from the macaque monkey area V4 and inferior temporal cortex while the animal was passively viewing complex images.

The model includes several layers that are meant to mimic visual areas V1, V2, V4 and IT cortex. We directly compared the responses of the model units against electrophysiological recordings obtained throughout all these visual areas. The model is able to account for many physiological observations in early visual areas. For instance, at the level of V1, model units agree with the tuning properties of cortical cells in terms of both frequency and orientation bandwidth, as well as peak frequency selectivity and receptive field sizes (see (Serre and Riesenhuber, 2004)). Also in V1, we observe that model units in the C1 layer can explain responses of a subpopulation of complex cells obtained upon presenting two oriented bars within the receptive field (Lampl et al., 2004). At the level of V4, model C2 units exhibit tuning for complex gratings (based on the recordings from (Gallant et al., 1996)), and curvature (based on (Pasupathy and Connor, 2001)), as well as interactions of multiple dots (based on (Freiwald et al., 2005)) or the simultaneous presentation of two-bar stimuli (based on (Reynolds et al., 1999), see (Serre et al., 2005) for details).

Here we focus on one comparison between C2 units and the responses of V4 cells. Figure 3 shows the side-by-side comparison between a model C2 unit and V4 cell responses to the presentation of one-bar and two-bar stimuli. As in (Reynolds et al., 1999) model units were presented with either 1) a reference stimulus alone (an oriented bar at position 1, see Figure 3**A**), 2) a probe stimulus alone (an oriented bar at position 2) or 3) both a reference and a probe stimulus simultaneously. We used stimuli of 16 different orientations for a total of $289 = (16 + 1)^2$

total stimulus combinations for each unit (see (Serre et al., 2005) for details). Each unit's response was normalized by the maximal response of the unit across all conditions. As in (Reynolds et al., 1999) we computed a selectivity index as the normalized response of the unit to the reference stimulus minus the normalized response of the unit to one of the probe stimuli. This index was computed for each of the probe stimuli, yielding 16 selectivity values for each model unit. This selectivity index ranges from −1 to +1, with negative values indicating that the reference stimulus elicited the stronger response, a value of 0 indicating identical responses to reference and probe, and positive values indicating that the probe stimulus elicited the strongest response. We also computed a sensory interaction index which corresponds to the normalized response to a pair of stimuli (the reference and a probe) minus the normalized response to the reference alone. The selectivity index also takes on values from −1 to +1. Negative values indicate that the response to the pair is smaller than the response to the reference stimulus alone (i.e., adding the probe stimulus suppresses the neuronal response). A value of 0 indicates that adding the probe stimulus has no effect on the neuron's response while positive values indicate that adding the probe increases the neuron's response.

As shown in figure 3**B**, model C2 units and V4 cells behave very similarly to the presentation of two stimuli within their receptive field. Indeed the slope of the selectivity vs. sensory interaction indices is about 0.5 for both model units and cortical cells. That is, at the population level, presenting a preferred and a non-preferred stimulus together produces a neural response that falls between the neural responses to the two stimuli individually, sometimes close to an average.1 We have found that such a "clutter effect" also happens higher up in the hierarchy at the level of IT, see (Serre et al., 2005). Since normal vision operates with many objects appearing within the same receptive fields and embedded in complex textures (unlike the artificial experimental setups), understanding the behavior of neurons under clutter conditions is important and warrants more experiments (see later section 3.2.4 and section 4.2). In sum, the model can capture many aspects of the physiological responses of neurons along the ventral visual stream from V1 to IT cortex (see also (Serre et al., 2005)).

