Learning and Vision Machines

BERND HEISELE, ALESSANDRO VERRI, AND TOMASO POGGIO

Invited Paper

The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial. In this paper, we review our approach to the problem of visual perception based on supervised learning. After a brief presentation of the theoretical background, we focus on some of the engineering applications of statistical learning to computer vision and discuss the main open problems and directions of our future research.

***Keywords*—Morphable models, object categorization, object detection, object recognition, pattern classification, support vector machines, visual learning.**

I. INTRODUCTION

Learning is a key for understanding intelligent systems. Because seeing is intelligence, learning is an essential feature in the study of visual systems from both the viewpoint of visual neuroscience and computer vision. While visual neuroscience concentrates on the mechanisms allowing the cortex to adapt its circuitry and learn a new task, computer vision aims at devising effectively trainable systems. Vision systems that *learn and adapt* represent one of the most important trends in computer vision research and may provide the only solution to the development of robust and reusable vision systems.

Manuscript received May 31, 2001; revised February 15, 2002. This work was supported by the Office of Naval Research under Contract N00014-93-1-3085, Office of Naval Research (DARPA) under Contract N00014-00-1-0907, the National Science Foundation (ITR) under Contract IIS-0085836, and by the National Science Foundation under Contract IIS-9800032 and in part by the Central Research Institute of Electric Power Industry, Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation, and The Whitaker Foundation.

B. Heisele was with the Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA 02142 USA. He is now with Honda R&D Americas Inc., Boston, MA 02111-1208 USA (e-mail: BHeisele@oh.hra.com).

A. Verri is with the Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA 02142 USA and also with the INFM, Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova, 16146 Genova, Italy (e-mail: verri@disi.unige.it).

T. Poggio is with the Mc Govern Institute, Center for Biological and Computational Learning, AI Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142 USA (e-mail: tp@ai.mit.edu).

Publisher Item Identifier 10.1109/JPROC.2002.801450.

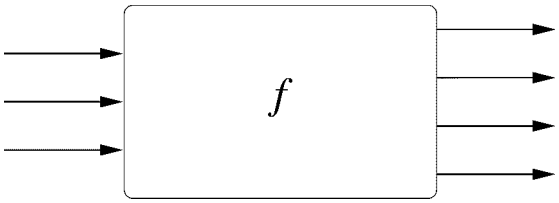

Fig. 1. In the learning-from-examples paradigm, we learn a function f from ℓ input–output pairs $(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, \ell$ called the training set. When the vectors $\mathbf{y}_i$ represent images, the system learns to perform image synthesis—i.e., a computer graphics task, depending on the control parameters $\mathbf{x}_i$. When images are the inputs $\mathbf{x}_i$, instead, the system learns to perform the inverse problem of image analysis, i.e., a computer vision task.

II. SUPERVISED LEARNING PARADIGM

In this paper, we review our approach over the last ten years in the area of supervised learning as applied to visual perception in machines. The research in this area is based on the mathematical foundations of learning and paralleled by neuroscience studies. The interactions between these research directions can be more complex than an ideal continuous loop from theory to feasibility demonstrations to biological models feeding back into new theoretical ideas. For instance, in 1990, ideas from the mathematics of learning theory—radial basis function networks—suggested a model [1] for biological object recognition. Psychophysical data [2] suggested and physiological experiments in cortex [3] found evidence for the neurons tuned to single views of a specific learned object that were predicted by the model. It was only later that the same general idea found its way in computer graphics applications that we will briefly describe in Section VII.

This paper discusses applications of supervised learning—or *learning-from-examples*—see Fig. 1. The problem of learning-from-examples can be set in a mathematically well-founded framework [4], [5] within probability theory, statistics, and function approximation. With a recent paper by Cucker and Smale [6], statistical learning is entering the core of mathematics.

A. Identification, Categorization, and Parameter Estimation

The recognition of visual objects is a perceptual and cognitive task fundamental to vision. It is performed effortlessly by the brain countless times every day while satisfying two essential requirements: invariance and specificity. In face recognition, to give an example of a three-dimensional (3-D) object, we can recognize a specific face among many, while being rather tolerant to changes in viewpoint, scale, illumination, and expression. We distinguish between two main tasks: identification and categorization. We use the term *categorization* (or detection) to designate *between-class* object classification and the term *identification* for *within class* object classification. A third task important to recognition is image analysis or *parameter estimation*: humans, for instance, are able to estimate facial expression, characteristics of gait, etc., with ease and precision. In this paper, we use word recognition for either of these tasks.

Our approach is to consider the recognition problem as a supervised learning problem. We start with a very simplified description. Fig. 1 shows the learning module. Its input is an image, its output is a discrete label for: 1) the class of the object in the image (e.g., is it a face?); 2) its individual identity (e.g., is it my friend's face?); or 3) a real number (e.g., estimating the degree of happiness in a face). The learning module is trained with a set of examples, which are a set of input–output pairs, or images previously labeled. In this setting, the distinction between identification and categorization is mostly semantic. In the case of categorization, the range of possible variations seems larger, since the system must generalize not only across different viewing conditions but also across different exemplars of the class (such as male/female, for example). The difficulty of the task, however, does not depend on identification versus categorization but on parameters such as the size and composition of the training set and how much of the variability required for generalization is covered by the training examples. For instance, the simple learning module just described would be incapable of identifying an individual face from any viewpoint if trained with only a single view of that face. Conversely, the same module, if trained with a large set of examples covering the relevant variability, may become capable of performing the task.

The class of learning methods we used in our applications includes classification and regression techniques such as regularization networks (RNs) and support vector machines (SVMs). SVMs for classification correspond to systems that are trained to recognize objects, such as faces, for example—either for identification or categorization—in images and videos. SVMs for regression and RNs, instead, may estimate continuous variables from images, such as parameters of facial expressions.

B. Plan of the Paper

In the Section III, we will provide a very brief overview of statistical learning theory. The main framework is regularization theory: SVMs are special cases of regularization methods for appropriate choices of the loss function. We will then discuss our work on the problems of object categorization, object identification, and image analysis. We have trained our systems with different types of objects, but in this paper we will mainly focus on faces throughout the three types of recognition: categorization, identification, and image analysis. SVM classifiers are the core engine in the trainable vision systems described in Section IV for categorization. The system for face identification of Section V is based on a hierarchy of SVM classifiers. The estimation methods described in Section VI, instead, rely on SVM for regression and also on morphable models. The final section (Section VII) will consider other applications of the learning approach to visual perception and directions of future research including image synthesis for graphics and extensions to time with action recognition. We will conclude arguing for the possibility to perform scene understanding by using a dictionary of classifiers trained to categorize different classes of objects.

III. Theoretical Background

This section provides a very brief account of statistical learning theory (SLT) [7], [4]. Following [5], we take the viewpoint of Regularization Theory [8]. In supervised learning, a machine chooses a function which best describes the relation between the inputs and the outputs. The central question of SLT is how well the chosen function generalizes on previously unseen inputs.

A. Regularized Solutions

We are interested in learning schemes leading to solution of the form

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \tag{1}$$

where the $\mathbf{x}_i, i = 1, \ldots, \ell$ are the input examples, K is a certain symmetric positive definite function named kernel, and α_i is a set of parameters to be determined form the examples. A solution f as in (1) is an example of a *regularized* solution and is found as the minimizer of functionals of the type

$$\Phi[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_K^2 \tag{2}$$

where V is a *loss function* which measures the goodness of the predicted output $f(\mathbf{x}_i)$ with respect to the given output $y_i, \|f\|_K^2$ a smoothness, or *regularizing*, term which is the norm in the reproducing Kernel Hilbert space (RKHS) defined by the kernel K, and λ is a positive parameter controlling the relative weight between the data and the regularizing term. For a fixed λ, the regularized solution can be thought of as the function of minimum RKHS norm approximating the data within some degree of accuracy. The choice of the loss function determines different learning techniques, each leading to a different learning algorithm for computing the coefficients α_i in (1).

