A Real-Time Adaptive Image Classification System

ABSTRACT

In the past five years the new field of statistical learning has led to improved classification algorithms. For instance the trainable systems developed over the last 2 years in the area of object detection - such as the MIT system based on a Support Vector Machine classifier and the CMU system based on a naive Bayes approach, are the best so far for detecting and identifying objects in realistic images. However, they are still not as robust and flexible as the human visual system. Therefore, we propose to turn to neurobiology for cues. We plan to start from the classification systems we have developed at MIT and CMU and improve them using insights from the psychophysics of human vision, the neuroscience of cortex and from two very recent models of object recognition and attention in visual cortex developed at MIT and CalTech, respectively. Both models are consistent with multiple physiological and psychophysical data. The first one accounts for invariance and specificity of recognition, including robustness to clutter; the second one for the computation of salient image regions – exploiting image information that includes different color bands. Together they suggest a parallel, hierarchical architecture to improve our existing classifiers. We will extend these biological models and their machine implementations to deal with motion and temporal patterns with hyperspectral imagery. Thus we believe that there is now an interesting opportunity to design and develop an artificial system that leverages the processing principles of higher cortical areas. Potential applications of the associated software architecture include remote sensing and robotic sensing. We propose to:

- **Augment the simple classification stage of our biological model of object recognition with a more powerful mathematical abstraction: the Support Vector Machine (SVM) classifier that we have extensively used and tested in the context of ATR applications;**
- **Explore whether the hierarchical architecture suggested by biological models can improve our best classifiers for object detection and increase their adaptability and learning speed;**
- **From new insights in cortical mechanisms of learning develop and implement new techniques for the learning of object specific features to improve ATR performance;**
- **Following our model of attention, add biomimetic bottom-up and top-down attention-like signals to allow our classification systems to deal effectively with large multispectral imagery and to allow for on-line adaptability and learning;**
- **Explore neural models of computation in terms of parallel hardware architecture and implement novel on-line learning algorithms for SVM;**
- **Extend the static classification stage of the present HMAX model and our object detection computer systems to architecture capable of exploiting motion and hypespectral information and;**
- **Implement prototype software architectures designed for surveillance, ATR and robotic sensing applications and transfer technology through Boston-based Alphatech and Brook's Cog project at the Artificial Intelligence Laboratory at M.I.T. (for robotics sensing).**

A Technical Proposal

A.1 Introduction

A.1.1 New Capabilities for Automatic Interpretation of Multisensor Data

One of the main reasons that work like ours is urgently needed is the challenge presented by new sensors under development – such as high-resolution hyperspectral imaging reconnaissance sensors: How can we manage and interpret the vast flood of information that these sensors will provide? The key technologies needed are technologies for searching, detecting, classifying and extracting information from huge amounts of data in order to support reconnaissance and decision-making. These technologies involve shallow understanding of the contents of images and data sequences. The "understanding" does not need to be complete, but it does need to be sufficient to automatically detect and cue an operator to the presence of targets or relevant changes in the massive amounts of image data that is produced. Furthermore, it must be able to adapt to the rapidly changing structure of the signals and varying environmental conditions, such as weather, day/night, noise/clutter conditions, terrain, camouflage, and target statistics.

We believe that a set of recently developed statistical learning methods augmented by very recent insights in the neuroscience of the visual system will be one of the key conceptual tools to meet that challenge. These methods can be applied to model the probabilistic structure of data but much more importantly to develop *classifiers* that can extract parts of the data – for instance parts of an image – by assigning them correctly to meaningful classes. Commercial applications of these technologies will also be possible since they can be applied equally well to images, movies, and audio recordings as well as to scientific data objects (genetic sequence data, molecular structures, galactic spectra, and so on). Finally, a key advantage of learning methods is that they can be customized by fast training and executed continually.

A.1.2 Anticipated Outcome

Our general goal is to develop a new *technology* for the classification, organization, and search of imagery by harnessing new statistical learning theories and algorithms and insights from biology. In the process we expect to prototype key system components and to develop scientific insights. What can the DoD expect from our proposed research? The answer is twofold: first we will do basic research towards exploiting our increasing understanding of the brain together with advances in statistical learning theory to make smarter and more capable machines, ultimately better – for instance faster – than the human brain in highly complex reconnaissance tasks. Clearly, we are still a long way from understanding the brain or even the visual system at all its levels of complexity. But, recent and ongoing progress in the neuroscience of vision already allows us to improve some of the best existing algorithms for object detection. Thus the second outcome of our project will consist of better classification system for complex imagery and sequences of data from multispectral sensors.

Thus the anticipated payoff of this research will be

- to substantially improve and extend our existing learning algorithms and associated representations with insights from recent work in neuroscience for applications such as categorization of complex imagery and video;
- to design a computer vision version of the HMAX [50] model of cortex, embedding two significant extensions: on-line adaptability through top-down signals and task- and object-dependent learning of features;
- to add to HMAX the attentional capabilities provided by the CalTech/USC model of human attention to increase its invariance to large-range translations and its speed and robustness to clutter;

- to design a real-time prototype of the SVM classifier with on-line incremental learning—and to evaluate implementations in parallel neural-like computer architectures;
- to demonstrate a software prototype of the system on one or more specific applications with assessment data in the domain of surveillance or ATR, including hyperspectral imagery (1000 wavebands, at 10 meter spatial resolution) where some of the main challenges are thruput and high-level, smart processing to help analysts.

These goals require a multi-investigator team that brings together experts on (a) learning algorithms and learning theory, (b) computational neuroscience of the visual system (c) object detection in images (d) design of optimal algorithms for parallel architectures. We have assembled the required team. We describe the team and its responsibilities in section D.

A.1.3 Research Objectives, Scientific Approach and Relation to Work in the Field

Traditional classification systems – as used for instance in ATR and surveillance – typically need expert tuning and even explicit programming for different tasks and often for different environmental conditions. In the case of vision, the inclusion of cluttered background can drastically decrease object detection and recognition performance. Furthermore, existing systems are mostly limited to static objects and cannot deal with recognition of actions and temporal sequences. All these challenges must be overcome in order to create systems that can be really useful for tasks such as remote surveillance, reconnaissance and target recognition.

Currently, no recognition systems other than biological ones are capable of robustly classifying objects for different tasks and in a variety of natural settings. This is still true even for the improved classification algorithms recently developed in the field of statistical learning. For instance the trainable systems demonstrated over the last 2 years in the area of object detection - such as the MIT system based on a Support Vector Machine classifier and the CMU system based on a naive Bayes approach, are the best so far in tasks of object detection in images (as measured by performance of the CMU face database). However, they are still not as robust and flexible as the human visual system. Accordingly, we propose in this interdisciplinary project to extend theoretical models and experimental studies of object and action classification in neuroscience, to develop and test software implementation, and to explore the feasibility of an implementation in software and parallel hardware. The objective is to develop and prototype real-time adaptive and trainable systems, modeled after biology, that recognize static and dynamic patterns, in the presence of significant background clutter, under variable environmental and task conditions. We will focus on the visual domain but the basic architecture will be capable of dealing with other sensors as well. The system will be assessed on appropriate data and applications provided by ONR (one or more of ATR, surveillance/reconnaissance and robotics tasks).

A.2 Outline of the Approach

In section A.3.2 we describe two of the trainable computer systems we have developed recently for object detection in images, at MIT and CMU, respectively. We then describe insights from biology and in particular from a recent model of object recognition in inferotemporal (IT) cortex (A.3.3) and a model of visual attention (A.3.4). Section A.4 deals with how we propose to improve our existing computer systems by using cues from biology. In particular, section A.4.1 discusses research to evaluate whether the hierarchical architecture suggested by the biological model can improve the SVM classification performance for noisy images with clutter. The more general question has to do with hierarchical learning systems, their theoretical motivation and their performance. Since our present biological model of IT cortex lacks the ability to learn features from experience, we plan (see section A.4.2) to explore new techniques for the learning of spatio-temporal features to increase classification performance while adapting to different environmental conditions such as weather. Section A.4.3 describes plans to add top-down and bottom-up attention-like biomimetic pathways to allow for on-line adaptability and learning. Section A.4.4 describes our plans to explore models of computation based on parallel architectures and also novel on-line learning algorithms for SVM. Section A.4.5 describes how to extend the static classification stage of the present biological model with architecture capable of tracking difficult targets. Finally, section A.4.6 deals with the implementation of the system and its demonstration on classification tasks on visual and hyperspectral imagery.

A.3 Technical Background

A.3.1 Statistical Learning Theory

A key problems in ATR and related reconnaissance tasks is the classification problem: how to assign the correct class y to the "image" $\mathbf{x}$ of an object – that is the set of measurements provided by the available sensors in use. Conceptually, this process can be thought of in terms of a very large – possibly infinite – table that provides for each possible object appearance its correct class. Of course the table itself is not known; what is known are some of the entries and their labels – called together training examples. In this view, the problem is how to best approximate the whole table from those few entries. There are two main approaches: in the first the conditional probability densities for each class are estimated and then used to choose the most likely label for new data. In the second approach, the mapping from the input patterns $\mathbf{x}$ to the labels y is learned directly from the examples without the stage of density estimation. In our past work we have used both approaches to develop good image classifiers for object detection. In the first we have used a rather classical "naive Bayes" technique: we will describe the details in a later section dedicated to the CMU object detection system. Here we sketch briefly the theory underlying the second approach used for the MIT system.