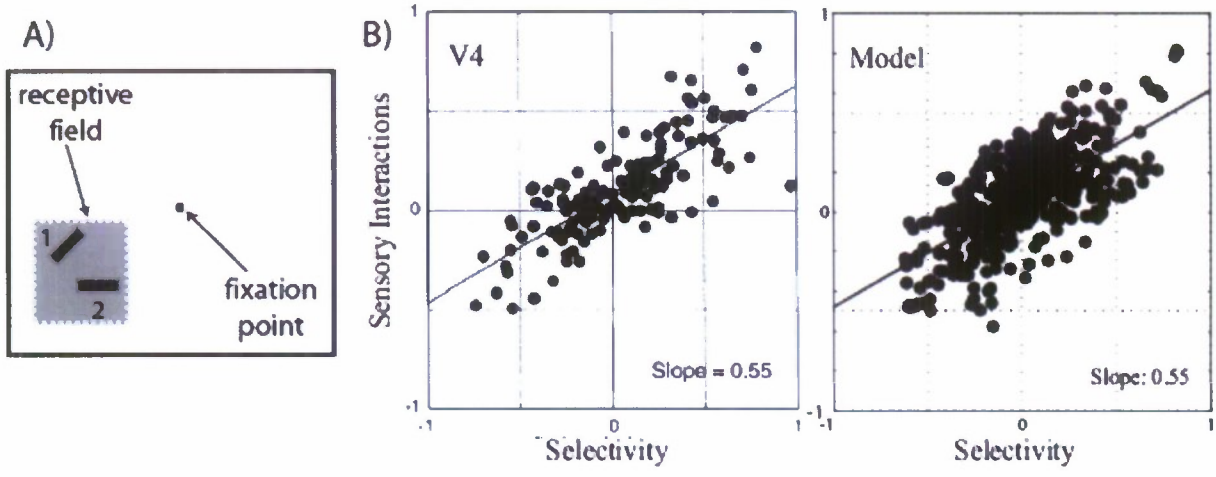

Figure 2: A quantitative comparison between model C2 units and V4 cells. **A)** Stimulus configuration (modified from Figure 1A in (Reynolds et al., 1999)): The stimulus in position 1 is denoted as the reference and the stimulus in position 2 as the probe. As in (Reynolds et al., 1999) we computed a selectivity index (which indicates how selective a cell is to an isolated stimulus in position 1 vs. position 2 alone) and a sensory interaction index (which indicates how selective the cell is to the paired stimuli vs. the reference stimulus alone), see text and (Serre et al., 2005) for details. **B)** Side by-side comparison between V4 neurons (left, adapted from Fig. 5 in (Reynolds et al., 1999)) while the monkey attends away from the receptive field location and C2 units (right). Consistent with the physiology, the addition of a second stimulus in the receptive field of the C2 unit moves the response of the unit toward that of the second stimulus alone, i.e., the response to the clutter condition lies between the responses to the individual stimuli.