B. Statistical Learning Theory

We consider two sets of random variables $\mathbf{x} \in X$ and $y \in Y$ related by a probabilistic relationship. Though not necessary in the following, we restrict the analysis to the case in whinch $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}$. We are provided with *examples* of this probabilistic relationship, that is, with a data set $D_\ell \equiv \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^{\ell}$ called *training set*, obtained by sampling ℓ times the set $X \times Y$ according to $p(\mathbf{x}, y)$. Given the data set D_ℓ, the "problem of learning" consists of providing an *estimator*, or a function $f : X \to Y$, that can be used to predict a value y for each $\mathbf{x} \in X$. In vision, X could be the set of all possible images, Y the set $\{-1, 1\}$, and $f(\mathbf{x})$ an *indicator function* which specifies whether image $\mathbf{x}$ contains a certain object $(y = 1)$, or not $(y = -1)$ (see, for example, [9]).

The fundamental problem of SLT is to find the estimator minimizing a measure of the average amount of error, the so-called *expected risk*, defined as

$$I[f] \equiv \int_{X,Y} V(y, f(\mathbf{x}))p(\mathbf{x}, y)\, d\mathbf{x}\, dy.$$

Typically, f_0, the minimizer of $I[f]$ also called *target* function, belongs to some large space $\mathcal{F}$ and cannot be found in practice, because the probability distribution $p(\mathbf{x}, y)$ is unknown, and only a sample of it, the data set D_ℓ, is available. A popular *induction principle* which can be put into practice for finding an approximation of the minimizer of the expected risk in the presence of finite number of examples is the *empirical risk minimization* (ERM) induction principle [4]. The ERM principle consists of using the data set D_ℓ to build a stochastic approximation of the expected risk, which is usually called the *empirical risk*, defined as

$$I_{\text{emp}}[f; \ell] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)).$$

Straight minimization of the empirical risk in $\mathcal{F}$ can be problematic. First, it is an *ill-posed* problem [8] in the sense that there are usually many functions, possibly infinitely many functions, minimizing the empirical risk. Second, even assuming that a unique minimizer for I_{emp}, say $\hat{f}$, can be found, ERM can lead to *overfitting*. This means that $I_{\text{emp}}[\hat{f}; \ell]$ can be very close to zero, while $I[\hat{f}]$, the expected risk of $\hat{f}$, can be much larger than the optimal expected risk $I[f_0]$.

SLT bounds the distance between the empirical and expected risk of *any* function with inequalities of the type

$$I[f] < I_{\text{emp}}[f] + \varphi\left(\sqrt{\frac{h}{\ell}}, \eta\right) \tag{3}$$

where φ is an increasing function of h/ℓ and η and the bound holds true with a probability of at least η. The quantity h in inequality (3) measures the "capacity" of the space in which the empirical risk is minimized, named *hypothesis space*. Appropriate capacity measures are defined in the theory, the most popular one being the VC-dimension [10] or scale-sensitive versions of it [11], [12]. For more details and examples of exact forms of φ, we refer the reader to [10], [4], and [12]. Intuitively, if the capacity of the hypothesis space is very large and the number of examples is small, then the distance between the empirical and expected risk can be large and overfitting is very likely to occur.

Inequality (3) suggests an alternative method for achieving good generalization: instead of minimizing the empirical risk, find the best tradeoff between the empirical risk and the *complexity of the hypothesis space* measured by the second term in the right-hand side (r.h.s.) of inequality (3). This observation leads to the principle of *structural risk minimization* (SRM). The idea of the SRM principle is to define a nested sequence of hypothesis spaces $H_1 \subset H_2 \subset \cdots \subset H_M$ of increasing capacity, minimize the empirical risk in each hypothesis space, and choose, among the solutions found, the one with the best tradeoff between the empirical risk and the capacity as given by the r.h.s. of inequality (3). The statistical properties of SRM are described at length in [13] and [4].

C. Learning as Functional Minimization

We now consider hypothesis spaces which are subsets of an RKHS [14]. An RKHS is a Hilbert space of functions f of the form $f(\mathbf{x}) = \sum_{n=1}^{N} a_n \phi_n(\mathbf{x})$, where $\{\phi_n(\mathbf{x})\}_{n=1}^{N}$ is a set of given, linearly independent basis functions, and N is not necessarily finite. An RKHS is equipped with a norm which is defined as

$$\|f\|_K^2 = \sum_{n=1}^{N} \frac{a_n^2}{\lambda_n}$$

where $\{\lambda_n\}_{n=1}^{N}$ is a decreasing sequence of strictly positive real numbers whose sum is finite. The λ_n and the basis functions $\{\phi_n\}_{n=1}^{N}$ define the symmetric positive definite kernel function

$$K(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{N} \lambda_n \phi_n(\mathbf{x}) \phi_n(\mathbf{y}).$$

A nested sequence of spaces of functions in the RKHS can be constructed by bounding the RKHS norm of functions in the space. This can be done by defining a set of constants $A_1 < A_2 < \cdots < A_M$ and considering spaces of the form

$$H_m = \{f \in \text{RKHS} : \|f\|_K \leq A_m\}.$$

It can be shown that the capacity of the hypothesis spaces H_m is an increasing function of A_m (see, for example, [5]). The solution of the learning problem is found by searching for the minimum of functionals like the functional $\Phi[f]$ in (2). To establish a connection with SRM, the key issue is the choice of the optimal m^* identifying the hypothesis space in which the structural risk is minimized or the choice of the optimal value for the regularization parameter λ in (2). These two problems, as discussed in [5], are related, and the SRM method can in principle be used to choose the optimal λ [4]. In practice, instead of using SRM, other methods are used such as cross validation [14], generalized cross validation, finite prediction error, and the MDL criteria (see [4] for a review and comparison).

An important feature of the minimizer of $\Phi[f]$ is that for a broad range of loss functions the minimizer has the same general form of (1) [14]. Notice that in (1) the function f is expressed as as a linear combination of kernels centered in each data point. Using different kernels we get functions such as Gaussian radial basis functions ($K(\mathbf{x},\mathbf{y}) = \exp(-\beta\|\mathbf{x}-\mathbf{y}\|^2)$), or polynomials of degree d ($K(\mathbf{x},\mathbf{y}) = (1+\mathbf{x}\cdot\mathbf{y})^d$) [15], [4].

We now turn to discuss a few learning techniques based on the minimization of functionals of the form (2) by specifying the loss function V.

D. Regularization Networks

The approximation scheme of RNs arises from the minimization of the functional in (2) with the quadratic loss function V defined as

$$V(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2.$$

For RN, it is possible to show (see, for example, [15]) that the coefficients α_i in (1) of the minimizer of (2) satisfy the following linear system of equations:

$$(K + \lambda I)\boldsymbol{\alpha} = \mathbf{y}$$

where I is the identity matrix, $\mathbf{y} = (y_1, \ldots, y_\ell), \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)$, and $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Since the coefficients α_i satisfy a linear system, the function f can be rewritten as

$$f(\mathbf{x}) = \sum_{j=1}^{\ell} y_j b_j(\mathbf{x}) \tag{4}$$

with $b_j(\mathbf{x}) = \sum_{i=1}^{\ell}(K+\lambda I)^{-1}_{ij} K(\mathbf{x}_i, \mathbf{x})$. Equation (4) gives the dual representation of RN. Notice the difference between (1) and (4): in the first, the coefficients α_i are learned from the data while in the second one the basis functions b_i are learned, the coefficient of the expansion being equal to the output of the examples. We refer the reader to [15] for more details on the dual representation.

E. Support Vector Machines

We now discuss SVM [16], [4]. We distinguish between real output (regression) and binary output (classification) problems. The method of SVM regression is obtained by defining the loss function V as

$$V(y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\epsilon \tag{5}$$

where the function $|\cdot|_\epsilon$, called ϵ-insensitive loss, is defined as

$$|t|_\epsilon \equiv \begin{cases} 0, & \text{if } |t| < \epsilon \\ |t| - \epsilon, & \text{otherwise.} \end{cases}$$

The method of SVM for classification, instead, is obtained through the loss function V

$$V(y, f(\mathbf{x})) = (1 - yf(\mathbf{x}))_+ \tag{6}$$

where $(t)_+ = t$ if $t > 0$ and zero otherwise.