Statistical learning and support vector machines

In recent years a new field – statistical learning – has taken form [70, 71]. It combines different fields of mathematics: theory of probability, statistics, function estimation and pattern recognition. It has developed a set of rich results, which provide the foundations for understanding algorithms for supervised learning and for developing new ones. The field has already developed algorithms that have been quite successful in difficult classification tasks such as OCR, text search, bioinformatics, object detection (see for instance [71, 8], and also `http://svm.first.gmd.de/links.html`; for our own work see [38, 34, 37, 36]).

The general theoretical framework is regularization [42, 77, 65, 16, 71]. In this framework a new approach to classification consists of Support Vector Machines (SVM). It involves minimization of a regularization functional of the general form (see for instance [13]):

$$\frac{1}{l}\sum_{i=1}^{l} V(y_i, f(\mathbf{x}_i)) + \lambda\Omega(f) \tag{1}$$

where $V(\cdot,\cdot)$ is a *loss function*, $(\mathbf{x}_i, y_i)$ are the l training data, $\Omega(f)$ is a regularization term on the set of functions f and λ is a regularization parameter. Regularization theory provides the basis for understanding, analysis, and further development of these methods. An important property is their surprising ability to work relatively successfully in very high dimensional spaces with a small number of training examples. A second feature is related to recent glimpses of how to couple more formally the classifier with prior knowledge about a domain. In order to classify complex data objects (images, multiple spectral bands, video sequences), learning approaches must go beyond simple (local) feature vector representations and exploit prior knowledge of the domain to generate more expressive features. In Support Vector Machines the representation and the kernel can be determined from a prior probabilistic model – when available – both in classification [25] and regression [43, 77].

We have recently shown [13] that standard regularization [74, 75, 76, 77] and SVM for both regression (SVMR) and classification (SVMC) correspond to the minimization of equation (1) over the choice of function f for $\Omega(f) = \|f\|_K^2$, where $\|f\|_K^2$ denotes the norm of f in the Reproducing Kernel Hilbert Space (RKHS) induced by the kernel K, and for different choices of V: $V(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$ yields Standard (L_2) Regularization Networks (RN); $V(y_i, f(\mathbf{x}_i)) = |y_i - f(\mathbf{x}_i)|_\epsilon$, where $|\cdot|_\epsilon$ is Vapnik's ϵ-insensitive norm, yields SVMR; $V(y_i, f(\mathbf{x}_i)) = |1 - y_i f(\mathbf{x}_i)|_+$, where y_i here takes values $-1, 1$, yields SVMC (see below). It is remarkable that the minimizer of the regularization functional using *any* of the three loss functions has the same general form $f(\mathbf{x}) = \sum_{i=1}^{l} c_i K(\mathbf{x}, \mathbf{x}_i) + b$, but interestingly different properties. In particular, a SVM classifier [9, 4, 71, 13] is thus defined in terms of a *kernel function* K that measures the similarity of the input to be classified $\mathbf{x}$ with all the l training examples $\mathbf{x}_i$, since K defines a scalar product in the associated RKHS. The decisions are based on the *discriminant function*, $f(\mathbf{x}) = \sum_{i=1}^{l} c_i K(\mathbf{x}, \mathbf{x}_i)$, which is a linear combination of kernels with the weights c_i indicating the relevance of each training example to the classification task. Coefficients c_i are found by minimizing the dual of the functional $min_f \frac{1}{l} \sum_{i=1}^{l} |1 - y_i f(\mathbf{x}_i)|_+ + \lambda \|f\|_K^2$ which is of the form of equation 1. Usually only a few of the c_i are non-zero. Therefore the classifier is *sparse* in the sense that only a few training examples – the support vectors – are needed in the solution [70, 13] with the advantage of a compact classifier after training. The kernel function plays a central role in incorporating prior domain knowledge such as invariances into the classifier, defining the similarity metric used for classification. A SVM classifier is the core engine of the systems developed at MIT.

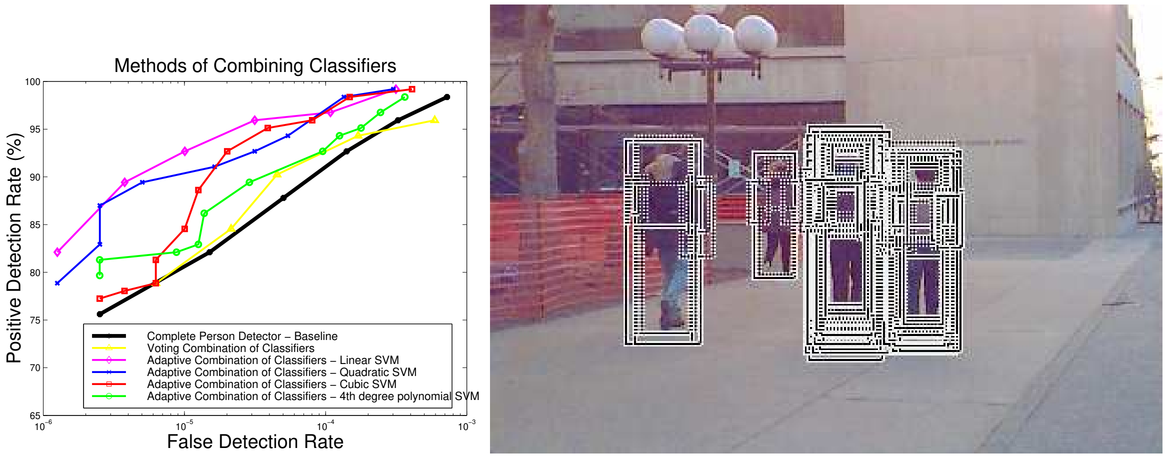
A.3.2 Object Detection Systems: How Computer Vision is Benefiting from Neuroscience

There is a large body of recent work in object detection; of particular relevance is the work on face detection in static images. Recently, example-base approaches have achieved a remarkable degree of success in this field [62], [31], [52], [69], and [35]. These view-based approaches can handle detecting faces in cluttered scenes, and have shown a reasonable degree of success when extended to handle non-frontal views. Sung and Poggio (MIT) and then Rowley and Baluja (CMU) developed view-based systems for object categorization based on learning techniques. Those already impressive face detection systems were improved by Osuna (using SVM classifiers), by Papageorgiou (extensions to pedestrians and cars), and more recently by Schneiderman (2000) and Heisele [18]. All these systems were mainly limited to one view of the object of interest. They were in part motivated by an earlier model based on Gaussian RBF networks as the learning mechanism (Poggio and Edelman, 1990) of IT cortex, later supported by physiological data, which suggested (and demonstrated on very simple images) that view-based detectors could be combined to provide view-invariant recognition. Schneiderman and Kanade (2000) have now indeed shown that view-based detectors can be combined to yield good view-invariant performance for the face and car detection problem in complex, cluttered images (see later).

The MIT object detection system

During the past five years we have developed and implemented at MIT a *trainable* object detection system – based on a core SVM classifier – that automatically learns to detect objects of a certain class in unconstrained scenes [38, 34, 37, 36]. We have demonstrated our system in the tasks of finding faces, pedestrians and cars. Unlike previous approaches to object detection that rely heavily on handcrafted models and motion information [20, 51, 81, 17], our system learns the model of the object of interest from examples. Three years ago we installed a version of our system specifically trained to detect pedestrians in an experimental car of Daimler-Chrysler (in the Esslingen research facility) [36]. The installed version runs successfully at 10 frames/sec on a Pentium and is a promising component of a future on-board automotive vision system that will inform the driver about pedestrians, track surrounding vehicles, and read street signs [14].

A key component of our object detection system, in addition to the learning engine, is the signal representation, that is the choice of the *features* that are used as input to the *classifier*. This is in general a very difficult problem, which needs prior information about the specific domain, and which has been seldom attacked in a general way. In our case we have used an overcomplete, multiscale set of the absolute values of Haar wavelets as the basic dictionary with which to represent the images [34]. Images are transformed from the space of pixels to that of an overcomplete dictionary of Haar wavelet features, which provides a rich description of the pattern.