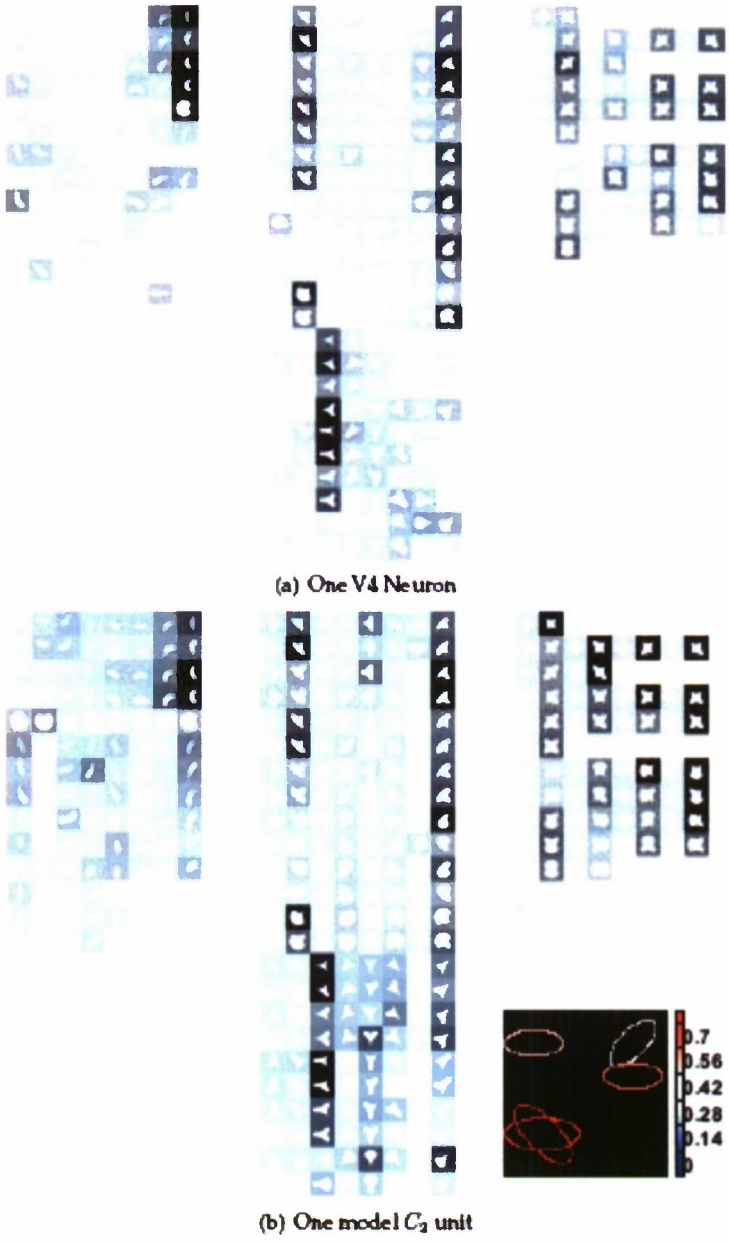

Figure 3: A comparison between the response of a single V4 neuron (corresponding to Fig. 4A in (Pasupathy and Connor, 2001)) **(a)** and a single model C2 unit **(b)** over the boundary conformation stimulus set. The gray level of the stimulus background indicates the response magnitude to each stimulus (the darker the shading the stronger the response). The model unit was picked from the population of 109 model C2 units under study. Both units exhibit very similar pattern of responses (overall correlation r = 0.78). The fit between the model unit and the V4 neuron is quiet remarkable given that there was no fitting procedure involved here for learning the weights of the model unit: The unit was simply selected from a small population of 109 model units learned from natural images and selected at random. The inset on the lower right end of the figure at the bottom describes the corresponding receptive field organization of the C2 unit. Each oriented ellipse characterizes one subfield at matching orientation. Color encodes for the strength of the connection between the subfield and the unit.

Decoding object information from IT and model units

We recently used a simple linear statistical classifier to quantitatively show that we could accurately, rapidly and robustly decode visual information about objects from the activity of small populations of neurons in anterior inferior temporal cortex (Hung et al., 2005). In collaboration with Chou Hung and James DiCarlo at MIT, we observed that a binary response from the neurons (using small bins of 12.5 ms to count spikes) was sufficient to encode information with high accuracy. This robust visual information, as measured by our classifiers, could in principle be decoded by the targets of IT cortex such as prefrontal cortex to determine the class or identity of an object (Miller, 2000). Importantly, the population response generalized across object positions and scales. This scale and position invariance was evident even for novel objects that the animal never observed before (see also (Logothetis et al., 1995)). The observation that scale and position invariance occurs for novel objects strongly suggests that these two forms of invariance do not require multiple examples of each specific object. This should be contrasted with other forms of invariance, such as robustness to depth rotation, which requires multiple views in order to be able to generalize (Poggio and Edelman, 1990).

We examined the responses of the model units to the same set of 77 complex object images seen by the monkey. These objects were divided into 8 possible categories. The model unit responses were divided into a training set and a test set. We used a one-versus-all approach, training 8 binary classifiers, one for each category against the rest of the categories, and then taking the classifier prediction to be the maximum among the 8 classifiers (for further details, see (Hung et al., 2005; Serre et al., 2005)). Similar observations were made when trying to identify each individual object by training 77 binary classifiers. For comparison, we also tried decoding object category from a random selection of model units from other layers of the model. The input to the classifier consisted of the responses of randomly selected model units and the labels of the object categories (or object identities for the identification task). Data from multiple units were concatenated assuming independence.

We observed that we could accurately read out the object category and identity from model units. In Figure 3**A**, we compare the classification performance, for the categorization task described above, between the IT neurons and the C2b model units. In agreement with the experimental data from IT, units from the C2b stage of the model yielded a high level of performance (> 70% for 100 units; where chance was 12.5%). We observed that the physiological observations were in agreement with the predictions made by the highest layers in the model (C2b, S4) but not by earlier stages (S1 through S2). As expected, the layers from S1 through S2 showed a weaker degree of scale and position invariance.

The classification performance of S2b units (the input to C2b units, see Figure 1) was qualitatively close to the performance of local field potentials (LFPs) in IT cortex (Kreiman et al., 2006). The main components of LFPs are dendritic potentials and therefore LFPs are generally considered to represent the dendritic input and local processing within a cortical area (Mitzdorf, 1985; Logothetis et al., 2001). Thus, it is tempting to speculate that the S2b responses in the model capture the type of information conveyed by LFPs in IT. However, care should be taken in this interpretation as the LFPs constitute an aggregate measure of the activity over many different types of neurons and large areas. Further investigation of the nature of the LFPs and their relation with the spiking responses could help unravel the transformations that take place across cortical layers.

The pattern of errors made by the classifier indicates that some groups were easier to discriminate than others. This was also evident in the correlation matrix of the population

responses between all pairs of pictures (Serre et al., 2005; Hung et al., 2005). The units yielded similar responses to stimuli that looked alike at the pixel level. The performance of the classifier for categorization dropped significantly upon arbitrarily defining the categories as random groups of pictures.

We also tested the ability of the model to generalize to novel stimuli not included in the training set. The performance values shown in Figure 3 are based on the responses of model units to single stimulus presentations that were not included in the classifier training and correspond to the results obtained using a linear classifier. Although the way in which the weights were learned (using a support vector machine classifier) is probably very different in biology (see (Serre, 2006)), once the weights are established the linear classification boundary could very easily be implemented by neuronal hardware. Therefore, the recognition performance provides a lower bound to what a real downstream unit (e.g., in PFC) could, in theory, perform on a single trial given input consisting of a few spikes from the neurons in IT cortex.

