Learning techniques for adaptive image analysis and object detection

Tomaso Poggio and Federico Girosi

1 Executive Summary

We had proposed to extend our research in learning and networks techniques for object detection with the goal of applying them to target detection. The motivation of our research is the belief that learning and adaptation techniques are needed for robust, flexible and environment-independent vision-based applications such as automatic driving. The goal is therefore to develop techniques for image analysis – in part inspired by the human visual system – which adapt and learn and do not need significant reprogramming for different situations.

The research carried out last year (1996) comprised two main parts.

1. A theoretical characterization of the complexity of the learning
2. Techniques to learn flexible templates of targets from a set of images

For the next year we plan to develop a real-time vision architecture for estimating parameters from images. Our initial testbed will be in the realm of trainable man-machine interfaces.

2 Project 1 (1996): Prior Information and Generalized Questions

The Problem:

Empirical learning from examples has to enable generalization to new situations to be useful. In addition to the standard empirical examples the ability to generalize is based on generalized information, usually called prior information. While in principle prior information should be seen as empirical data, just like standard empirical examples, the more complex nature of generalized information usually prevents a unique treatment of those different kinds of informations in existing learning algorithms. To achieve a systematic treatment of prior information and examples on equal footing and to allow the incorporation of all forms of generalized data in the learning process one has to face the two problems of 1. empirical measurement of generalized data, and 2. of constructing learning algorithms for generalized data.

Motivation:

Despite the importance of nonlocal information for any generalization ability there is nearly no discussion of nonlocal empirical measurements in the learning literature. Therefore nonlocal information is mostly implemented quite ad hoc using some smoothness or symmetry considerations. But, besides smoothness or symmetry there can be many more nonlocal data available like differences, (local) averages, Fourier components or typical building blocks. As the number of available example data is very often small compared to the complexity of the task under consideration the use of prior or generalized information is crucial to achieve a reasonable generalization behavior for most real world applications [12]. This shows clearly the importance of research on empirical nonlocal measurements and algorithms for generalized data.

Previous Work:

Usually nonlocal prior information is included by restricting the space of approximating functions or more general by weighting/penalizing functions using extra (regularization) terms added to the training error[11]. There are many variants of dealing with these extra terms [8], e.g. analytical solutions [7], creating virtual examples [4] [6] [10] or directly include the additional terms in numerical minimization procedures like backpropagation [1]. But as prior information is normally not part of the error or lossfunction of interest the corresponding (regularization) constants multiplying the additional terms must be determined in the finite case by crossvalidation with respect to the true error or lossfunction (without additional terms). In all these approaches the treatment of prior information is somewhat different from that of the standard training examples.

Bayesian Decision Theory [3] is general enough to deal with generalized data, but mostly used in connection with the Maximum a Posteriori Approximation and Gaussian noise assumptions, being equivalent to the minimization of error with additional terms. Recently Monte Carlo Integration methods have been applied [9] [5] making a full Bayesian treatment in many cases numerically feasible. Bayesian Decision Theory is therefore a possible starting point for a systematic treatment of generalized information.

Approach:

Data are formalized as question answer pairs. A set of basic questions x and corresponding probabilities of answers $p(y|x, f^0)$ completely characterizes a state f^0 of nature. Generalized information are answers to generalized questions. A generalized question is any probabilistic functional of the state of nature whose f^0–dependence can be expressed by the $p(y|x, f^0)$. Local data or standard training examples are pairs $(x, f(x))$ corresponding to δ–like functionals. Examples for nonlocal data include differences (local smoothness, scalar symmetries) $f(x) - f(x')$ and sums (averages) $f(x) + f(x')$ as well as (arbitrary) nonlinear combinations like extrema or power spectra. With that formalism it can be shown that starting from a so called "tabula rasa" or factorial state of knowledge no learning algorithm can generalize using only local examples, a generalization of No Free Lunch Theorems [14].