(a) (b)

Figure 1: *Figure (a) shows ROC curves comparing the performance of detectors based on different methods of combining the individual component detectors. The black 'Baseline' corresponds to the "old" system that detects the pattern of a complete person in an image. Figure (b) shows results from the component-based person detection system. The solid boxes outline the complete pedestrian, where the dashed rectangles are the components.*

The architectural overview of our system is as follows. The system takes as input a set of scaled images of the object class that have been aligned and a set of scaled patterns that are not in our object class. An intermediate representation that encapsulates the important information of our object class is computed for each of these patterns, yielding a set of positive and negative feature vectors, which are then used to train a classifier to differentiate between in-class and out-of-class patterns. We have used a Support Vector Machine classifier (SVM). To detect objects in out-of-sample images, the system slides a fixed size window over an image and uses the trained classifier to decide which patterns show the objects of interest. To achieve multi-scale detection, we iteratively resize the image and process each image size using the same fixed size window.

Our preliminary experiments on people detection using a component-based approach yield promising results. The current system used in the experiments consists of four SVM-based systems, each detecting separately a component of the "person" object: head, legs and two arms. In Figure 1, we show the ROC curves that benchmark the ability of individual component detectors – each can be considered a *pedestrian expert* to correctly flag a person. Examples of detections are shown in Figure 1. The solid lines indicate the person detected and the dashed boxes outline the individual components. As Figure 1 shows (b), the real power of the component-based detection system emerges when we combine the component detectors in different ways to do people detection.

In more recent work we introduced [18] a similar component-based approach to face detection. Again, there are two main ideas behind component-based detection of objects. First, many object classes can be well described by a few characteristic components and their geometrical relations while it might be more difficult to describe them by their global appearances. Second, the patterns of some object parts might vary less under pose changes than the pattern belonging to the whole object. Our component-based system is structured with distinct example based detectors that are trained to find separately components of an object such as the eyes, nose and mouth of a face. After ensuring that these components are present in the proper geometrical configuration, a second example based classifier combines the results of the component detectors to classify a pattern as either face or non-face. The architecture of the component-based face detection system using eyes, nose and mouth is illustrated in Figure 2. We have shown in experiments on face detection that the component-based approach is more robust against pose changes and partial occlusions than comparable systems performing detection on the whole face, paralleling the

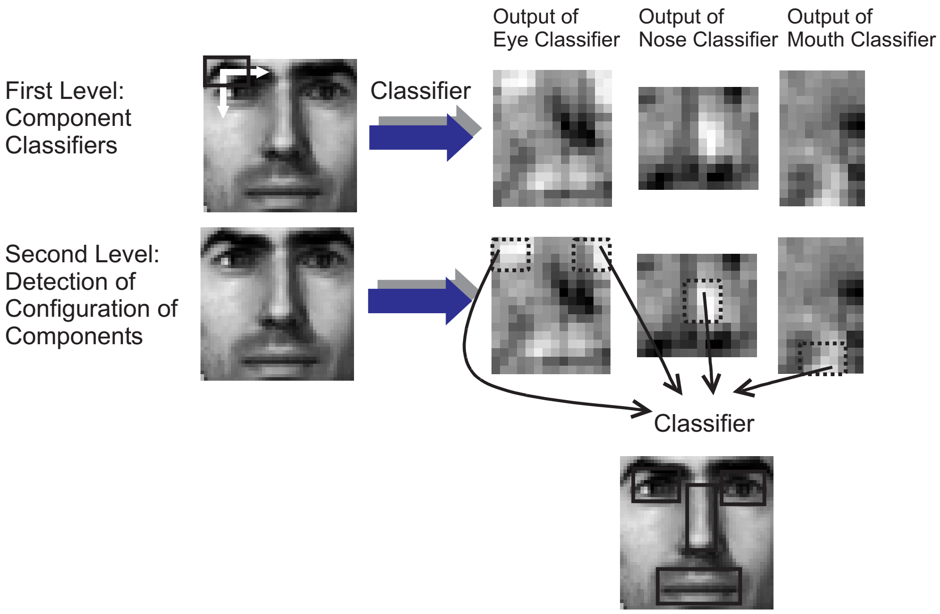


Figure 2: *System overview of the component-based classifier. On the first level, windows of the size of the components (solid lined boxes) are shifted over the face image and classified by the component classifiers. On the second level, the maximum outputs of the component classifiers within predefined search regions (dotted lined boxes) are fed into the geometrical configuration classifier.*

results obtained for people detection.

The CMU object detection system

At the highest level, the CMU object detection system [57, 55, 56] operates in a similar fashion to the MIT SVM system. Like the SVM system it is based around a classifier that distinguishes the object from non-object. In particular, it classifies the input as object only when the object appears at a specified size and alignment within the input window. Therefore, to perform detection on an entire image, we scan this input window over an exhaustive range of positions and scales.

Similar to the latter MIT system [18], the CMU classifier is based around a decomposition of appearance into components. We treat these components as statistically independent, i.e., as in a naive-Bayes classifier. However, our approach at CMU differs from other component-based approaches in the quantity and variety of components we use. In particular, the existing CMU system decomposes the input into components as follows. A series of local operators, f_i, are applied to the input window. Each of these operators could be thought of as a filter giving an output for every shifted position within the input window. (Each of these local operators is the result of vector quantization on various collections of linear wavelet filter outputs). The collection of all these outputs is the "components." All together, these components cover the full extent of the object with overlap. The idea is to use as much information as possible in making the classification. Paradoxically, we can do so without paying a prohibitive penalty in computational cost. We take advantage of the fact that adjacent input windows overlap and that we can share many computations among these windows. We also share computations over successive octaves in the search across scale. Moreover, we use a coarse-to-fine heuristic strategy in evaluating the operators. We begin by first applying coarse resolution local operators. We then only continue evaluation of the finer resolution operators if it is warranted by the outcome of the coarser evaluation.

The CMU classifier is based on statistical models that describe the behavior of these local operators. A unique discrete statistical model is used. First, each local operator, i, was designed to generate a discrete valued output, $pattern_i$, typically with about 10,000 different values. With this representation we can characterize the statistical behavior of each operator in a table. Each entry in the table records the

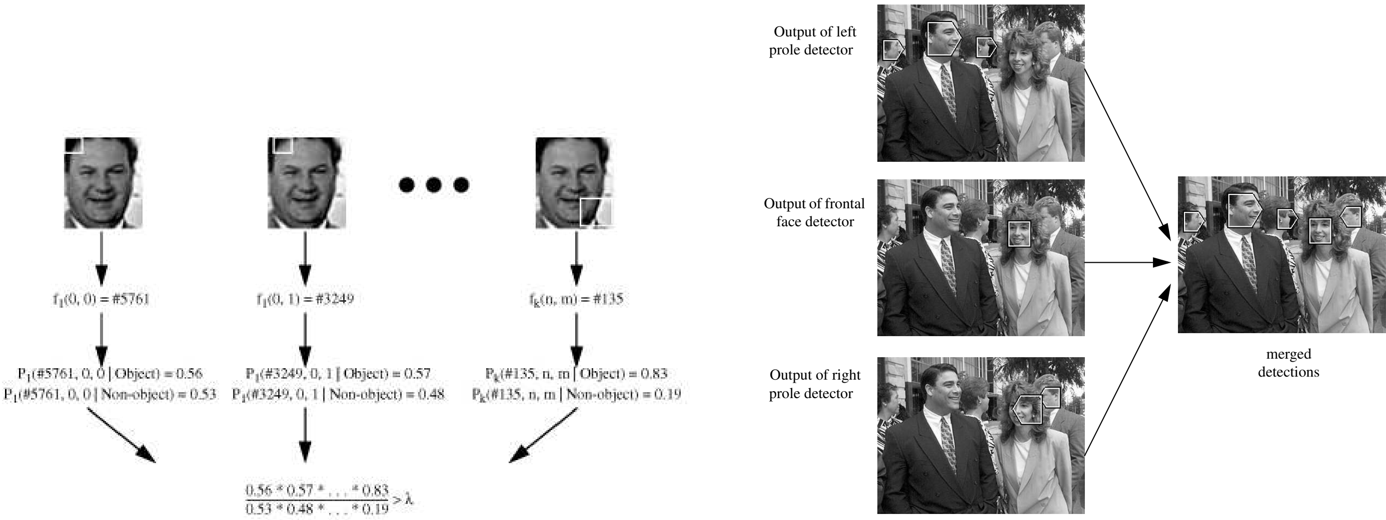


Figure 3: *Left: the CMU classification algorithm. Right: combining view-based detectors to detect faces over a range of orientations.*

probability of occurence of a specific output value. To determine these probabilities we simply count how often each output value occurs over a set of training images and divide that number by the total number of samples in this set. We build two such tables for each operator. One table, $P_i(pattern_i, x, y|Object)$, is computed from a set of training images of the object. The other table, $P_i(pattern_i, x, y|NonObject)$, is computed from a set of images that do not contain the object; we call this the $NonObject$ class. Note that we compute these probability distributions as a function of position, x, y, on the object. By doing so the CMU classifier implicitly captures the geometry of the components on the object.

The CMU classifier then uses these probability tables to classify an input window as illustrated in the figure below. We first compute the output of all local operators. We then look up the $Object$ and $NonObject$ probabilities associated with each output value. We multiply the object probabilities and divide them by the product of the non-object probabilities. We compare the resulting ratio to a threshold. If it is greater than the threshold we classify the input as $Object$. Otherwise we classify it as $NonObject$. This test is called a likelihood ratio test and is statistically optimal provided the statistical models are accurate.