Overall, we observed that the population of C2b model units yields a read-out performance level that is very similar to the one observed from a population of IT neurons.

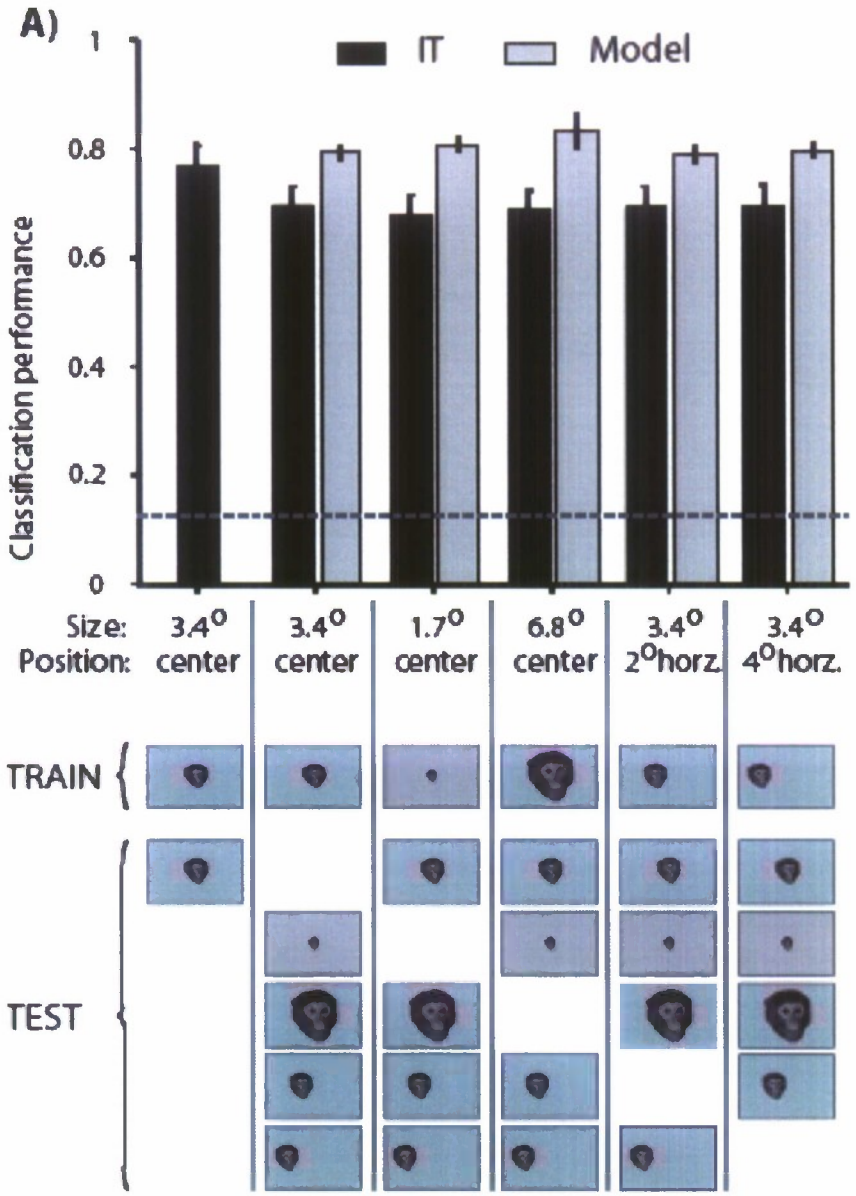

Figure 4: Classification performance based on the spiking activity from IT neurons (black) and C2b units from the model (gray). The performance shown here is based on the categorization task where the classifier was trained based on the category of the object. A linear classifier was trained using the responses to the 77 objects at a single scale and position (shown for one object by "TRAIN"). The classifier performance was evaluated using shifted or scaled versions of the same 77 objects (shown for one object by "TEST"). During training, the classifier was never presented with the unit responses to the shifted or scaled objects. The left-most column shows the performance for training and testing on separate repetitions of the objects at the same standard position and scale (this is shown only for the IT neurons because there is no variability in the model which is deterministic). The second bar shows the performance after training on the standard position and scale (3.4 degrees, center of gaze) and testing on the shifted and scaled images. The dashed horizontal line indicates chance performance (12.5%, 1 out of 8 possible categories). Error bars show standard deviations over 20 random choices of the units used for training/testing.

3. Comparison between the model and other state-of-the-art computer vision systems

CalTech-101

We compared the model to the SIFT features (Lowe 1999; Lowe 2004) on the CalTech-101 database. As illustrated on Fig. 4, the model C2 features exhibit higher performance.