In both cases, it turns out that the coefficients α_i can be found by solving a quadratic programming (QP) problem with linear constraints. The regularization parameter λ appears only in the linear constraints, and for each coefficient α_i we have $0 \le \alpha_i \le 1/(2\lambda\ell)$. The QP problem is nontrivial since the size of the matrix of the quadratic form is equal to $\ell \times \ell$ and the matrix is dense. A number of algorithms for training SVM have been proposed: some are based on a decomposition approach where the QP problem is attacked by solving a sequence of smaller QP problems [17], and others are based on sequential updates of the solution [18].

A remarkable property of SVMs is that loss functions (5) and (6) lead to *sparse* solutions. This means that, unlike in the case of RNs, typically only a small fraction of the coefficients α_i in the expansion (1) are nonzero. The data points $\mathbf{x}_i$ associated with the nonzero α_i are called *support vectors*. If all data points which are not support vectors were to be discarded from the training set, the same solution would be found. In this context, an interesting perspective on SVM is to consider its information compression properties. The support vectors represent the most informative data points and compress the information contained in the training set: for the purpose of, say, classification, only the support vectors need to be stored, while all other training examples can be discarded.

In classification, the inverse of the RKHS norm of the solution has a very interesting geometrical interpretation. The inverse of the RKHS norm equals the *margin* [4], a quantity measuring the distance of the closest point in the training set from the separating surface. Therefore, an SVM looks for a separating surface which leaves the closest point as far as possible by controlling the norm (i.e., the smoothness) of the solution.

Of the various bounds which can be computed for the generalization performance of SVMs (see [4], for example), we consider a *leave-one-out* bound. An almost unbiased upper bound L on the expected risk of an SVM trained on ℓ data points drawn according to a probability $p(\mathbf{x}, y)$ is given by [4]

$$L = \frac{1}{\ell} E\left[\min\left(m_\ell, \frac{r_\ell^2}{\rho_\ell^2}\right)\right] \tag{7}$$

where $E[\cdot]$ denotes the expectation over the probability $p(\mathbf{x}, y), m_\ell$ the number of support vectors, r_ℓ is the radius of the smallest sphere containing the support vectors, and ρ_ℓ is the margin of the SVM trained on ℓ data points.

IV. Object Detection and Categorization

A major task in visual scene analysis is to detect objects in images. A possible way to perform this task is to shift a search window over an input image and categorize the object in the window with a classifier. A main problem with this approach is the large range of possible variations within a class of objects. As already mentioned, the classifier must generalize not only across different viewing and illumination conditions but also across different exemplars of a class, such as faces of different people for face detection. To simplify the categorization task, most vision systems use sets of binary classifiers, one for each object category. In this section, we only consider the binary categorization task where the classifier has to separate objects of one particular class (e.g., faces) from all other objects.

Fig. 2. Results of detecting people using Haar wavelets as inputs to an SVM classifier [24].

Fig. 3. Matching with a single template. (a)The schematic template of a frontal face is shown. Slight rotations of the face in (b) the image plane and (c) in depth lead to considerable discrepancies between template and face.

A. Global Approach

Face detection is an active area of research (see [19]–[23], for example). Over the last few years, we developed an example-based learning system for object detection in the context of face, people, and car detection [24]. The system used Haar wavelet features as inputs to an SVM classifier. In the training step, the system took a set of aligned and normalized images of the object class (positive examples) and a set of patterns that were not in the object class (negative examples). We calculated the Haar wavelet features for each pattern and then trained the SVM to distinguish between positive and negative examples. In the testing phase, the system slid a fixed size window over an input image and used the trained classifier to decide which patterns show the objects of interest. At each window position, we extracted the same set of Haar wavelet features as in the training step and fed them into our classifier; the classifier output determined whether or not that pattern contained an in-class object. To achieve multiscale detection, we iteratively resized the image and processed each image size using the same fixed size window. Results of our system for people detection are shown in Fig. 2.

B. Component Based Approach

In the above system, the whole object was represented by one feature vector which was fed into a single classifier. This global approach proved to work well for detecting objects under fixed viewing conditions. However, problems occur when the viewpoint and the pose of the objects vary, especially when the training set does not cover all the viewing variations in the test set. This is illustrated in Fig. 3 for a face detection system that is trained on frontal, upright faces and tested on rotated faces. The result of training a linear classifier on frontal faces can be represented as a single face template, schematically drawn in Fig. 3(a). Even for small rotations, the template clearly deviates from the rotated faces as shown in Fig. 3(b) and (c). In order to overcome this problem, we developed a component-based approach [25] where the object is decomposed into a set of components that are interconnected by a flexible geometrical model. While their relative positions change, each component varies less under pose changes than the pattern belonging to the whole object. Fig. 4 illustrates the component-based idea for face detection. The face is decomposed into a set of components: left and right eye, nose, and mouth. Training a linear classifier on each component results in a set of component templates which are schematically drawn in Fig. 4(a). For small rotations, the changes in the components are small. Since the geometrical model is flexible, we can slightly shift the components in order to achieve a better match with the rotated faces [see Fig. 4(b) and (c)].

Fig. 4. Matching with a set of component templates. (a) The schematic component templates for a frontal face are shown. Shifting the component templates can compensate for slight rotations of the face in (b) the image plane and in (c) depth.

The two main issues arising in the implementation of this component-based approach are how to include information about the geometrical relation between components into the classification process and how to choose a set of relevant components. We now discuss how we deal with these two issues in some detail and then illustrate some experimental results.

C. Geometrical Classifier

We developed a two-level classification system for face detection that implies geometrical relations between components similar to the one proposed in [26] for people detection. An overview on the system is shown in Fig. 5. On the first level, component classifiers independently detect components of the face. We used linear SVM classifiers, each of which was trained on a set of extracted facial components and on a set of randomly selected nonface patterns. The components were automatically extracted from synthetic face images generated from 3-D head models in which pointwise correspondences were known [27]. On the second level, the classifier checks if the geometrical configuration of the detected components corresponds with the learned geometrical model of a face. The inputs to the geometrical configuration classifier are the maximum SVM outputs of the component classifiers within rectangular search regions. They have been calculated from the mean and standard deviation of the components' locations in the training images. We also provide the geometrical classifier with the positions of the detected components relative to the upper left corner of the search window.

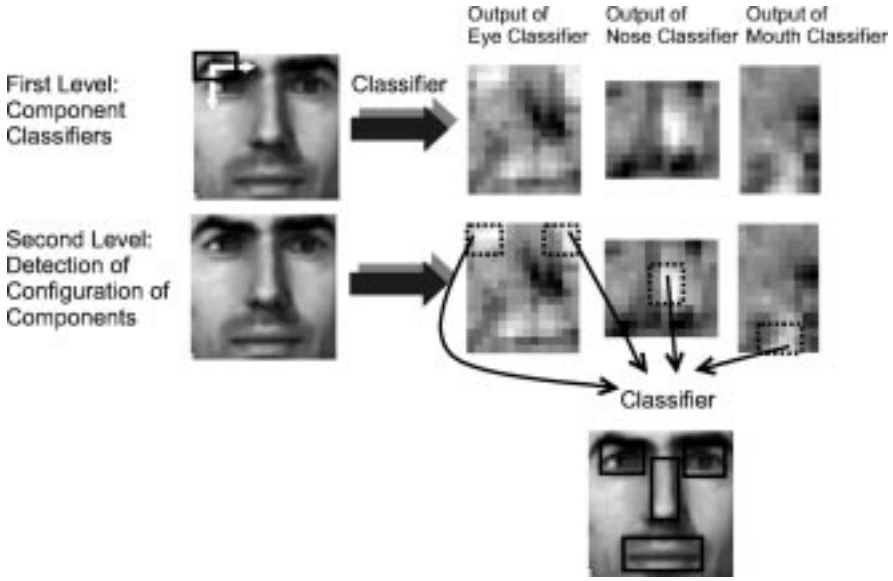

Fig. 5. System overview of the component-based classifier using four components. On the first level, windows of the size of the components (solid lined boxes) are shifted over the face image and classified by the component classifiers. On the second level, the maximum outputs of the component classifiers within predefined search regions (dotted lined boxes) and the positions of the components are fed into the geometrical configuration classifier.