Figure 4: *Results in detection of human faces and automobiles.*

One of the advantages of a components based approach, as mentioned above [18], is its effectiveness under small amounts of variation in object pose. This flexibility allows us at CMU to use a "view-based" approach to detect three-dimensional objects. We train separate detectors corresponding to different viewpoints. For faces we use three viewpoints (right profile, frontal, left side) and for cars we use fifteen viewpoints. At first, we apply these detectors separately. We then merge their results as shown in the figure below. Using this approach we have achieved detection results that robustly span a large continuous range of object orientation. This view-based approach is well supported as a mechanism in natural vision systems by biological evidence [12, 45] and psychological evidence [3] and was suggested by the model of Poggio and Edelman (1990).

The CMU system has been applied to three objects: human faces, passenger cars, and doorknobs. For each of these objects, the classifier is based on the same underlying functional form. We simply gather statistics from different sets of images in each case. Below we show some representative results on these images at a fixed value of the detection threshold. In particular, the system has achieved the most accurate results to date in frontal face detection and the most robust results in profile face detection. Also, the CMU car detector is the first of its kind that spans a wide range of orientations and has been tested to work robustly under a wide range of image variation: lighting, shadowing, reflection, unconstrained background scenery.

A.3.3 Insights from a Model of Recognition in Visual Cortex

Some hallmarks of the human visual system are: i) invariance to image transformations (for example, we can recognize a specific face among many, while being rather tolerant to changes in viewpoint, scale, illumination, and expression), ii) robustness to clutter (target objects can be detected in complex real world scenes), and iii) speed (EEG studies [64] have provided evidence that the human visual system is able to solve an "animal/no animal" detection task in unconstrained images within 150ms, which is on the order of the signal latency in the cortical ventral visual stream [40], thought to be crucial for object recognition). These are highly desirable qualities also for computer vision systems, whose performance still lags far behind their biological counterpart. Our recent model of object recognition in visual cortex, HMAX, has suggested computational mechanisms key to biological object recognition that differ substantially from current computer vision approaches [63, 50]. The model is consistent with a number of physiological and psychophysical data and leads to a comprehensive perspective on the biological processes of object identification and classification (which we shall both summarily refer to as "recognition") [47].

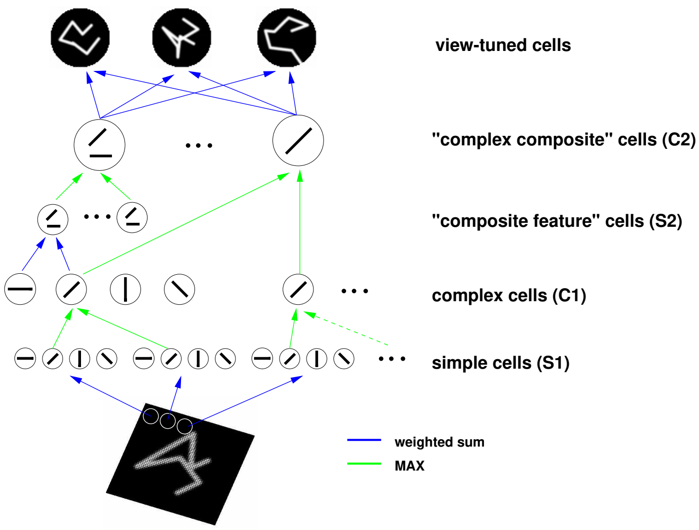


Figure 5: *The HMAX architecture [50]. It consists of a hierarchy of layers with linear (template matching, blue arrows) and non-linear operations ("MAX" operation, green arrows). The non-linear MAX operation — which selects the maximum of the cell's inputs and uses it to drive the cell's output — is key to the model's properties. These two types of operations respectively provide pattern specificity and invariance (to translation, by pooling over afferents tuned to different positions, and scale (not shown), by pooling over afferents tuned to different scales).*

The model is based on a hierarchical feedforward architecture (Fig. 5) that starts out from units tuned to bar-like stimuli with small receptive fields, similar to so-called simple cells in primary visual cortex. At the top of the hierarchy are view-tuned units, as found in inferotemporal cortex, one of the highest areas in the ventral visual stream. These view-tuned units respond specifically to a view of a complex object while showing tolerance to changes in stimulus scale and position. Feature complexity and transformation invariance are built up through a succession of feature complexity-increasing "S" layers and invariance-increasing "C" layers (with the nomenclature following Fukushima's Neocognitron [15]). While feature complexity is increased by a template-match operation, HMAX uses a new pooling mechanism — a MAX operation — to increase invariance in the C layers by pooling over units tuned to the same feature but at different scales and positions.

This MAX pooling operation also has interesting implications for the recognition of objects in cluttered scenes [49]. To correctly recognize multiple objects in clutter, two problems must be solved: i) features must be robustly extracted, and ii) based on these features, a decision has to be made about which objects are present in the visual scene. The MAX operation can perform robust feature extraction: A MAX pooling cell that receives inputs from cells tuned to the same feature at, e.g., different locations, will select the most strongly activated afferent, i.e., its response will be determined by the afferent with the closest match to its preferred feature in its receptive field. Thus, the MAX mechanism effectively isolates the feature of interest from the surrounding clutter. Hence, to achieve robustness to clutter, a unit tuned to a specific object view should only receive input from feature detectors that are strongly activated by that unit's preferred stimulus (i.e., those features that are relevant to the definition of the object) and thus less affected by clutter (which will tend to activate the afferents less and will therefore be ignored by the MAX response function). More complex and more invariant features are built in HMAX through a succession of several S and C layers to obtain an (overcomplete) dictionary of hierarchical features that show tolerance to changes in scale and position of their components. Note that features retain no global position information. Rather, an object is coded by a sufficiently large set of (overlapping) features that constrain the representation to uniquely encode complex objects, similar to the "Wickelfeatures" [80] proposed in speech recognition (see [29, 30]).

It is interesting to contrast this approach to recent computer vision algorithms: A key operation in several systems for the recognition of objects [52, 61] is to scan a window across an image, through both position and scale, in order to analyze at each step a subimage – for instance by providing it to a classifier that decides whether the subimage represents the object of interest. Such algorithms have been successful in achieving invariance to image plane transformations such as translation and scale, over ranges of a few to thousands of pixels. A mechanism like the MAX operation that locally and automatically selects a relevant subset of inputs provide scale invariance and short-range translation invariance over tens up to hundreds of pixels. Long-range translation invariance can be obtained through additional attentional mechanisms. HMAX will in fact profit from an attentional system as described in the next section: since HMAX operates on scene regions of the order of V4 receptive field size (2-4 degrees of visual angle), an attentional system that selects salient regions over the whole visual field will provide long-range translation invariance.

A.3.4 Insights from a Model of Visual Attention

A large body of psychophysical evidence on attention can be summarized by postulating two forms of visual attention [39]. One is driven by the input image itself. This saliency-form of attention is rapid, operates in parallel throughout the entire visual field, and helps mediate popout (our tendency to automatically shift attention towards highly conspicuous objects). Unless subjects close their eyes, sleep or have brain damage, this form of attention cannot be turned off. The second type of attention depends on the exact task and on the subject's visual experience, takes longer to deploy and is volitionally controlled using top-down feedback. Normal vision employs both processes simultaneously. In biological systems, the selection of where to attend has traditionally been attributed to the so-called "dorsal processing pathway," in contrast to the identification of objects being performed primarily in the "ventral pathway." Bottom-up, saliency-based attention is believed to be primarily computed in the dorsal pathway.

Additionally, attention modulates neuronal activity along the object recognition (ventral) pathway that the MIT group has considered in their HMAX architecture. This modulation can take several forms and enhances the activity of neurons encoding for salient objects (mediated by bottom-up pathways that impact inferior-temporal (IT) cortex and related cortical areas), sharpens competition between neurons encoding for nearby or overlapping stimuli [28] and also carries attribute-specific information mediated by top-down processes from working-memory [10].

Most models of visual search, whether involving overt eye movements or covert shifts of attention, are based on the concept of a saliency-map [27]; that is an explicit, viewer-centered map that encodes the saliency or conspicuity of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the next target "attended" by the system. Inhibiting this location automatically allows the system to attend to the next most salient location.