Analysis shows that data which enable generalization 1. necessarily has to include nonlocal data depending on different x–values at the same time, 2. most times are not included directly in the error or lossfunction of interest and 3. often are common to a family of similar tasks. This leads to a classification of generalized information according to the three aspects:

1. local vs. nonlocal data (generalization aspect)
2. directly tested vs. not directly tested data (application aspect)
3. task specific vs. common data (transfer aspect)

As these aspects are independent many nonstandard situations can be described not covered by classical learning scenarios. Nonlocality requires more than one function evaluation (e.g. backpropagation step) for every update during minimization. The case of available information not directly related to intended future applications is covered by the Bayesian Decision Theory, while methods of classical statistics have problems when the training set is not sampled according to the distribution of the test set. The tranfer aspect in form of a hierachical capacity control [2] will probably be of large practical importance in many future applications. For illustration take the ability of humans to recognize a person's face after having seen a photo of it. Such a number of examples clearly is too small to learn this task from scratch as it strongly restricts the maximal capacity of the learning machine one can use. But there can be related (sub)tasks for which a lot more training examples are available, like some more faces of different persons, some examples of a rotated face of another person, some animal faces and drawings, an even larger sample of approximately lateral symmetric objects, and a huge sample of images containing macroscopic objects.

Difficulty:

Numerical minimization within a full Bayesian model can imply huge numerical integrations when no simplifying approximations are available. Those cases require large computational resources.

Impact:

As nonlocal information is necessary for any generalization ability the results could be interesting for most real world learning problems. If the number of available local examples itself only allows learning machines with too low a capacity compared to the task complexity, any additional nonlocal data could be helpful. The development of a general framework helps comparing different algorithms including the discussion of their inherent approximations.

References

[1] Abu–Mostafa, Y. (1993) A method for learning from hints. *Advances in Neural Information Processing Systems* **5**, 73–80, San Mateo, CA:Morgan Kauffmann.

[2] Baxter, J. (1995) Learning Internal Representations. PhD thesis, Dept. of Math. and Statistics, The Flinders University of South Australia.

[3] Berger, J.O. (1985) Statistical Decision Theory and Bayesian Analysis. (Second Ed.), New York: Springer.

[4] Beymer, D. & Poggio, T. (1995) Face Recognition From One Example View. M.I.T., Artificial Intelligence Laboratory, Memo No.1536.

[5] Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995) Bayesian Data Analysis. London: Chapman & Hall.

[6] Girosi, F. & Chan, N.T. (1995) Prior Knowledge and the Creation of Virtual Examples for RBF Networks. *IEEE Workshop on Neural Networks for Signal Processing*, September 1995, Cambridge, MA.

[7] Girosi, F., Jones, M. & Poggio, T. (1995) Regularization Theory and Neural Networks Architectures. *Neural Computation*,**7** (2), 219–269.

[8] Leen, T.K. (1995) from Data Distributions to Regularization in Invariant Learning. *Neural Computation*,**7**, 974–981.

[9] Neal, R.M. (1993) Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG–TR–93–1, Dept.] of Comp. Sc., Univ. of Toronto, Canada.

[10] Pomerleau, D. (1991) *Neural Computation* **3**, 88.

[11] Tikhonov, A.N. (1963) Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4**, 1035–1038.

[12] Vapnik, V.N. (1982) Estimation of Dependences Based on Empirical Data. New York: Springer.

[13] Verri, A. & Poggio, T. (1986) Regularization Theory and Shape Constraints. M.I.T., Artificial Intelligence Laboratory, Memo No.916.

[14] Wolpert, D.H. (1996) The Lack of Apriori Distinctions between Learning Algorithms. To appear in *Neural Computation.*

3 Project 2 (1996): Model-based Matching

The Problem: The problem we are trying to solve is to model classes of objects (such as faces) in the computer and then match this model to novel images containing an instance of the object class. To do this we propose a learning based approach which models object classes using a number of example images of the class.

Motivation: Our approach is based on modeling object classes as a linear combination of prototype images. The motivation for this model comes from the linear class concept of Poggio and Vetter (1992) and Vetter and Poggio (1995). They showed that linear transformations can be learned exactly from a small set of examples in the case of linear object classes. Since many object transformations can be approximated by linear transformations, this is a very useful result. As for practical applications which motivate this work, there are a number of important applications including man-machine interfaces, image analysis, video teleconferencing, image compression and object verification.