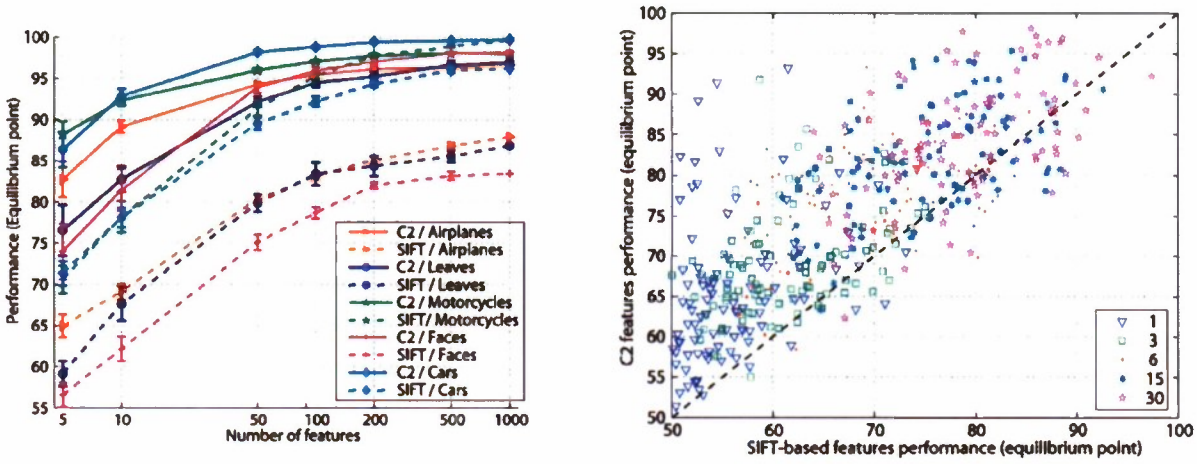

Figure 5: Comparison with SiFT features on the calTech-101.

MIT face and car database

We compared the performance of the C2 units to two computer vision systems that were developed in the lab. The two benchmarks are also hierarchical (a first layer of SVM classifiers detect object components and a second layer check for their configuration). Model C2 units outperform both systems.

Datasets	Benchmark	Model
MIT-CBCL faces (Heisele et al 2002)	90.4% correct	95.9% correct
MIT-CBCL cars (Leung 2004)	75.4% correct	95.1% correct

StreetScene database

Rigid-objects:

For comparison, we also implemented four other benchmark systems. Our most simple baseline detector is a single-template Grayscale system: Each image is normalized in size and histogram equalized before the gray-values are passed to a linear classifier (gentleBoost). Another baseline detector, Local Patch Correlation, is built using patch-based features similar to [45]. Each feature fi is associated with a particular image patch pi, extracted randomly from the training set. Each feature fi is calculated in a test image as the maximum normalized cross correlation of pi within a subwindow of the image. This window of support is equal to a rectangle three times the size of pi and centered in the image at the same relative location from which pi was originally extracted. The advantage of the patch-based features over the single-template approach is that local patches can be highly selective while maintaining a degree of position invariance. The system was implemented with N = 1,024 features and with patches of size 12 X 12 in images of size 128 X 128. The third benchmark system is a Part-based system as described in (Leibe et al, 2004). Briefly, both object parts and a geometric model are learned via image patch clustering. The detection stage is performed by redetecting these parts and allowing them to vote for objects-at-poses in a generalized Hough transform framework. Finally, we compare to an implementation of the Histogram of Gradients (HoG) feature of (Dalal & Triggs, 2005), which has shown excellent performance on these types of objects. All benchmark systems were trained and tested on the same data sets as the SMFs-based system. They all use gentleBoost except (Leibe et al, 2004).

The ROC results of this experiment are illustrated in Fig. 5. For the two (C1 and C2) SMFs-based systems, the Grayscale as well as the Local Patch Correlation system, the classifier is GentleBoost, but we found very similar results with both a linear and a polynomial-kernel SVM. Overall, for all thre object categories tested, the SMFs-based system performs best on cars and bicycles and second behind HoG on pedestrians (the HoG system was parameter-tuned in (Dalal & Triggs, 2005) to achieve maximal performance on this one class). Finally, for this recognition task, i.e., with a windowing framework, the C1 SMFs seem to be superior to the C2 SMFs.

Textured-objects:

We implemented four benchmark texture classification systems. The Blobworld (BW) system was constructed as described in (Carson et al, 1999.) Briefly, the Blobworld feature, originally designed for image segmentation, is a six-dimensional vector at each pixel location; three dimensions encode color in the Lab color space and three dimensions encode texture using the local spectrum of gradient responses. We did not include the color information for a fair comparison between all the various texture detection methods.