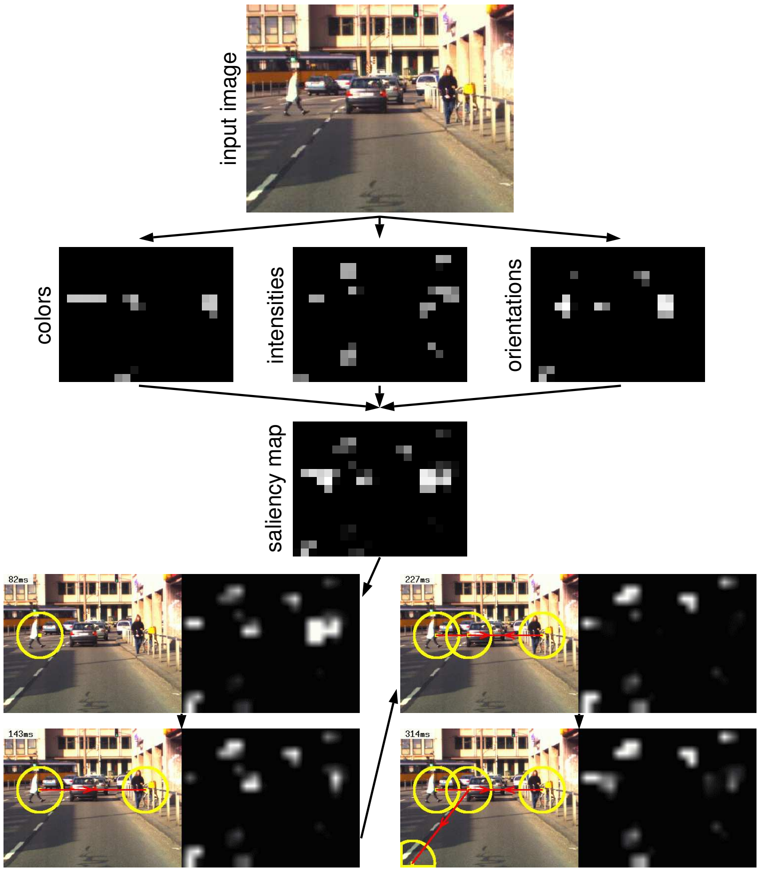


Figure 6: *Overview of the bottom-up attentional selection model developed by Itti and Koch at Caltech (Itti is now at USC). The input image is decomposed into seven channels (intensity contrast, red/green and blue/yellow double color opponencies, and* $0°$*,* $45°$*,* $90°$ *and* $135°$ *orientation contrast) at six spatial scales, yielding a total of 42 feature maps. Here the six intensity maps, twelve color maps, and 24 orientation maps are combined into the three summary maps shown. After iterative spatial competition for salience within each map, which allows only a sparse number of locations to remain active, all maps are combined into the unique saliency map. The saliency map is scanned by the focus of attention in order of decreasing saliency, through the interaction between a winner-take-all neural network (which selects the most salient location at any given time) and an inhibition-of-return mechanism (which transiently suppresses the currently attended location from the saliency map, as shown on the four bottom frames). In the example shown, the system selected the two isolated pedestrians as being the two most salient locations in the image, mostly because of their strong responses in the orientation channel. See http://iLab.usc.edu/bu for an interactive demonstration of system performance with several hundred images.*

The Caltech group (Itti has now moved to USC, while Koch remains at Caltech) implemented such a model and applied it successfully to a wide range of target detection tasks, using synthetic and natural stimuli (see below). As discussed later, we propose to incorporate bottom-up and top-down attentional processes into the object recognition systems discussed here (*i.e.*, the MIT and the CMU systems) to enable them to more efficiently identify objects in cluttered, natural scenes. In our biologically-inspired system [24, 23], a multi-spectral (e.g., color) input image is decomposed into a set of multiscale neural "feature maps" which extract local spatial discontinuities in the modalities of color, intensity and orientation. The number of feature channels used depends on the number of spectral bands available in the input. Each feature map is endowed with non-linear spatially competitive dynamics, so that the response of a neuron at a given location in a map is modulated by the activity in neighboring neurons. Such contextual modulation, inspired from recent neurobiological findings on non-classical receptive fields [59], has proven remarkably efficient at extracting salient targets from cluttered backgrounds in visible-band color imagery. All feature maps are then combined into a unique scalar "saliency map" which encodes for the salience of a location in the scene irrespectively of the particular feature which detected this location as conspicuous (Figure 6). A winner-take-all neural network detects the point of highest salience in the map at any given time, and draws the focus of attention towards this location. In order to allow the focus of attention to shift to the next most salient target, the currently attended target is transiently inhibited in the saliency map (via an *inhibition-of-return* mechanism that has been extensively characterized in humans [46, 26]). The interplay between winner-take-all and inhibition-of-return ensures that the original scene is scanned in order of decreasing saliency by the focus of attention, and generates the model's output in the form of spatio-temporal attentional scanpaths. The system performs remarkably well at picking out salient targets from cluttered environments. Experimental results include the reproduction by the model of human behavior in classical visual search tasks (popout versus conjunctive search, and search asymmetries; see [66, 23]); a demonstration of very strong robustness of the salience computation with respect to image noise [24]; the automatic detection of traffic signs and other salient objects in natural environments [21]; the detection of pedestrians in natural scenes (see below for preliminary results). Finally, the model's performance at detecting military vehicles in the high-resolution *Search2* NATO database of color rural scenes exceeded human performance, in terms of the estimated number of locations that need to be visited by the attentional searchlight before the target is located [22].

The system shows a large degree of adaptability, using supervised training to learn the optimal set of weights with which all the feature maps contribute to the scalar saliency map [21]. This permits the system to be trained to reliably detect traffic signs, soda cans, or pedestrians, despite the high intra-class variability in such targets [21].

This model, originally motivated by neuroscience, can strongly contribute to the overall performance of a trainable, general purpose ATR architecture. Of particular interest for the present proposal is a preliminary study, in which the SVM-based algorithm for the automatic detection of pedestrians of Papageorgiou and Poggio at MIT [37] was combined with the saliency algorithm of Itti and Koch at Caltech. The attentional system was used for rapid selection of a small number of candidate image locations, and only at those locations the full SVM-based target recognition was attempted. The computational time required by the attentional system was negligible compared to that required by the target recognition system (a few seconds versus over one hour for an exhaustive search at all locations and scales by the SVM algorithm); yet, the attentional system demonstrated a strong tendency to attend to target pedestrians. This tendency was further refined using supervised training. Our preliminary study showed that if SVM-based target recognition was limited to the first five attended locations, the overall detection rate was only degraded by 5% while accelerating the entire process by more than 3-fold.

A.4 Proposed Research

Taking these pieces of work as starting points, we propose to develop and test them on complex real-world classification tasks in the domains of surveillance and ATR, while stripping out the purely biological details. We plan to leverage the two biological models described earlier. In particular, we will add a more sophisticated classification stage at the top stage of the recognition model, based on the Support Vector Machine learning technique that we have successfully applied to real problems of object detection. We have recently invented an on-line incremental algorithm for SVM learning that is fast and memory-efficient, making real-time implementation feasible. This will allow us to learn representations of diverse object classes and spatio-temporal structures corresponding to dynamic activities. We will also extend in an important direction the HMAX model by introducing schemes for adaptively learning relevant features. Theoretical work in our labs is under way to develop and refine to versatile strategies for feature learning. These strategies will prove useful for creating autonomous and custom-trainable classification systems that minimize the need for manual programming. The model of visual attention described earlier suggests the details of another key enhancement to our recognition system – both from the point of view of biology and of DoD applications – by adding top-down control signals in order to maintain robust classification performance even under significant viewing transformations, clutter and a variety of different environmental and task conditions.

These ideas, explored in an interdisciplinary fashion will allow us to make significant headway on the important and challenging problem of robust and adaptable object classification and target recognition, for instance in hyperspectral imagery. To tie all these endeavors together, to test the validity of the ideas and also to pave the way to image processing systems, we shall study the feasibility of real-time, parallel architectures. The next sections provide details on each of the main points of our research plan.

A.4.1 Hierarchical Architectures for Object Detection

The classifiers used in some of our systems – such as Support Vector Machines – correspond to one-layer non-hierarchical architecture. Are hierarchical architectures justifiable in terms of statistical learning theory? The superficial answer from the theory seems negative. Of course, hierarchical classifiers are nothing new: tree-based classifiers are some of the most used classifiers and tree-structured algorithms have been used for the related tasks of clustering, density estimation and Vector Quantization (VQ) (see [11] for a number of (asymptotic!) theoretical results). However, at the moment, the approach based on Vapnik's statistical learning theory does not offer any general argument in favor of hierarchical machines for classification. This is somewhat of a puzzle since the organization of cortex – in particular visual cortex – is strongly hierarchical. As we described earlier, Riesenhuber and Poggio [48] have demonstrated that a hierarchical model of cortex can account quantitatively for several physiological data on IT neurons likely to be directly involved in visual object recognition. Thus the key question for machine vision applications is whether this translates to superior recognition performance.

As a first step, we propose empirical tests to find out whether a hierarchical organization such as in HMAX – complemented by a SVM classifier in the last stage – can perform better on real image classification tasks than the full-object MIT system. Furthermore, our component-based systems – both MIT and CMU – are clearly representing the first step of a hierarchy: we plan to understand the relation between them and the HMAX model of the cortex and how to design components-based architectures at multiple, recursive levels as in cortex.