D. Learning Components

As mentioned above, a second main issue of this component-based approach is how to choose the set of discriminatory object parts. For the class of faces, an obvious choice of components would be the eyes, the nose, and the mouth. For other classes, it might be more difficult to define a set of intuitively meaningful components. Instead of manually choosing the components it would be more sensible to choose the components automatically based on their discriminative power and their robustness against pose and illumination changes. Training a large number of classifiers on components of random size and location is one way to approach the problem of automatically determining components. The components can be ranked and selected based on the training results of the classifiers, e.g., the bound on the expected error probability. An alternative to using large sets of arbitrary components is to specifically generate discriminative components. Following this idea, we developed a method that automatically determines rectangular components from a set of synthetic face images. The algorithm starts with a small rectangular component located around a preselected point in the face (e.g., center of the left eye). Since the point-by-point correspondences between the 3-D head models are known, we can locate the same facial point in all face images. The component is extracted from all synthetic face images to build a training set of positive examples. We also generate a training set of nonface patterns that have the same rectangular shape as the component. After training an SVM on the component data, we determine the performance of the SVM based on a rough estimate $\tilde{L}$ of the upper bound L in (7) given by

$$\tilde{L} = \frac{1}{\ell}\frac{R^2}{\rho^2} \tag{8}$$

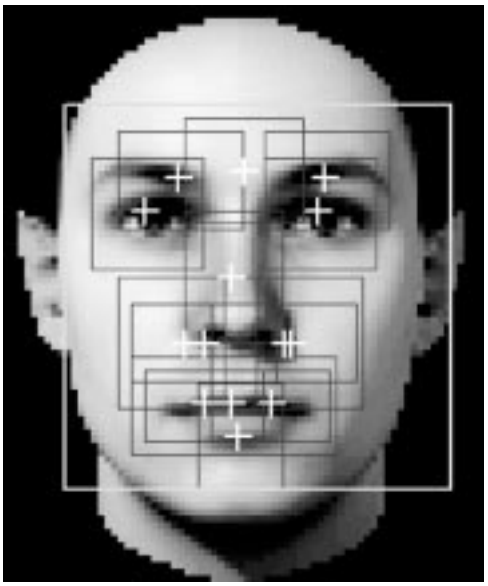

Fig. 6. The 14 learned components. The crosses mark the centers of the components.

where ℓ is the number of training patterns, R is the diameter of the smallest sphere containing the data points in the training set—and not just the support vectors as in (7)—and ρ is the margin of the classifier. After determining $\tilde{L}$, we enlarge the component by expanding the rectangle by one pixel into one of the four directions (up, down, left, right). Again, we generate training data, train an SVM, and determine $\tilde{L}$. We do this for expansions into all four directions and finally keep the expansion which decreases $\tilde{L}$ the most. This process is continued until the expansions into all four directions lead to an increase of $\tilde{L}$. In our experiments, we started with 14 seed regions located in the vicinity of the eyes, nose, and mouth. Fig. 6 shows the results obtained after the component growing stage.

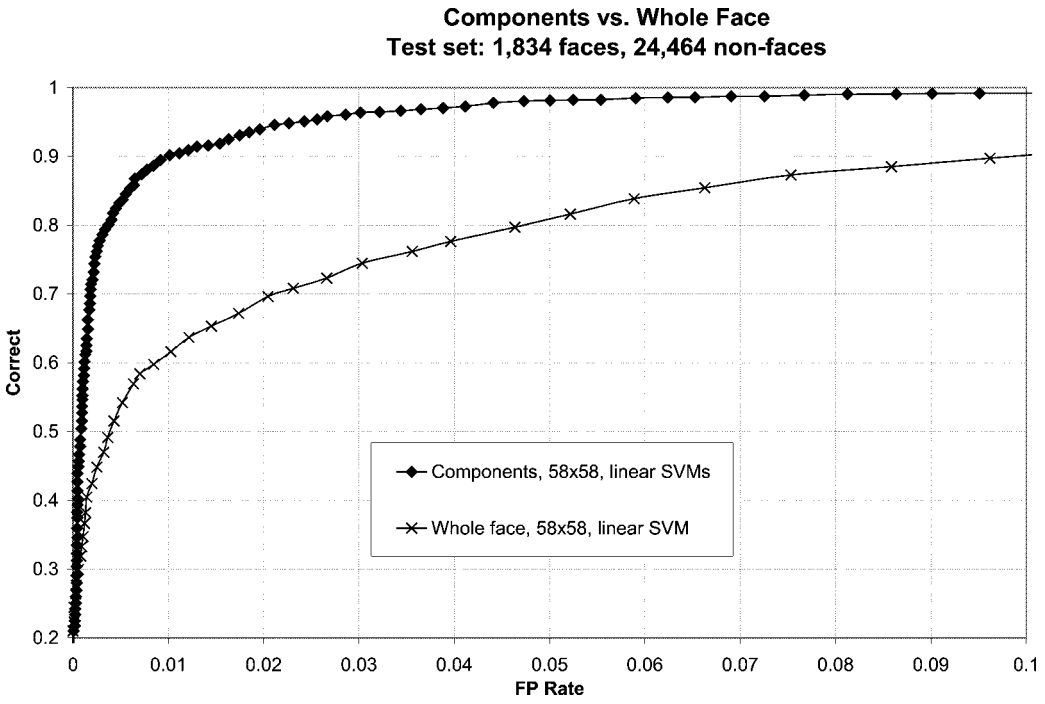

Fig. 7. ROC curves for a linear SVM whole face classifier and a component classifier consisting of 14 linear SVM component classifiers and a linear SVM geometrical configuration classifier. The false-positives (FPs) are given relative to the number of nonface images.

E. Experiments

In our experiments, we compared the component-based system to a classifier trained on the whole face pattern. The component system consisted of 14 linear SVM classifiers for detecting the components and a single linear SVM as a geometrical configuration classifier. The training data for both classifiers included 3590 synthetic gray face images and 13 655 nonface gray images 58×58 pixels in size. The positive test set consisted of 1154 gray images of real faces, the resolution varying between about 40×40 and 90×90 pixels. It included faces rotated between about $-45°$ and $45°$ in depth. The negative test set consisted of 24 464 randomly collected nonface patterns 58×58 in size. The comparison between a linear SVM whole face classifier and the component-based system with 14 component classifiers and a linear geometrical configuration classifier is shown in Fig. 7. The component system clearly outperforms the whole face systems. Some detection results generated by the component system are shown in Fig. 8.

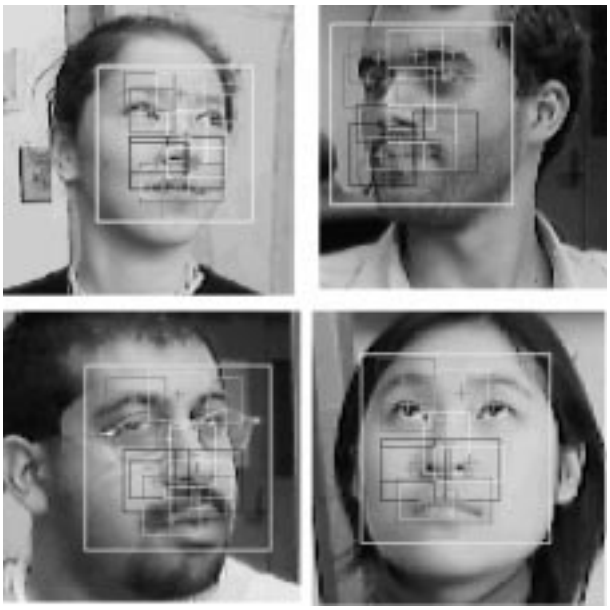

Fig. 8. Faces detected by the 14 component system. The lower right image is from the CMU test set [21].