The systems labeled T1 and T2 are based on (Renninger & Malik, 2004). In these systems, the test image is first processed with a number of predefined filters. T1 uses 36 oriented edge-filters arranged in five degrees increments from 0 degrees to 180 degrees. T2 follows (Renninger & Malik, 2004) exactly by using 36 Gabor filters at six orientations, three scales, and two phases. For both systems independently, a large number of random samples of the 36-dimensional edge response images were taken and subsequently clustered using k-means to find 100 cluster centroids (i.e., the textons). The texton image was then calculated by finding the index of the nearest texton to the filter response vector at each pixel in the response images. A 100-dimensional texton feature vector was then built by calculating the local 10 X 10 histogram of

nearest texton indexes. Finally, the Histogram of edges (HoE) system was built by simply using the same type of histogram framework, but over the local 36-dimensional directional filter responses (using the filters of T1) rather than the texton identity. Here, as well, learning was done using the gentleBoost algorithm (again a linear SVM produced very similar results). The within-class variability of the texture-objects in this test is considerably larger than that of the texture classes usually used to test texture-detection systems, making this task somewhat different. This may explain the relatively poor performance of some of these systems on certain objects.

As shown in Fig. 6, the SMFs-based texture system seems to consistently outperform the benchmarks (BW, T1, T2, and HoE). C2 compared to C1 SMFs may be better suited to this task because of their increased invariance properties and complexity.