In general, the theoretical issues surrounding hierarchical systems are wide open – and likely to be of paramount importance for the next major development of efficient classifiers for information extraction, ATR and reconnaissance. We have seen that the architecture of the visual system is hierarchical. In addition, cognitive science shows that humans are able to construct hierarchical representations, whereby constituents are used and reused in an essentially infinite variety of relational compositions. An upper case L consists of two suitably arranged and suitably sized line segments. The line segments are themselves nameable and useable objects, and in fact could be used in any of a nearly infinite variety of other compositions.

Why hierarchies? There are reasons of *efficiency*– computational speed and use of computational resources. For instance, the lowest levels of the hierarchy may represent a dictionary of features that can be shared across multiple classification tasks. Hierarchical system usually decompose a task in a series of simple computations at each level – often an advantage for fast implementations. There are also more fundamental issues of *speed of learning*. A *key* difference between our best classifiers and human learning is the number of examples required in tasks such as object detection. The MIT car detection system uses thousands of examples, the CMU system more than a million (including virtual examples, see later); in contrast, it seems that monkeys are able to learn the task with just a few examples (of course it is not easy to assess the role of prior experience). Statistical learning theory shows that difficulty of a learning task can be quantified through a "complexity" measure such as VC dimension [73]. This complexity determines how many training examples are needed to achieve a given level of generalization error. A part of the generalization error depend on the quantity $\frac{h}{l}$, where h is a measure of the complexity of the classifier and l is the number of training examples. Thus the complexity of the learning machine sets the speed limit and sample complexity for learning. If a task can be decomposed into low-complexity learning task, for each layer in the representation, then each layer only requires a small number of training examples. For instance, each layer may consist of a set of classifiers with low VC-dimension, but the hierarchical system may be able to represent a rich and complex set of hypotheses. Thus, instead of capacity control (which controls h to be small enough depending on l) or locality control (which keeps h small and controls the local number l), hierarchical machines decompose the problem in subsets of the input space in a hierarchical architecture: h is smaller for each of the classifiers in each layer AND the number of examples needed by each classifier is thus small.

Notice, however, that not all classification tasks have a hierarchical representation. More precisely, as Stuart Geman has pointed out, not all categories can be represented in a hierarchy with a fixed and bounded number of levels and a fixed and bounded complexity at each level. Roughly speaking, the issue is under which conditions a function of many variables can be approximated by a function of a small number of functions of subsets of the original variables. Cognitive science suggests that what humans can learn can be represented by hierarchies that are locally simple.

The issue of multilearning – learning multiple classification tasks – is also likely to be important in a theory of hierarchical recognition systems. In a multi-stage learning scheme, common features can be learned for different recognition tasks together with general invariances. The challenge of course lies in finding a learning scheme that describes how visual experience drives the development of features at lower levels.

Visual neuroscience, as summarized by models such as HMAX, suggests another property that hierarchical visual representations must have: invariance to position, scale and (to a lesser extent) image-plane rotations. These invariances can be built at the level of each layer of a hierarchy, like the HMAX model. In such hierarchical systems, "receptive field size" increases towards higher levels together with feature complexity: there is in fact an interesting trade-off between the complexity of a feature and the spatial density of similar features. This is somewhat similar to the well-known trade-off between size of visual filters and "Panum's area" for stereo and motion correspondence (see section A.4.5). The hierarchical structure of HMAX suggests how to break down the learning problem and first focus on the learning of features within one layer, and subsequently extend the system to the learning of feature hierarchies and the concurrent learning of features and short-range invariances. We propose to start with a two-layer version of the model that contains only one pooling stage. On one end of the complexity spectrum of learning schemes will be simple unsupervised feature extraction algorithms like vector quantization. On the other end are more elaborate methods [68] (see also [19, 78]) in which features are learned through the selection of significant components common to different examples of a class of objects.

In summary, we plan to do empirical tests of the performance enhancement that can be achieved in the CMU and MIT system by using hierarchical representations such as in the biological HMAX model. In particular, we will extend the learning architecture to several hierarchy levels, as in the full HMAX model. In addition to the independent scanning over scale and position in each feature channel via the MAX operation in the two-layer model, features in the multi-level model are tolerant to local deformations, which is expected to further increase the robustness of the feature dictionary with concomitant advantages for recognition in cluttered real-world scenes. More fundamentally we aim to develop a theory of hierarchical visual recognition systems, in the framework of statistical learning. In the next section, we will describe the first application-relevant steps we plan to do for learning features and components from data, as well as our plans to extend the technique of virtual examples.

A.4.2 Learning Features and Components and Generating Virtual Examples

The previous section argues that our long-term goal is a system in which the features used at the various levels of a hierarchy (higher up in the hierarchy the term components may be more appropriate than features) are all learned automatically from data, across different learning tasks and quickly adapted for new tasks and different environmental conditions. As first steps towards this goal we will experiment with different approaches to learning components and features from data. In parallel we will also continue to work on a complementary approach that is key to training the classifiers we have developed so far: the use of *virtual examples.* Training sets are always too small for the problems of object detection given the high dimensionality of images. Prior information can be used to enlarge the training set by generating virtual examples by transforming the real ones with the use of "legal" transformations. In the case of face and people detection, for instance, legal transformations are scaling, translation, some rotation in the image plane, blurring, other affine transformations. Both the MIT and the CMU system use virtual examples, with the latter using an especially large number of them; virtual examples are *the* key to its very good performance.

Learning features and components

In the previous experiments with the MIT component-based system, we manually selected the basic features such as grey-levels or wavelets and the components such as eyes, the nose and the mouth (for the face detection system). In the CMU system, we choose the filters (referred to as *operators*) based on behavior we expect in the imagery. In particular, we expect there to be strong statistical dependencies with locality in position, frequency, and orientation and design the local operators accordingly. It would be highly desirable that the system could choose the filters and the components automatically based on their discriminative power and their robustness against pose changes. We plan to explore several different techniques, including both supervised and unsupervised methods for learning these operators as a function of the specific characteristics of the object we are trying to detect. Among the unsupervised techniques, we plan to use methods such as learning vector quantization for unsupervised learning – which we have used already in recent work on our biological HMAX model. At CMU we plan to explore discriminant based methods such as the Foley-Sammon and the Fukunaga-Koontz as supervised methods.

In the following we describe in more details two general approaches for learning components from examples: the first is based on *selecting* from a dictionary of possible components; the second approach *synthesizes* "optimal" features from the data.

Selecting from a Dictionary of Components: In one implementation of the idea, we arbitrarily defined a number of components and let the geometrical configuration classifier learn to weight the components according to their relevancy. At MIT we carried out an experiment with 16 non-overlapping components of size 5×5 evenly distributed on the 19×19 face pattern. As in previous experiments the component classifiers were SVMs with 2nd-degree polynomial kernels and the geometrical configuration classifier was a linear SVM. Since the geometrical configuration classifier calculates the sum over the weighted outputs of the component classifiers, the weight is a reasonable measure for the relevancy of a component. The training errors of the component classifiers give information about the discriminative power of each component. Some of the components are located on the cheeks of the face. They contain only few gray value structures, which are reflected in the comparatively high error rates. The 16 component classifier achieved similar detection rates on the test set as the five component system of Figure 2.

In a second experiment, we plan to generate a large number of random components and then select the best components according to measures derived from SVM learning theory. For selecting features a well-motivated approach is based on recent work done at MIT and ATT for performing feature selection based upon minimizing bounds on the leave-one-out error for Support Vector Machines (SVMs) [79, 7, 32]. We have done preliminary experiments with these algorithms on both toy problems as well as problems of face detection, pedestrian detection, and DNA microarray analysis. The basic idea is to rescale the features, which are input to the classifier, and then eliminate dimensions corresponding to small-scale parameters.

We can parametrize the kernel function for the SVM as

$$K_\sigma(\mathbf{x},\mathbf{y}) = K((\mathbf{x}\cdot\sigma),(\mathbf{y}\cdot\sigma)) = (\Phi_\sigma(\mathbf{x})\cdot\Phi_\sigma(\mathbf{y})) \tag{2}$$

for any K where the vector σ scales features. Feature selection is performed by the following iterative procedure:

1) Solve the standard SVM problem using the kernel (2) with σ initialized as a vector of ones. This gives us one over the margin $W(\alpha^0,\sigma)$, for the support vectors compute the radius $R(\beta^0,\sigma)$. 2) Minimize $R^2W^2(\alpha^0)$ with respect to σ by gradient decent.

3) Eliminate the k-dimensions corresponding to small values of σ_k.

4) Return to step (1) and repeat until all dimensions left are significant or a local minima is reached.
In the case of pedestrians and faces we used as input features the principal components [58] with good results. In this case the kernel is parameterized as $K_\sigma(\mathbf{x}, \mathbf{y}) = K((\Lambda \mathbf{x} \cdot \sigma), (\Lambda \mathbf{y} \cdot \sigma)) = (\Phi_\sigma(\mathbf{x}) \cdot \Phi_\sigma(\mathbf{y}))$ for any K where the vector σ scales features and Λ is the matrix of principal components.