Previous Work: Cootes and Taylor (1992) have proposed a similar model for object classes. They also use a linear combination of prototypes as an object model. Their work differs from ours in the algorithm they used to fit a model to an image, and in the use of a sparse set of control points as opposed to a full pixelwise correspendence matrix for their models. The work of Ullman and Basri (1989) and Shashua (1992) provided strong motivation for this work. They showed that any view of a single object can be represented as a linear combination of just three views of the object. Our work is concerned with extending this idea to object classes. Another piece of related work is that of Beymer (1995) who used a similar approach to model faces.

Approach: This research focuses on two main problems: modeling object classes and matching the model to a novel image. Our approach to the first problem is to model object classes as a linear combination of example images called prototypes. Linearly combining images means treating the images as vectors. In order to linearly combine these vectors, we must have pixelwise correspondence between the images. Otherwise, the linear combination would not form a vector space. So, in addition to the prototype images as input, we also require the pixelwise correspondences as input. The correspondences are represented as a flow fields from one prototype chosen as the reference prototype and each of the other prototypes. Given this, our model for object classes consists of two components, namely shape and texture. The shape of a model image is a linear combination of prototype shapes (flow fields). Analogously, the texture of a model image is a linear combination of prototype textures (grey level images). The model thus contains two parameters per prototype: one for shape and one for texture. Adjusting these parameters allows one to create a large variety of model images by morphing the reference image according to the linear combination of shape and texture specified by the model parameters.

Our approach to the problem of matching a model to a novel image is the following. We first define an error measure between the input novel image and the current guess for the best fitting model image. This error is simply the L2 error between the pixels (grey level values) of the novel and model images. To obtain the model image, we must morph the reference image according to the model parameters. This produces an image against which we can compare the novel input image. Given this error function we find the best matching model image by optimizing the error with respect to the model parameters using a stochastic gradient descent algorithm.

To illustrate our model and matching algorithm, figure 1 shows a few prototypes from a database of frontal views of faces. Figure 2 shows a few novel faces which were not used as prototypes and the resulting best match after fitting the model. One can see that the model is able to match novel faces well.

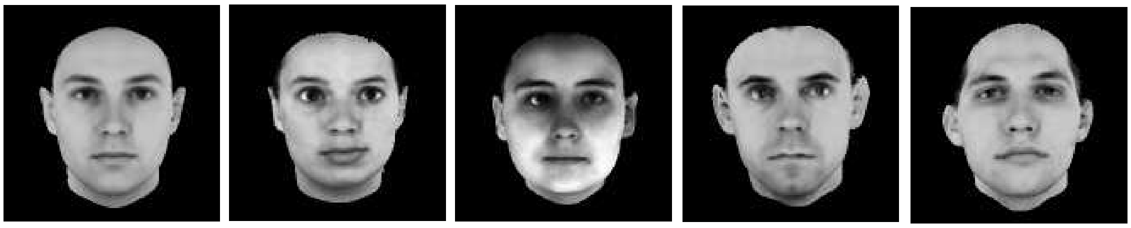

Figure 1: Five of the one hundred prototype images that define the object class of frontal views of faces.