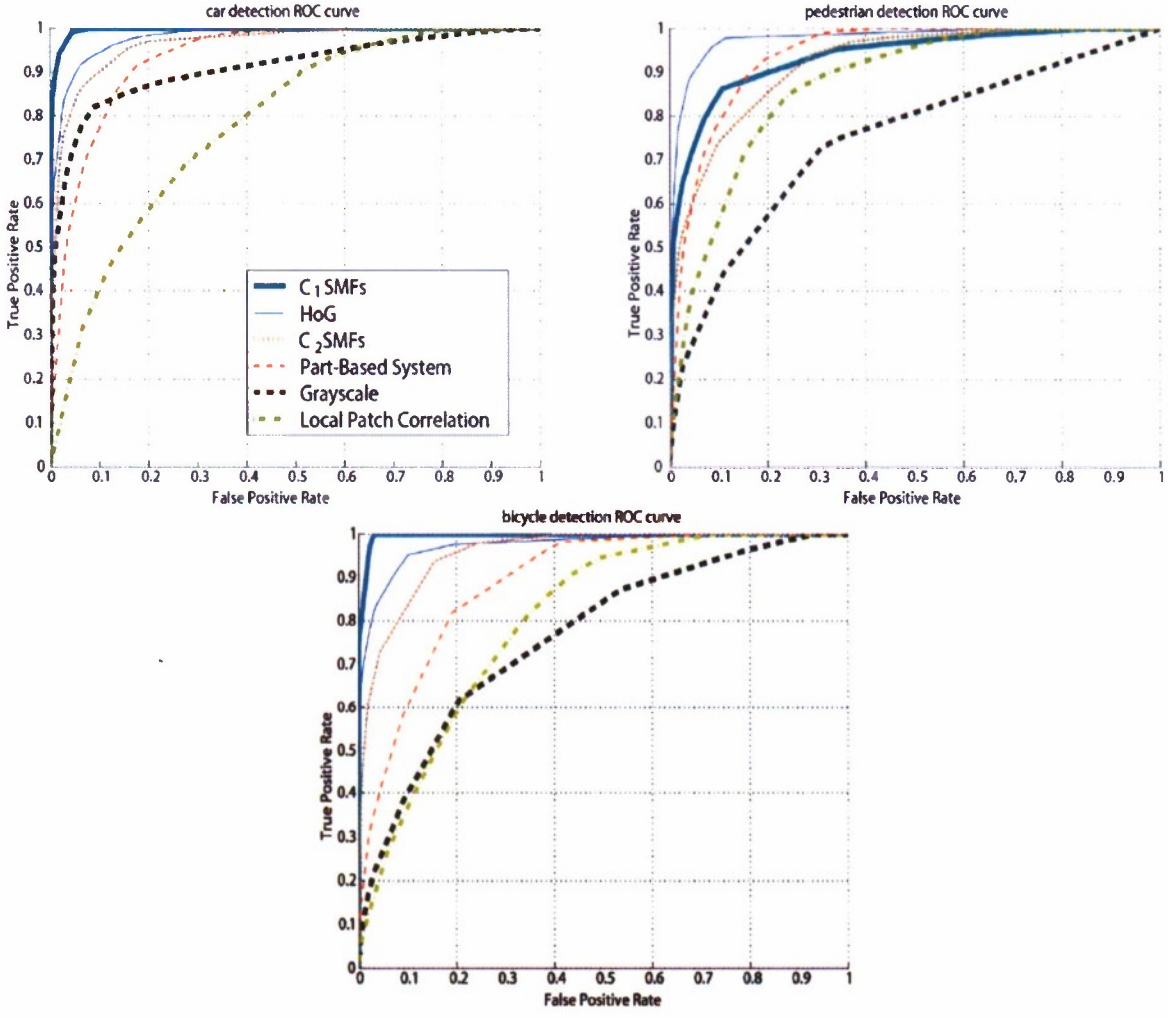

Figure 6: Comparison between the model (C1 SMFs and C2SMFs) and other state of the art systems on the MIT StreetScene database for the recognition of rigid objects.

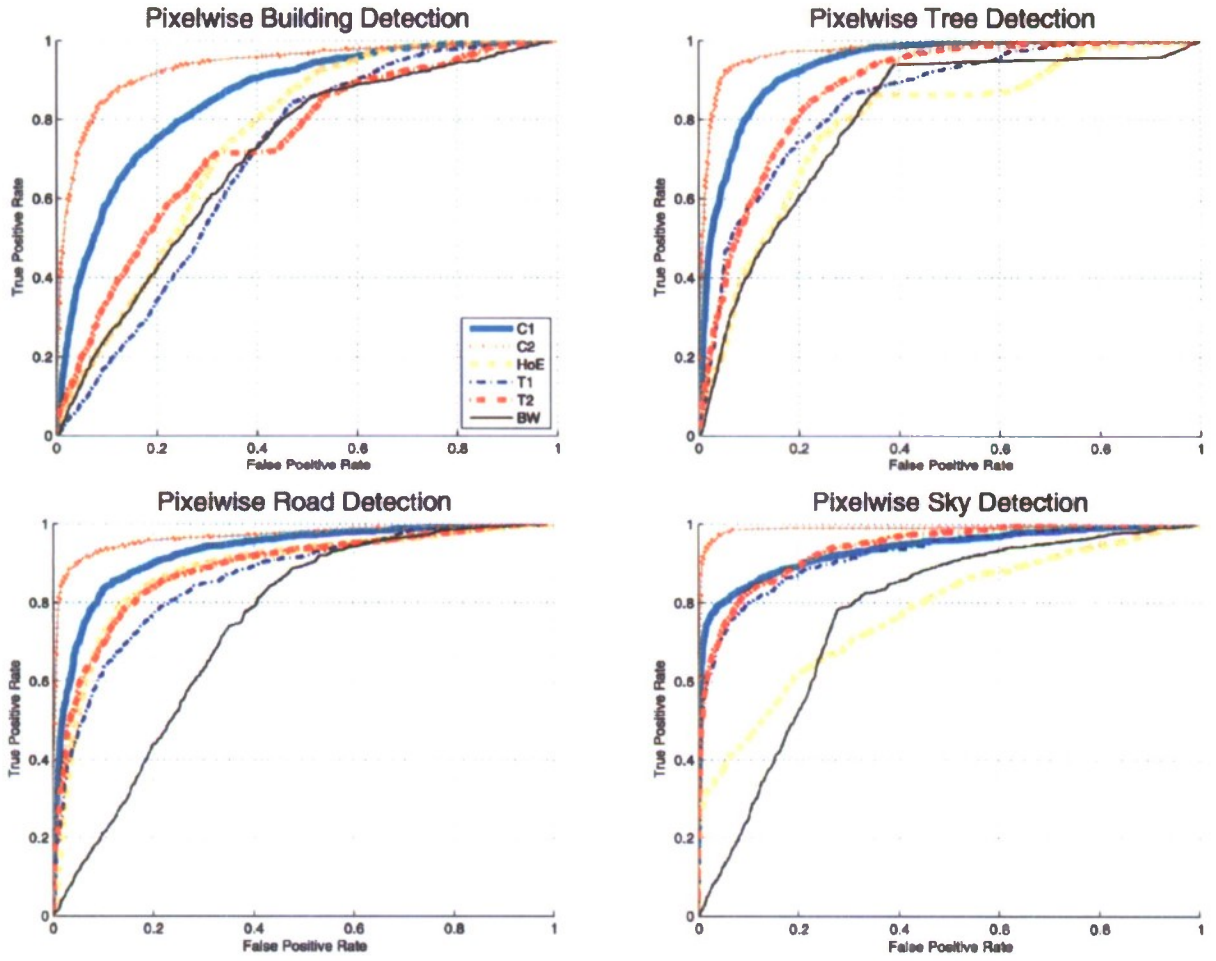

Figure: Comparison between the model (C1 and C2) with other state of the art systems on the MIT StreetScene database for the recognition of textured-objects.