Synthesizing Components: The second approach takes the opposite route: instead of selecting components from a dictionary, it systematically grows components for the object of interest starting from random seed points on the surface of an available 3D object model. The termination of the component growing is again determined by a measure derived from learning theory. In preliminary experiments at MIT, we could show that for face detection a component-based system using learned components clearly outperforms a comparable whole face detection system. Detection results for a system with 14 learned components are shown in Figure 7. Preliminary results show that this classifier may perform significantly better than any other classifiers (including the best so far – CMU Schneiderman's) when trained and tested with high resolution images (60x60).

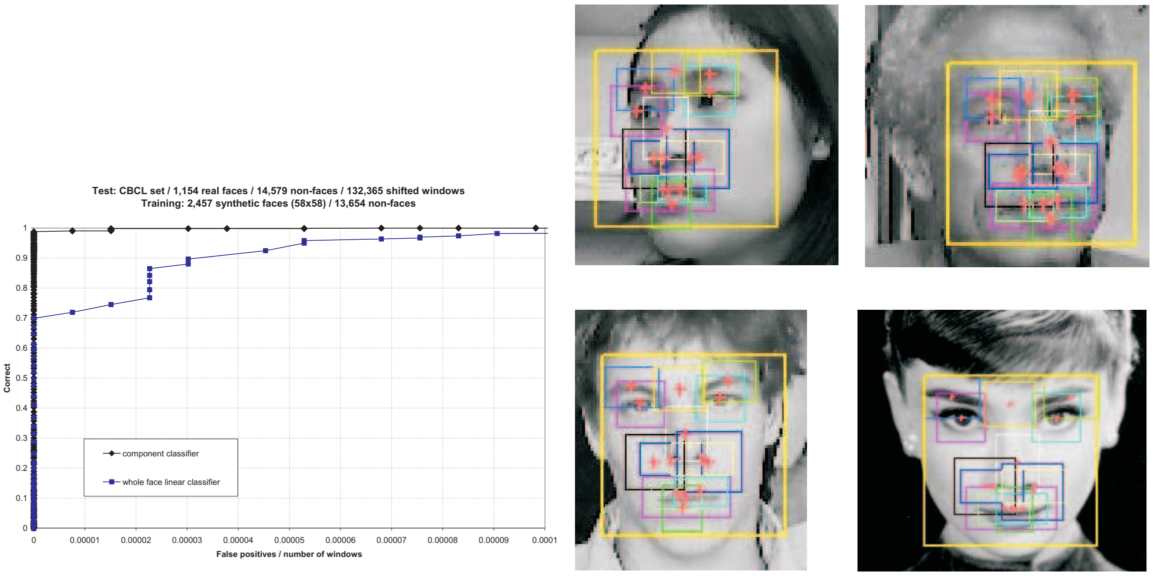


Figure 7: *Left: ROC curve of the (high-resolution) system based on learned components vs the (low-resolution) without learning. Right: examples of component-based face detection using 14 learned components.*

Virtual Examples

The idea of virtual example is a natural one. It was developed in different ways by Abu-Mustafa and Poggio and Vetter [44] among others. They showed how to use prior knowledge about the problem domain to create virtual examples and expand the training set. Niyogi, Girosi and Poggio [33] review the basic techniques (see also the review by [2]) and show that virtual examples are mathematically equivalent, under some conditions, to incorporating the prior knowledge in a regularizer (in fact the choice of a kernel is closely related to the choice of the features, see an early section). Thus, there is an elegant theoretical justification of the virtual example technique for approaches such as SVM classifiers; the justification is simpler for the Bayes approach. Sung and Poggio [61] (1994) used the virtual example technique in their object detection system (for instance for face detection) and since then virtual examples have been used in all object detection systems. Both the present MIT and CMU systems use virtual examples. Schneiderman at CMU has exploited them in the most extensive way so far: the CMU face detection system uses in the order of 10^6 (1M) virtual examples by applying *legal* transformation such as blurring, changing scale and other affine transformations to a few thousands "real" examples.

We plan to explore the theoretical issue of prior information, invariances in recognition, kernels and virtual examples. We also plan to extend our techniques to generate virtual examples. For instance, at MIT we have begun to use 3D models of faces – obtained from examples through the new technique ([2]) of *morphable models* – to generate training examples of our classifier under an almost infinite number of different conditions of identity, illumination and viewpoint. A similar approach could be used for other objects and other sensors – for instance for hyperspectral imagery of certain types of objects.

A.4.3 Biomimetic Bottom-up and Top-down Attention-like Pathways

We propose to extend our trainable system for bottom-up, saliency-based visual attention to multispectral and dynamic image sequences. Because such extension is computationally non-trivial, we will first implement our algorithms on parallel clusters of CPUs. We then propose to study two broad problems posed by multispectral and multiframe datasets: first, the overall architecture of the model needs to be enhanced to exploit multiple spectral bands in a more integrated manner; second, salience information specific to moving objects needs to be computed.

High-Performance Computing: In the past, we have not been able to handle multi-spectral imagery and animated sequences because of limitations in processing power. Indeed, the performance of our vision systems depends heavily on the use of non-linear spatially competitive mechanisms, for attentional selection with the Caltech/USC algorithm as well as for object recognition with the MIT/CMU algorithms. However, we can now attack these problems, as we have just completed the deployment at USC of a small parallel cluster of high-speed Linux computers (16 nodes at 733MHz, 4.5 GByte total RAM, 1/2 TByte total disk, 300 Mbps fully switched network). One part of the proposed research at USC will hence consist of the implementation of our current and future algorithms to such high-performance computing clusters. Beyond allowing real-time processing of visible-band video sequences, such high-performance implementation will be mandatory if we want to process multi-gigabyte datasets comprising multiple spectral bands.

So far, the saliency model has been used with relatively small numbers of channels, usually seven (two for color opponencies, one for intensity, and four for orientation), processed in separate computation streams. In principle, nothing prevents the extension of the model's architecture to many more channels. In a preliminary feasibility experiment, we extended the model to 256 orientation channels (spanning 180° in 0.7° steps). This increased computation time 40-fold on a standard color image but otherwise posed no problem. Because the expensive convolution operations in different channels are computed in parallel, and because our cluster is equipped with very high-speed interconnect (300 Mbps theoretical, 250 Mbps tested), we expect near-to-maximal speedup through coarse-grain parallel implementation of the algorithm. Being able to handle multi-gigabyte image sequences should become feasible using USC's cluster, in particular through the use of tools such as the "Parallel Virtual File System" from Clemson University, to stripe data across the compound 1/2 TByte of live storage on our 16 nodes, for parallel read/write operations.

Cooperative Computation across Channels: So far, the saliency algorithm processes different feature channels such as colors and orientations independently of each other. With images comprising large numbers of spectral bands, we expect to see significant redundancy in the ATR-relevant information contained in each band. We consequently propose to extend the saliency system to model more sophisticated interactions between spectral channels. To investigate this problem, we will implement connections across channels by which each feature map at a given scale can receive multiplicative excitatory or inhibitory input from another feature map at same or different scale. These connections will extend the interactive spatial competition for salience already implemented in the saliency model: at each time step, spatial interactions within each map will be iterated, followed by one iteration of interactions across maps. The supervised training algorithm, which we have used to learn the relative weights of the various feature maps, will be extended to include training of the weights by which the different channels interact. Because such extension will yield a highly adaptable system with several tunable parameters, one important aspect of the proposed research will also be the extension of our training algorithm to on-line adaptability. This will require a modification of the learning technique (currently using batch processing) and will directly benefit from the proposed integration between saliency computation and object recognition proposed in a later section (so that we will not have to manually outline targets to be learned as we have done in the past but can outline them automatically). With such extension, we expect the system to be able to exploit multi-spectral imagery in a much more sophisticated manner than is currently possible.

Computation of Motion-derived Salience: We will develop and implement a motion channel to be added to the existing feature channels of the bottom-up attention system. To follow our so-far successful biological inspiration, the motion channel will be derived from the *Spatio-Temporal Energy Model* proposed by Adelson and Bergen to model motion perception in primates [1]. This algorithm uses 3D spatio-temporal bandpass filters to extract optical flow in a given direction and at a given velocity (just like our existing orientation channels are based on 2D spatially bandpass filters to extract features in a given orientation). We will investigate efficient implementations of such 3D filters using 3D extensions to the steerable filter decomposition scheme [60]. Presently, a single node of USC's parallel cluster can compute the saliency map for a single 512×384 24-bit color (trispectral) frame (4.7 Mbit) in $\sim 1.8s$, and we have benchmarked the cluster's channel-bonded triple fast-ethernet network to sustained TCP transfer rates of 250 Mbps under normal load; we can thus transfer up to 150 bands per second at that resolution, such that our rate-limiting factor is CPU speed (1.8s / 16 nodes). With some optimization of our code for speed (which we have not done so far), we hence expect to be able to process color video sequences at 5-15 frames per second (depending on the additional computational complexity of the new motion channel). Images with a very large number of spectral bands may not be processed in real-time on our cluster, but the speedup obtained by parallelizing the algorithms will still be appreciable, especially during development and testing of computational hypotheses.