V. Object Identification

The task of object identification is to distinguish between exemplars of the same class. The main difficulty of this problem is that objects belonging to the same class might differ in details only. In contrast to object detection and categorization, we constrain the test images to be new images of only the objects used in training, e.g., only new face images of the same people used for training the face identification system. Before describing our work on object identification, let us briefly discuss the issue of multiclass classification.

A. Multiclass Classification With Binary Classifiers

The number q of classes of a typical problem of object identification can be very large. Given the binary nature of SVM, in order to obtain a unique classification, the adoption of some combination strategy of a number of SVMs is required. Currently, we have considered two basic strategies. The first is a one-versus-one approach in which $q(q-1)/2$ machines are trained. Each SVM separates a pair of classes and the pairwise classifiers are arranged in trees, where each

tree node represents an SVM. The second is a one-versus-all approach, in which q SVMs are trained. Each of the SVMs separates a single class from all remaining classes [16], [29].

A bottom-up tree similar to the elimination tree used in tennis tournaments, for example, has been used in 3-D objects recognition [30] and face recognition in [31]. A top-down tree structure has been recently studied in [32]. Regarding the training effort, the one-versus-all approach is preferable since only q SVMs have to be trained compared to $q(q-1)/2$ SVMs in the pairwise approach. In the presence of a relatively small number of examples and a large number of classes, the pairwise approach has the advantage of keeping a better balance between the size of the training examples of the two classes, avoiding bias in the SVM performance. The run-time complexity of both strategies is similar: the pairwise approach requires the evaluation of $q-1$ SVMs, the one-versus-all approach that of q SVMs. Recent experiments on person recognition show similar classification performances for the two strategies [33].

B. Previous Work

Similarly to object categorization, one of the most important problems for object identification is pose invariance. In [30], we studied a simple appearance-based system for 3-D object identification. The system was assessed on the Columbia Object Image Library (COIL) database, that is, on 7200 images of 100 objects (72 views for each of the 100 objects). As explained in [34], the COIL objects were positioned in the center of a turntable and observed from a fixed viewpoint. For each object, the turntable was rotated 5° per image. In our experiments, the color images were transformed into gray-level images of reduced size (32×32 pixels) by averaging the original images over 4×4 pixel patches. The training set used in each experiment consists of 36 images (one every 10°) for each object. The remaining 36 images for each object were used for testing.

Given the relatively small number of examples and the large number of classes, we used the one-versus-one approach to multiclass classification. Given a subset of 32 of the 100 objects, a linear SVM was trained on each pair of objects using the gray values at each pixel as components of a 32×32 input vector. Recognition, without pose estimation, was then performed on a one-versus-one basis. The filtering effect of the averaging preprocessing step was sufficient to induce very high recognition rates even in the presence of large amount of random noise added to the test set and for a slight shift of the object location in the image.

A major limitation of the above system was due to the global structure of the classifier, ill-suited to deal with occlusions and clutter. Once again, in order to circumvent this problem, one could look at approaches based on object components. In [35], we performed face recognition by independently matching templates of three facial regions (both eyes, nose, and mouth). The configuration of the components during classification was unconstrained since the system did not include a geometrical model of the face. A similar approach with an additional alignment stage was proposed in [36]. In the following, we describe some of our current work in this area concentrating on face identification which is one of the most relevant identification tasks in computer vision with respect to commercial applications.

C. Face Identification

Face identification is a classic problem of computer vision. Various approaches have been proposed including eigenfaces [37], [38], linear discriminant analysis [39], hidden Markov models [40], elastic graph matching [41], and, more recently, SVM [31], [42]. Following the idea of component-based face detection, we built a face identification system [43] that is more robust against pose changes than common systems using global approaches. We extracted facial components from the face image using the component-based detector described in the previous section. The components were normalized in size, combined into a single feature vector, and fed into the identification classifier. As in the global approach, we ended up with one feature vector as input to the identification classifier. However, in the component-based approach, each feature was attached to a location in the face (e.g., the mouth left corner) rather than to a fixed x–y location in the image.

Since in face recognition both the number of examples and the number of classes can be rather large, we opted for combining SVMs according to the one-versus-all strategy. We trained a linear SVM for every person in the database. Each SVM was trained to distinguish between all images of a single person $(y = 1)$ and all other images in the training set $(y = -1)$. Given q people and q corresponding SVMs, each one associated with one person, the class label y of a face pattern $\mathbf{x}$ is computed as follows:

$$y = \begin{cases} n, & \text{if } d_n(\mathbf{x}) + t > 0 \\ 0, & \text{if } d_n(\mathbf{x}) + t \leq 0 \end{cases}$$
$$\text{with } d_n(\mathbf{x}) = \arg \max_{1 \leq i \leq q} \{d_i(\mathbf{x})\} \tag{9}$$

where $d_i(\mathbf{x})$ is the distance of $\mathbf{x}$ from the hyperplane of the SVM which was trained to identify person i. The classification threshold is denoted as t. We added the rejection class denoted by class label 0 for cases where $\mathbf{x}$ is too close to the decision surface to perform a reliable identification.

We performed experiments on a database including images of five different people. The training set consisted of 8593 gray face images. The resolution of the face images originally ranged between 80×80 and 130×130 pixels with rotations in depth up to about $\pm 40°$. The test set included 974 images of the same five subjects. The rotations in depth was again up to about $\pm 40°$.

In order to evaluate the component-based system, we compared it to a global system. For the global system, we simply normalized the extracted face images in size and converted the gray values into a feature vector. The feature vector was then classified by a system of five linear SVM classifiers. For the component-based face recognizer, we first ran the component-based detector over each image in the training set and extracted the components. From the 14 original components, we kept only 10, removing those that strongly overlapped with other components. To generate the input to our

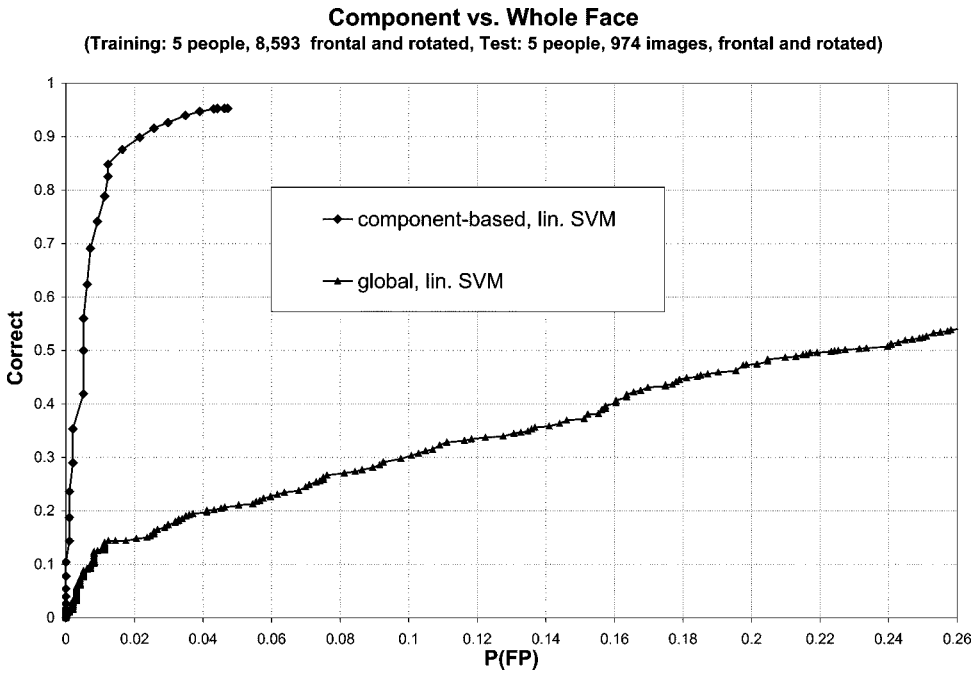

Fig. 9. ROC curves for face identification for the component-based system and the global system.

face recognition classifier, we normalized each of the components in size and combined their grey values into a single feature vector which was fed into five linear SVM classifiers. Fig. 9 shows the ROC curves for the two systems. Each point on the curve corresponds to a different value of the classification threshold t in (9). At the end points of the ROC curves the rejection rate is 0.