4. Automatic recognition of actions in videos: a tool for behavioral phenotyping

During the last year, we developed a prototype system for the recognition of basic rodent behaviors (see Fig. 1 for examples). The work was based on a model of the dorsal pathway in the visual cortex which also outlines a computer implementation and its state-of-the-art performance in the recognition of human actions.

With the McGovern funding, we were able to collect about 100 hours of mouse monitoring to train and test the system. We video recorded singly housed mice from an angle perpendicular to the side of the cage (Figure below). In order to train a robust detection system, we used at least six different camera angles in our training (and test) set, all of which had slightly different lighting conditions. In addition, we utilized mice of different size, gender, and coat color. Several summer students manually annotated these videos. Table 1 gives a comparison between the performance of our system and comparisons with human labelers and with the Clever Sys. Commercial system. Our system achieves near-human level performance on this task and is significantly better than an existing commercial system. A demo of the system can be found at http://techtv.mit.edu/videos/1838.

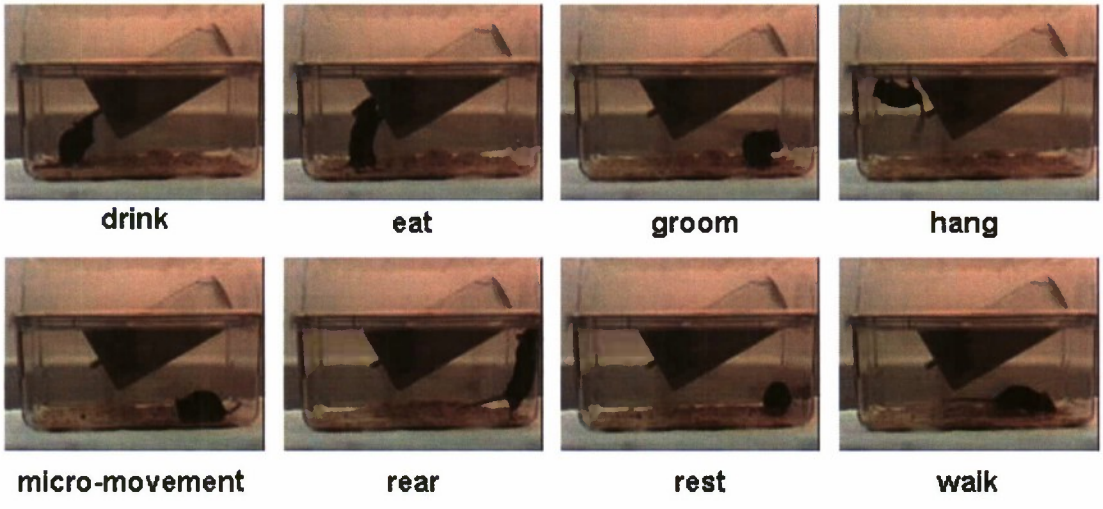

Figure 1: Snapshots taken from representative videos for 8 types of behavior that the system was trained to recognize.

	Our system	Clever Sys. Commercial system	Inter-human agreement
Performance	71.0	56.0	71.6

Table 1: Performance of the system (percent frames correctly classified) and comparison with an available commercial system and human as measured by the agreement on the labeling performed by two independent labelers.

5. Hierarchical Kernel Machines

We have developed a mathematical framework to analyze hierarchical kernel machines motivated by the architecture of the primate visual cortex. The main motivations for the project are two: 1) primates seem to be able to learn complex tasks from far fewer examples than our present non-hierarchical kernel–based learning algorithms predict, e.g. they are able to solve the "poverty of stimulus" problem 2) a preliminary computational model at MIT of the feedforward flow of information in visual cortex performs well on difficult recognition tasks compared to existing computer vision systems. What is needed now is an approach based on a *mathematical theory* – to explain why a hierarchy is needed, under which conditions, and to provide a framework for optimizing the learning architecture and its parameters. A theory, such as the one of which we have the foundations, will explain why hierarchical models work as well as they do and what the computational reasons for the hierarchical organization of cortex are, leading to potentially significant contributions to outstanding challenges in machine learning and computer science. The development of new powerful learning techniques such as the hierarchical kernel machines we propose can be of pervasive importance for many capabilities of the DoD, because machine learning is becoming the common mathematics language across different areas of computer science and because of the reliance of the DoD on computers and algorithms. Our theory should be relevant to help preprocess and interpret the huge flow of electronic information -- such as images -- provided by different types of sensors. Navigation, surveillance and intelligence are just three of the areas that could be hugely impacted by the development of novel learning techniques, inspired by the brain and based on solid mathematical foundations.