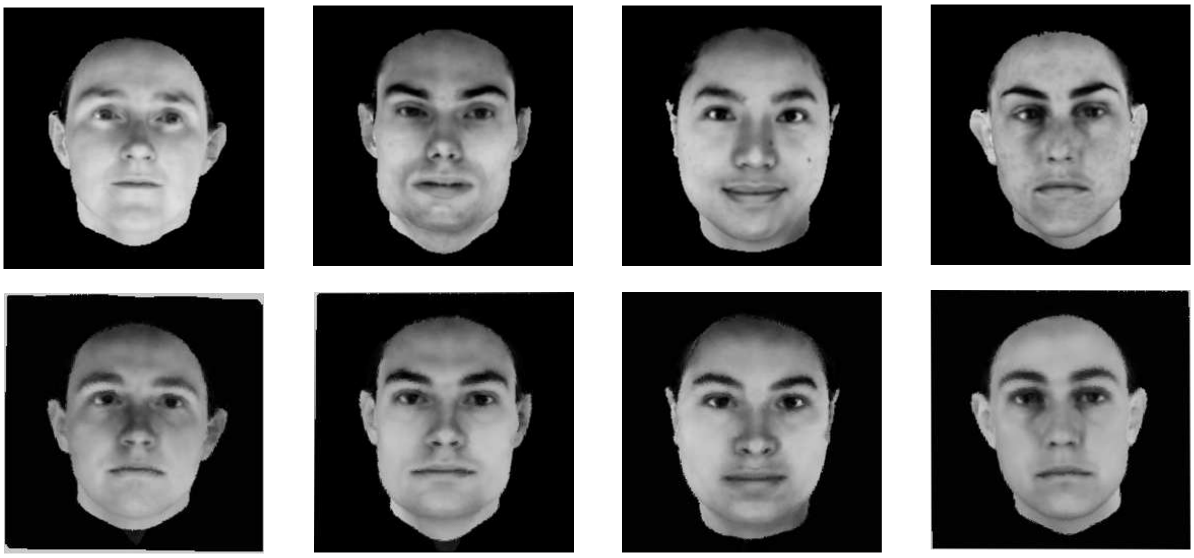

Figure 2: Top row has novel faces that the algorithm attempted to match with the model. Bottom row contains the resulting best fitting model image.

References

[1] Beymer. D. "Vectorizing Face Images by Interleaving Shape and Texture Computations", *AI Memo No. 1537*, MIT, 1995.

[2] Cootes, T.F. and Taylor, C.J. "Active Shape Models - Smart Snakes", In *Proceedings of the British Machine Vision Conference.* Springer-Verlag, 1992, pp. 266-275.

[3] Poggio, T. and Vetter, T. "Recognition and Structure from one 2D Model View: Observations on Prototypes, Object Classes and Symmetries", *AI Memo No. 1347*, MIT, 1992.

[4] Ullman, S. and Basri, R. "Recognition by Linear Combinations of Models", *AI Memo No. 1152*, MIT, 1989.

[5] Vetter, T. and Poggio, T. "Linear Object Classes and Image Synthesis from a Single Example Image", *AI Memo No. 1531*, MIT, 1995.

4 Future plan: an interactive interface for the real time analysis of facial expressions

The Problem: To develop a system capable of analyzing human faces and "mapping" the expressions therein onto a set of parameters which could be used to re-synthesize those expressions.

Motivation: The problem is motivated by its relevance to several computer vision and graphics applications like very low bandwidth video conferencing, virtual actor and video email. In these applications it is fundamental to be able to represent human facial features and expressions in a compact and robust form that can be stored or transmitted without appreciable degradation in quality of their recognition on re-synthesis.

Previous Work: The work done previously in this regard includes [6], [5] and [4]. In [6] a scheme for synthesizing expressions by interpolating between views of faces is used in a feedback loop to estimate the expression parameters that best describe the novel image. This method relies on repeated computation of optical flow which can be computationally intensive. The technique in [5] also relies on computation of optical flow but uses parametric flow models to classify them into expression classes. Another approach at classification of facial parts is attempted in [4] where the outputs of different band-pass components of the input image is used as input to a classifier.

Approach: We first attempt to model and analyze human lip and mouth patterns as a example of facial expressions. We formulate the analysis problem in the framework of learning by examples. For this, we need to gather a database of examples of mouth and lip movements. From these one can proceed in two different directions.

1. **Classification:** This involves learning the classification boundaries between the various expression classes. A promising technique that can accomplish this is the

support vector machine [7, 1]. It has been applied to face detection [2] and we propose to extend it to recognize finer details of expression. The input space in this case will be the pixel map of a block around the mouth of the subject or a set of wavelet functions. To track the position of the mouth quickly we propose to use a color-based segmentation scheme.

2. **Regression:** This involves constructing a smooth surface from a properly defined input space to the set of desired parameters. Regression using support vectors has been implemented [3] and appears promising provided a smooth regression surface exists. Thus the problem here is mainly of selecting a proper input feature space to the support vector regressor.