Fig. 9 clearly shows that the component system outperforms the whole face system. Some results of the component-based recognition system are shown in Fig. 10.

VI. Image Analysis

Learning techniques can also be used for addressing problems of image analysis, requiring the estimation of perceptually meaningful parameters. In the case of faces, for example, these parameters may be associated with mouth shapes and expressions. Following [44], in [45] we constructed a linear morphable model from examples of mouths through the use of optical flow and texture information. Mouths pose difficult problems due to the large visual difference between closed and open mouths. About 2000 images of mouths from one person were collected and 93 prototypes manually chosen to construct the morphable model. Pixelwise correspondences are computed between each prototype image and a standard reference image. The flow vector consists of the displacements of each pixel in the reference image relative to its position in the prototype. The texture vector consists of the prototype backward warped to the reference image. The parameter set, which is still largely redundant (despite the lower dimensionality of the model with respect to the dimensionality of the image space), is reduced by performing PCA on the example texture and flows. Only the first top three components of both texture and flow are retained (see Fig. 11), and six SVMs for regression with a Gaussian kernel are then

Fig. 10. Examples of component-based face recognition. Images in the first two rows were properly identified by the system. The last two images were misclassified due to strong rotation and facial expression.

used to learn the nonlinear mapping from a sparse subset of Haar wavelet coefficients [46] and each of the six matching parameters of the morphable model (b_1, b_2, and b_3 for the texture, and c_1, c_2, and c_3 for the flow). An example of the linear morphable model at work is shown in Fig. 11.

The SVMs for regression are trained on a set generated by estimating the true matching parameters using the computationally intensive stochastic gradient descent algorithm described in [44]. The direct parameter estimation obtained

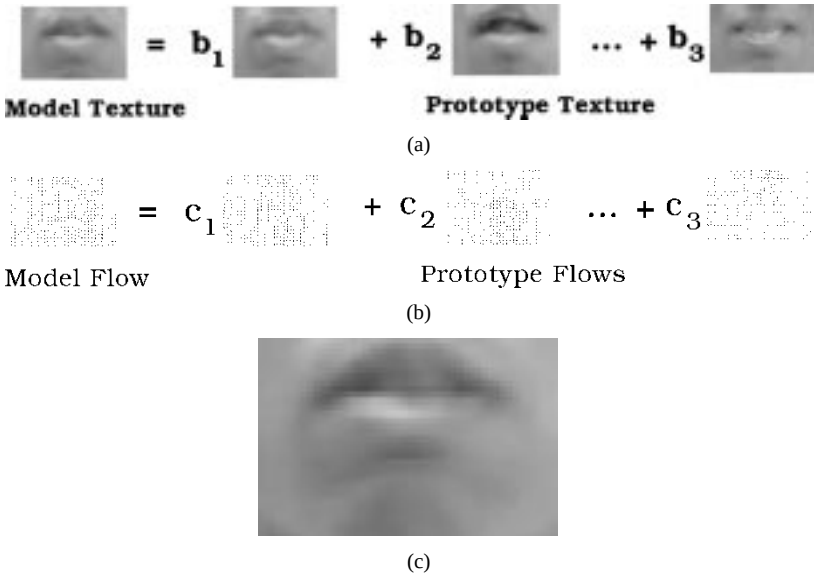

Fig. 11. LMM at work. (a) The model texture has a linear combination of texture prototypes. (b) The model flow has a linear combination of the flow prototypes. (c) The model image obtained by warping the model texture along the model flow.

by the SVMs makes possible a real-time implementation of the system bypassing the use of the stochastic gradient descent algorithm. In addition to computational efficiency, this strategy increases the stability of the estimation and the tolerance to outliers. This method bears some resemblance to the work by Cootes *et al.* [47] in which multivariate linear regression is used to learn the mapping between the error image (difference between the novel and the model image) and the appropriate perturbation of the model parameters to avoid the computation of gradients.

The obtained matching parameters can be used to solve recognition tasks like, for example, viseme (the visual analogs of phonemes) recognition [48]. Preliminary experiments in which linear SVMs are trained to recognize six different visemes from static images using the estimated parameters lead to significant improvements over more traditional classification techniques like k-*nearest-neighbors* [45]. So far, slightly better results have been obtained using wavelet representation as input to the SVM classifier but more extensive experiments are needed.

Linear morphable models are not restricted to two-dimensional (2-D) images (for an early description of the basic idea and mathematics, see [49]). Blanz and Vetter [27] successfully extended to 3-D the 2-D morphable models pioneered by [49]–[55], and [44]. The approach of Cootes *et al.* [47] has recently become essentially identical to these 2-D morphable models. Linear morphable models, as described in more detail in Section VII, can also be used to synthesize new images. A comparison of new mouth images with different expressions synthesized by using the gradient descent and support vector regression described before is shown in Fig. 12. As can be seen by inspection, the agreement with the novel images shown in the rightmost column is very good. An interesting open issue is how this scheme generalizes to images and mouths from different people.

VII. Discussion and Future Directions

We conclude here by briefly addressing a few issues that are relevant to the approach. They are either work already done or projects for the future.

A. Image Synthesis

The supervised learning paradigm outlined here can be applied to other domains as wel, beyond the area of vision. A closely related field is computer graphics. In analogy to the view-based paradigm for computer vision, we were led to the paradigm of image-based rendering which is just now becoming an important research direction in the graphics community. Consider the learning metaphor for vision: a system is trained to perform a visual task, such as estimating the pose of an object from images by training it with a set of example pairs, where the inputs are images and the desired estimates are the outputs. Imagine now training another system by interchanging the inputs with the outputs in the training set we described. If the system learns the task, then it should be capable of generating new images under user control without any 3-D physically based model and without simulating the physics of light as in traditional computer graphics. In other words, the system would do computer graphics in a somewhat unconventional way, in an image-based rendering paradigm.

Our first successful demonstration of using learning techniques for computer graphics tasks involved line drawing images of cartoon characters. The metaphor is that the computer learns to draw a cartoon character from a few example drawings provided by an artist [56]. We later extended the approach to deal with real images [52]. We are now developing an image-based text-to-visual-speech (TTVS) system [57]. The TTVS module takes as input unconstrained typed text and produces as output two synchronized streams: the audio stream and the video stream.