Difficulty: The selection of the proper feature space is a source of difficulty. Based on the previous work one expects optical flow of the lips to function as a good feature which however can be prohibitive in real-time applications. The approach suggested by [4] appears more suited to real-time applications and can be interpreted in a wavelet-based framework. In the case of classification, one is faced with the problem of the fine granularity of the classification needed in order to obtain a reconstruction of high visual quality. A finer granularity on the other hand may make estimation of classification boundaries difficult. In the case of regression, a smooth surface between the input pixel map and the expression parameters probably does not exist. In that case we need to map the mouth onto some other feature such as lip outline in order to obtain the required regression surface.

Impact: This research will have significant impact on the development of new trainable man-machine interfaces, that will allow a computer to understand the user in more natural ways.

References

[1] C.Cortes and V.Vapnik. Suppost-Vector Machine Networks. *Machine Learning (to appear)*, 1996.

[2] E.E.Osuna. Support Vector Machines: Training and Applications. *CBCL, MIT Tech. Report*, 1996.

[3] H.Drucker, C.J.C.Burges, L.Kaufman, A.Smola and V.Vapnik. Linear Support Vector Regression Machines. In *Some Conference*, 1996.

[4] H.P. Graf, T. Chen, E. Petajan and E. Cosatto. Locating Faces and Facial Parts. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, pages 41–46, Zurich, Switzerland, 1995.

[5] M.J.Black and Y.Yacoob. Recognizing Racial Expressions in Image Sequences using Local Parametrized Models of Image Motion. *Intl. Journal of Computer Vision (submitted)*, 1996.

[6] T.Ezzat. *Example based Analysis and Synthesis for Images of Human Faces.* Springer Verlag, 1995.

[7] V.Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, 1995.

PAPERS PUBLISHED IN REFEREED JOURNALS

Niyogi, P. and Girosi, F. "On The Relationship between Generalization Error, Hypothesis Complexity and Sample Complexity for Radial Basis Functions," Neural Computation, 8, 819-842, 1996.

Poggio, T. and Beymer, D. "Learning to See," IEEE Spectrum, 60-69, May 1996.

Sinha, P. and Poggio, T. "Is Human Face Recognition Better Characterized as Head Recognition (Correspondence), Nature, Vol. 384, No. 6608, 1996.

Sinha, P. and Poggio, T. "Role of Learning in Three-dimensional Form Perception, " Nature, Vol. 384, No. 6608, 460-463, 1996.

Beymer, D. and Poggio, T. "Image Representation for Visual Learning," Science, 272, 1905-1909, 1996.

MEMOS/TECHNICAL REPORTS

"Model-based Matching by Linear Combinations of Prototypes," (M. Jones and T. Poggio). AI Memo 1583/CBCL Paper 139, Massachusetts Institute of Technology, Cambridge, MA, November 1996.

"Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers," (B. Schoelkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio and V. Vapnik). AI Memo 1599/CBCL Paper 142, Massachusetts Institute of Technology, Cambridge, MA, December 1996.

BOOKS/BOOK CHAPTERS PUBLISHED

Poggio, T. and Beymer, D. "Regularization Networks for Visual Learning." In: Early Visual Learning, S. Nayar and T. Poggio (eds.), Oxford University Press, 43-66, 1996.

Poggio, T. and K.-K. Sung, K.K. "Networks that Learn for Image Understanding." In: Advances in Image Understanding, K. Bowyer and N. Ahuja (eds.), IEEE Computer Society Press, 226-240, 1996.

Grimson, W.E.L., Horn, B.K.P., Poggio, T. and the staff of the AI Laboratory. "Progress in Image Understanding at MIT." In: Proceedings of the Image Understanding Workshop, Morgan Kaufmann, San Francisco, CA, 65-74, 1996.

Romano, R., Beymer, D. and Poggio, T. "Face Verification for Real-time Applications." In: Proceedings of the Image Understanding Workshop, Morgan Kaufmann, San Francisco, CA, 1996

Early Visual Learning, S. Nayar and T. Poggio (eds.), Oxford University Press, 1996.

BOOKS/BOOK CHAPTERS IN PRESS

Jones, M. and Poggio, T. "Top-down Learning of Low-level Vision Tasks." In: Computer Vision and Pattern Recognition, submitted November 1996.

Poggio, T. "Networks that Learn and How the Brain Works." In: PSPUM/JERISON, D. Jerison, I.M. Singer and D.W. Stroock (eds.), American Mathematical Society, 1996.

INVITED PRESENTATIONS

Federico Girosi:

Artificial Intelligence Laboratory, MIT, Cambridge, MA, May 6, 1996. "How to Make Approximation Theory Easy with Statistics."

Michael Jones:

AI Student Seminar, MIT, Cambridge, MA, April 1996. "Model-Based Matching Using Prototypes."

Tomaso Poggio:

John Hopkins Univeristy, Baltimore, MD, December 2, 1996. "Object Recognition and IT Cortex."

New York University, New York, NY, November 14, 1996. "Image Representations for Visual Learning."

AT&T Research, Redbanks, NJ, November 12, 1996. "Learning and Networks: Graphics and Vision."

Boston Colloquium for the Philosophy of Science: The Future of Biology, Boston University, Boston, MA, October 2, 1996. "Modeling Systems that Learn."

IAPR 13th International Conference on Pattern Recognition, Vienna, Austria, August 25-29,1996. "Image Representations for Visual Learning."

Symposium Vom Neuron zur Informatik, Bochum, Germany, June 8, 1996. "Learning to See."

MIT, Center for Imaging Science, Woods-Hole Workshop, May 23-24, 1996. "Object Recogni-tion, Especially Face Recognition, with and without Feedback."

Machines that Learn, Snowbird, Utah, April 9-12, 1996. "A New Image Representation for Visual Learning."

ATR Human Information Processing Research Laboratories Symposium on Face and Object Recognition, Kyoto, Japan, January 22-25, 1996. "Virtual Examples and Example-based Correspondence."

MIT, Center for Imaging Science, Woods-Hole Workshop, May 23-24, 1996. "Object Recognition, Especially Face Recognition, with and without Feedback."

Machines that Learn, Snowbird, Utah, April 9-12, 1996. "A New Image Representation for Visual Learning."

ATR Human Information Processing Research Laboratories Symposium on Face and Object Recognition, Kyoto, Japan, January 22-25, 1996. "Virtual Examples and Example-based Correspondence."

Pawan Sinha:

Brown University, Sensation and Perception Seminar Series, February, 1996 "The Perception of Motion and Form."

NEC Research Institute, Princeton, March, 1996. "The Perception of 3D Structures in 2D Images."

STUDENTS

Graduate Students:

David Beymer, Electrical Engineering & Computer Science and Artificial Intelligence Laboratory

Emanuela Bricolo, Department of Brain & Cognitive Sciences and Center for Biological and Computational Learning

Nicholas Chan, Electrical Engineering & Computer Science and Center for Biological and Computational Learning

Theodoros Evgeniou, Electrical Engineering & Computer Science and Artificial Intelligence Laboratory

Tony Ezzat, Electrical Engineering & Computer Science and Artificial Intelligence Laboratory

Michael Jones, Electrical Engineering & Computer Science and Artificial Intelligence Laboratory

Stephen Lines, Electrical Engineering & Computer Science and Artificial Intelligence Laboratory

Edgar Osuna, Electrical Engineering & Computer Science and Operations Research

Constantine Papageorgiou, Electrical Engineering & Computer Science and Center for Biological and Computational Learning

Maximilian Riesenhuber, Department of Brain & Cognitive Sciences and Center for Biological and Computational Learning

Raquel Romano, Electrical Engineering & Computer Science and Artificial Intelligence Laboratory

Kah Kay Sung, Electrical Engineering & Computer Science and Artificial Intelligence Laboratory

Robert Thau, Brain & Cognitive Sciences and Artificial Intelligence Laboratory

Elaine Yiu, Electrical Engineering & Computer Science and Center for Biological and Computational Learning

Postdoctoral Fellows:

Michael Oren, Center for Biological & Computational Learning

Pawan Sinha, Center for Biological & Computational Learning

Partha Niyogi, Artificial Intelligence Laboratory