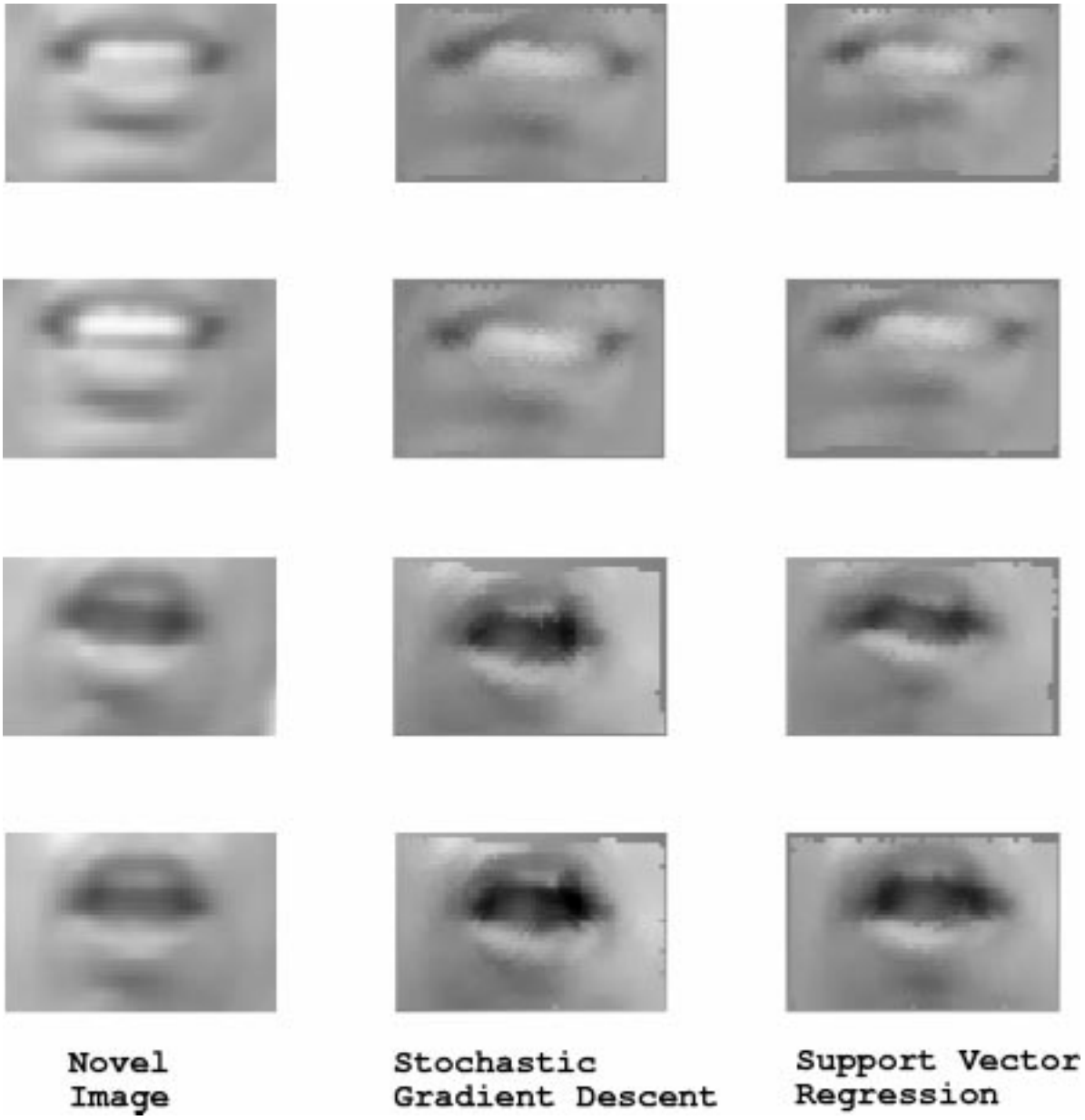

Fig. 12. Example matches of novel mouths using stochastic gradient descent and SVM regression from a test sequence.

The audio stream comprises the audible counterpart of the input sentence, while the video stream comprises the face animation and especially the visible mouth movements. Unlike other systems, our TTVS is based on synthesizing new video images from a small number (about 16) real viseme images, without using any 3-D or physical model of the face, thereby achieving high photorealism.

The above approach to image synthesis relies on the technique of morphable models which is related to regularization networks. A formal connection between classification/regression and the morphable models approach is provided by the dual representation of many regression and classification networks [15], [5]. The main difference between the two approaches is the explicit role of pixelwise correspondence in morphable models.

B. Adding the Temporal Dimension

The described framework can be extended over the time dimension in different ways. In [58], for example, we trained a system to recognize visual dynamic events, e.g., movements performed by different people, from a stream of images taken by a fixed camera. In this case, the system solves an object identification problem in which the object is the observed trajectory of a 3-D movement in the image plane. Unlike systems relying on quite sophisticated image processing and a complex description of the possible dynamic events (see [59]–[63], for example), this system does not attempt to characterize explicitly the properties of each event class but learns each class from examples.

Each event is represented by a feature vector built from the spatio-temporal changes detected in the observed image

sequence by means of a simple chain of low-level vision techniques. The system neither attempts to recover the 3-D structure nor assumes a prior model of the observed dynamic events. During training, a supervisor identifies and labels the events of interest among those automatically detected by the system. At run time, new events are detected and then classified. Given q dynamic event classes and a training set containing examples from each class, the system trains q polynomial SVMs in a one-versus-all fashion. The problem of varying time duration of the various events (either within class or between class events) is solved by a simple temporal normalization. Somewhat surprisingly, this simple procedure seems to be sufficient to deal with a large number of possible different events. The obtained results indicate that the system can be effectively trained from a rather small number of examples.

We explored a second interesting extension to time sequences in [64]. Extending morphable models in time, we have described videosequences—that is, synthesized or analyzed—as a linear combination of an appropriate number of prototypical example sequences. The feasibility of this approach has been demonstrated by synthesizing and analyzing biological motion (simulated and real).

C. Making Recognition More Efficient: Saliency and Attention

A key operation of the described detection systems is to scan a window across an image, through both position and scale, in order to analyze at each step a subimage. This subimage is then provided to a classifier that decides whether or not the object of interest is present. This is a computationally expensive strategy. A more efficient alternative, grounded in neurobiology, is for attention to focus on salient parts of the image and to have the output of the attentional spotlight provide the input to the object recognition module. In preliminary work [65], we have demonstrated the feasibility of exploiting saliency mechanism to speed-up the object detection system. Large speed-ups have been achieved, since evaluating the saliency of the image is far less computationally demanding than identifying objects.

In addition to merging bottom-up attention and object recognition, a future direction of research is to integrate top-down attention in our system using two mechanisms. If some attribute of the target is known in advance, for instance, that it is red or moves toward the viewer, then the saliency-computation can be biased appropriately.

Our final, integrated system should be able to rapidly identify interesting locations in the image, based on a plurality of rapidly adaptive top-down and bottom-up cues, and feed these to the object recognition module for identification.

D. A Dictionary for Classifiers for Scene Understanding?

The approach we described in part of this paper—training a classifier to detect a specific class of objects in the image—leads to the idea that ultimately a sufficiently complete dictionary of such classifiers may be used to attack the old problem of *scene interpretation*. We conjecture that the choice of which classifiers to run can be based upon the output of a relatively small number of these classifiers used to establish the contexual information about the observed scene.

ACKNOWLEDGMENT

The authors would like to thank V. Kumar for helpful discussions. This paper describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

REFERENCES

[1] T. Poggio and S. Edelman, "Network that learns to recognize 3-D objects," *Nature*, vol. 343, pp. 263–266, 1990.
[2] H. H. Bülthoff and S. Edelman, "Psychophysical support for a 2-D view interpolation theory of object recognition," in *Proc. Nat. Acad. Sci.*, vol. 89, 1992, pp. 60–64.
[3] N. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," *Current Biol.*, vol. 5, pp. 552–563, 1995.
[4] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[5] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Computat. Math.*, vol. 13, pp. 1–50, 2000.
[6] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2002.
[7] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.
[8] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC: Winston, 1977.
[9] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. Int. Conf. Computer Vision*, Bombay, India, 1998, pp. 555–562.
[10] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 17, pp. 264–280, 1971.
[11] M. Kearns and R. Shapire, "Efficient distribution-free learning of probabilistic concepts," *J. Comput. Syst. Sci.*, vol. 48, pp. 464–497, 1994.
[12] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," in *Symp. Foundations of Computer Science*, 1993.
[13] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.
[14] G. Wahba, *Splines Models for Observational Data*, ser. Series in Applied Mathematics. Philadelphia, PA: SIAM, 1990, vol. 59.
[15] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Computat.*, vol. 7, pp. 219–269, 1995.
[16] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learning*, vol. 20, pp. 1–25, 1995.
[17] E. Osuna, R. Freund, and F. Girosi, "An improved training algorithm for support vector machines," in *Proc. IEEE Workshop on Neural Networks and Signal Processing*, Amelia Island, FL, 1997, pp. 276–285.
[18] J. C. Platt, "Sequential minimal imization: A fast algorithm for training support vector machines," Microsoft Research, Tech. Rep. MST-TR-98-14, Apr. 1998.
[19] T. K. Leung, M. C. Burl, and P. Perona, "Finding faces in cluttered scenes using random labeled graph matching," in *Proc. Int. Conf. Computer Vision*, Cambridge, MA, 1995, pp. 637–644.

[20] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *Proc. IEEE 7th Int. Conf. Comput. Analysis of Images and Patterns (CAIP'97)*, pp. 456–463, 1997.
[21] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, 19987.
[22] T. D. Rikert, M. J. Jones, and P. Viola, "A cluster-based statistical model for object detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999.
[23] H. Schneiderman and T. Kanade, "A statistical method for 3-D object detection applied to faces and cars," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 746–751, 2000.
[24] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, pp. 15–33, 2000.
[25] B. Heisele, T. Serre, M. Pontil, and T. Poggio, "Component-based face detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 657–662, 2001.
[26] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 349–361, 2001.
[27] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3-D faces," in *Proc. SIGGRAPH*, 1999, pp. 187–194.
[28] H. A. Rowley, S. Baluja, and T. Kanade, "Rotation Invariant Neural Network-Based Face Detection,", Pittsburgh, PA, Computer Science Tech. Rep. CMU-CS-97-201, 1997.
[29] B. Schölkopf, C. Burges, and V. Vapnik, "Extracting support data for a given task," presented at the 1st Int. Conf. Knowledge Discovery and Data Mining, Menlo Park, CA, 1995.
[30] M. Pontil and A. Verri, "Support vector machines for 3D object recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 637–646, 1998.
[31] G. Guodong, S. Li, and C. Kapluk, "Face recognition by support vector machines," *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 196–201, 2000.
[32] J. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification," in *Advances in Neural Information Processing Systems*, vol. 12, 2000.
[33] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio, "People recognition in image sequences by supervised learning," MIT, AI Memo 1688, 2000.
[34] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-D object from appearance," *Int. J. Comput. Vis.*, vol. 14, pp. 5–24, 1995.
[35] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 1042–1052, 1993.
[36] D. J. Beymer, "Face recognition under varying pose," MIT, AI Memo 1461, 1993.
[37] L. Sirovitch and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Amer. A*, vol. 2, pp. 519–524, 1987.
[38] M. Turk and A. Pentland, "Face recognition using eigenfaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 586–591, 1991.
[39] P. Belhumeur, P. Hespanha, and D. Kriegman, "Eigenfaces vs fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 711–720, 1997.
[40] A. V. Nefian and M. H. Hayes, "An embedded HMM-based approach for face detection and recognition," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 19–24, 1999.
[41] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 775–779, 1997.
[42] K. Jonsson, J. Matas, J. Kittler, and Y. Li, "Learning support vectors for face verification and recognition," *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 208–213, 2000.
[43] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," in *Proc. 8th Int. Conf. Computer Vision*, Vancouver, BC, Canada, 2001, pp. 688–694.
[44] M. Jones and T. Poggio, "Multidimensional morphable models: A framework for representing and matching object classes," in *Proc. 6th Int. Conf. Computer Vision*, Bombay, India, 1998, pp. 683–688.
[45] V. Kumar and T. Poggio, "Learning Based Approach to Estimation of morphable model parameters," MIT, AI Memo 1696, 2000.
[46] ——, "Learning based approach to real time tracking and analysis of faces," presented at the Int. Conf. Automatic Face and Gesture Recognition, Grenoble, France, 2000.
[47] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. 5th Eur. Conf. Computer Vision*, 1998, pp. 484–498.
[48] T. Ezzat and T. Poggio, "Visual speech synthesis by morphyng visemes," *Int. J. Comput. Vis.*, vol. 38, pp. 45–57, 2000.
[49] T. Poggio and T. Vetter, "Recognition and structure from one 2-D model view: Observations on prototypes, object classes and symmetries," MIT, AI Memo 1347, 1992.
[50] T. Poggio and R. Brunelli, "A novel approach to graphics," MIT, AI Memo 1354, 1992.
[51] D. Beymer, A. Shashua, and T. Poggio, "Example-based image analysis and synthesis," MIT, AI Memo 1431, 1993.
[52] D. Beymer and T. Poggio, "Image representation for visual learning," *Science*, vol. 272, pp. 1905–1909, 1996.
[53] M. Jones, P. Sinha, T. Vetter, and T. Poggio, "Top-down learning of low-level vision tasks," *Current Biol.*, vol. 7, pp. 991–994, 1997.
[54] T. Vetter, M. Jones, and T. Poggio, "A bootstrapping algorithm for learning linear models of object classes," in *IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 40–46.
[55] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 733–742, 1997.
[56] S. Librande, *Example-Based Character Drawing*, Cambridge: M.S., Media Arts and Science Section, School of Architecture and Planning, Massachusetts Inst. Technol., 1992.
[57] T. Ezzat and T. Poggio, "MikeTalk: A talking facial display based on morphing visemes," in *Proc. Computer Animation Conf.*, Philadelphia, PA, June 1998, pp. 96–102.
[58] M. Pittore, M. Campani, and A. Verri, "Learning to recognize visual dynamic events from examples," *Int. J. Comput. Vis.*, vol. 38, pp. 35–44, 2000.
[59] C. Bregler, "Learning and recognizing human dynamics in video sequences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 568–574, 1997.
[60] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 780–785, 1997.
[61] I. Haritaoglu, D. Harwood, and L. Davis, "W^4S: A real time system for detecting and tracking people in 2.5 D," presented at the Proc. Eur. Conf. Computer Vision, Friburg, Germany, 1998, pp. 222–227.
[62] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schütte, and A. Wilson, "The KidsRoom: A perceptually-based interactive and immersive story environment," *PRESENCE: Teleoperators and Virtual Environments*, vol. 8, pp. 367–391, 1999.
[63] S. Intille, J. Davis, and A. Bobick, "Real-time closed-world tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 697–703, 1997.
[64] M. A. Giese and T. Poggio, "Morphable models for the analysis and synthesis of complex motion patterns," *Int. J. Comput. Vis.*, vol. 38, pp. 59–73, 2000.
[65] C. Papageorgiou, private communication.

Bernd Heisele received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Stuttgart, Stuttgart, Germany, in 1993 and 1999, respectively.

In 1999, he was awarded a Postd-Doctoral Fellowship by the DFG in Germany. From 1999 to 2001, he worked as a Post-Doctoral Researcher at the Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge. He subsequently joined Honda R&D America and is currently heading the Honda Research Laboratory, Cambridge, MA, where he is conducting research in computer vision. His research interests are learning-based object detection/recognition and motion analysis in image sequences.

Alessandro Verri received both the Laurea and the Ph.D. degrees in physics from the University of Genoa, Genoa, Italy, in 1984 and 1989, respectively.

Since 1989, he has been with the University of Genoa where he is a Professor in the Department of Computer and Information Science. He has been a Visiting Scientist and Professor at the Massachusetts Institute of Technology, Cambridge, INRIA, Rennes, France, ICSI, Berkeley, CA, and Heriot-Watt University, Edinburgh, U.K. He has published nearly 50 papers on stereopis, motion analysis in natural and machine vision systems, shape representation and recognition, pattern recognition, and 3-D object recognition. He is coauthor, with Dr. E. Trucco, of *Introductory Techniques for 3–D Computer Vision* (Englewood Cliffs, NJ: Prentice-Hall, 1998). Currently, he is interested in the mathematical and computational aspects of computer vision and statistical learning theory and he is leading a number of applied projects on the development of computer vision-based solution to industrial problems.

Dr. Verri has been and is on the Program Committee of the major international conferences in the area of image processing and computer vision and serves as referee of several leading journals in the field.

Tomaso Poggio received the doctorate degree in theoretical physics from the University of Genoa, Genoa, Italy, in 1970.

From 1971 to 1981, he held a tenured research position at the Max Planck Institute, Germany, after which he became a Professor at the Massachusetts Institute of Technology (MIT), Cambridge. Currently, he is the Uncas and Helen Whitaker Professor in the Department of Brain and Cognitive Sciences at MIT and a member of the McGovern Institute and of the Artificial Intelligence Laboratory. He is doing research in computational learning at the MIT Center for Biological and Computational Learning, which is his group. He has authored more than 2000 papers in areas ranging from psychophysics and biophysics to information processing in man and machine, artificial intelligence, machine vision, and learning. His main research activity at present is learning from the perspective of statistical learning theory, engineering applications, and neuroscience.

Dr. Poggio has received a number of distinguished international awards in the scientific community, is on the editorial board of a number of interdisciplinary journals, a fellow of the American Association for Artificial Intelligence as well as the American Academy of Art and Sciences, and an Honorary Associate of the Neuroscience Research Program at Rockfeller University.