It is unclear whether the spatial competition, which we have implemented in the other channels, is appropriate for the motion channel (experimental evidence supporting our approach so far has come from experiments using static stimuli). We will thus test our current approach for the combination of multiple feature channels into a unique scalar saliency map, and determine how it compares to a number of alternative approaches which we have previously studied on static images [21].

Grouping and Saliency: An interesting and non-trivial issue involves spatial grouping processes. As is well known in human vision, grouping cues, such as common motion, color, elongated contours and so on, provide a powerful means to increase the saliency of objects. Thus, three objects moving at the same speed across the sky are much more conspicuous than three independently moving objects and are often perceived as a single "group." The drawback in exploiting grouping for detecting salient object groups is that it requires spatial interactions across neighborhoods, a computationally expensive proposition. However, several authors have recently investigated local, iterative schemes for evaluating the saliency of extended contours and other objects. We plan to investigate such schemes to increase the performance of the saliency-based front-end by incorporating non-local cues.

Attention and Object Recognition: In preliminary work started under the impetus of a prior MURI project, we have successfully demonstrated feasibility of a synergy between the CalTech-USC visual attention front-end and object recognition back-ends such as MIT's. Thus, we now have the chance to combine both, giving rise to a single, very powerful system that mimics both the "ventral" object or *what* pathway as well as the "dorsal" *where* stream that originates in primary visual cortex.

A key operation in both the MIT and the CMU 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 the object of interest is present. This is a computationally very 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. Our preliminary work has shown large speed-ups, since evaluating the saliency of the image is far less computationally demanding than identifying objects. Notice that use of the HMAX architecture would automatically provide some initial invariance to position and scale, which will reduce the extent of scanning needed.

In addition to merging bottom-up attention and object recognition, we are also planning 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 towards the viewer, then the saliency-computation can be biased appropriately by feedback connections — that is, all red objects will be inherently more salient than green ones. It is known that attention can act in this manner in the monkey's visual system by biasing neurons with the appropriate stimulus selectivity [67]. We will also investigate the computational role of attentional signals to tune the feedforward recognition system for specific tasks. Such top-down control of recognition fits naturally in the HMAX framework, without sacrificing the speed advantage of the feedforward architecture: The MAX operation in the present HMAX model performs a scanning operation completely controlled by the incoming stimuli, without top-down influence. We propose to add the ability to perform selection to HMAX, driven by an attentional signal provided by the attentional system. That is, a tunable component will be added to the MAX function, that allows one to not just pick the most strongly activated afferent but the afferent closest to some particular value v which can be set using top-down signals. Thus, when looking for a certain object that, when presented in isolation, evokes a response of v in a certain feature channel this activation could be isolated even in the presence of more strongly activated afferents. An interesting question in this context, also with respect to neurophysiology, is how far down in the hierarchy the effect of cueing has to reach to be effective.

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. We believe that this ambitious goal of mimicking for the first time most of the principal elements of the primate visual system, and doing it for multi-spectral images in close to real-time, is a realistic and feasible one.

A.4.4 Neural Models of Computation and On-line Learning Algorithms for SVM

On-line incremental and decremental SVM learning

The task of training an SVM is equivalent to solving a quadratic programming (QP) problem in a number of variables equal to the number of training examples. The standard QP formulation is infeasible in a real-time application such as adaptive classification of hyperspectral imagery, since memory and CPU requirements would grow quadratically with time. Some efficient methods have been proposed to speed up the SVM training process dramatically [35, 41], however these still require several passes through the training set, infeasible in a real-time setting.

Recently we have succeeded in formulating a fast SVM learning algorithm, which trains the SVM incrementally, one data point at a time [6]. The *incremental learning* algorithm is functionally equivalent to other SVM learning algorithms, but it finds the optimal solution to the large-margin training problem with $N+1$ points from the N-point optimal solution and the extra data point, plus a number of learning state variables (almost) independent of the cumulative data size N. Furthermore, the adiabatic incremental learning algorithm is reversible, and *decremental SVM learning* [6] allows to efficiently compute an exact *leave-one-out* (LOO) estimate of the generalization error, while training. The LOO procedure estimates the generalization error directly on the training set (without a separate cross-validation set) by evaluating the "prediction" on each training point obtained by leaving that point out of the training process. Normally this requires N training sessions each on $N-1$ training data, or a suitable approximation with less computational burden [72]; decremental SVM learning obtains an exact LOO estimate with one pass through the data. A real-time estimate of the generalization error is extremely important so that the learning system can be monitored, and correspondingly tuned, for optimal classification performance. Our efforts at JHU on this front in the context of real-time complex image classification tasks will include:

1. On-line versions of incremental and decremental SVM learning: Although the incremental SVM training algorithm is on-line in spirit, at present real-time implementation requires excessive memory resources to store past data which could be needed later depending on future data. We will investigate the option of including "reserve vectors" in the support vector formulation, among with other techniques to reduce memory requirements such as replacing the Jacobian in the recursive QP formulation with a conjugate gradient method.

2. Automated tuning of global system parameters based on analytic LOO sensitivities: The LOO estimate from decremental training in the topmost SVM of the modular vision system can be used directly to tune system parameters in all layers of the architecture. We plan to make use of analytic derivatives of the decremental LOO generalization error to SVM parameters and input data to compute sensitivities and optimize generalization performance correspondingly. Sensitivities to system parameters in preceding stages of the architecture will be obtained from applying the chain rule on the data feeding into the SVM. This presents a powerful paradigm for optimizing the *entire* system architecture for maximum performance and minimal overtraining, not just the SVM. The question what features are important in the data is of paramount importance in any learning problem, and the LOO sensitivity method will allow to *dynamically* select features of interest, such as the right combination of spectral bands in SAR and other imagery, in on-line fashion.

3. A study of the complexity of the algorithms, in light of real-time implementation: Our efforts here will start by analyzing the CPU and memory requirements of the algorithms in software on conventional computers, and will then consider alternative realizations that make use of parallelism in a distributed architecture, such as DSP or special-purpose implementations. Of particular interest are neural models of computation that closely resemble the structure of biological information processing systems.

A real-time adaptive architecture

Our objective in this part of the effort (at JHU, mostly during the IV and V year) is to develop a custom-trainable, on-line adaptive, and yet versatile and reconfigurable system that outperforms the USC software emulation of the models by at least 2 orders of magnitude in speed, power dissipation, and density of integration. Our goal is a system that is capable of combining a multitude of sensory modalities, including hyperspectral and radar imagery, and adaptively selecting modalities and features of interest. The system will be configurable, capable of interfacing with and converting various image input formats at variable spatial scale. It will serve a variety of ATR and surveillance applications, ranging from detection of obscured targets in foliage on a fast-moving platform, to distributed networks of sensors in UAVs. We envisage demonstrating the following capabilities in this system: 1) Hierarchical, modular and reconfigurable architecture; 2) Real-time incremental learning and adaptation; 3) ability to incorporate contextual information in a top-down manner to improve object detection performance under different environmental and task conditions.

The modular adaptive vision system will consist of a core SVM processor, implemented using a combination of off-the-shelf and custom hardware, and peripheral processors for feature selection. The peripheral components will implement biological models of processing that yield a more invariant representation of features for classification by the SVM processor. These include the qualitative representation scheme and the HMAX model described above. Input to the system will be custom-selectable from either a standard digital video camera (*e.g.*, for surveillance applications) or an externally supplied image or video input stream (*e.g.*, for hyperspectral or SAR data). The system will also feature a bi-directional interface to a workstation for configuration and programming, performance assessment and data acquisition and analysis. The computationally expensive part of evaluating kernel functions in the SVM calls for special-purpose hardware to achieve a factor 100-10,000 improvement in speed over software implementations, necessary to run the algorithms in real-time (including on-line incremental learning) while ensuring good generalization performance under variable environmental conditions (background clutter, illumination variations, occlusions). Our team has previously developed biomimetic sensory processing systems in custom VLSI [5, 54] that meet these stringent computational throughput and power dissipation requirements. One team member (GC) also has an ongoing ONR Young Investigator Program (and PECASE) that addresses the computational efficiency of neural models for signal processing implemented in silicon.

In years 4 and 5 of this effort, we will prototype the hardware using a combination of neuromorphic analog and conventional digital signal processing techniques, and demonstrate the system on a surveillance application, and adaptive image and hyperspectral data classification tasks, of interest to the Navy.

A.4.5 Tracking in Clutter: Computation of Correspondence with Complexity Hierarchies

Consistent with the hierarchical structure proposed before, we are planning to exploit hierarchical representations for the efficient calculation of motion correspondence. Motion correspondence information can be used for the tracking of individual objects or targets in presence of clutter, but also for steering the attentional selection process.

Given a large number of potentially very noisy channels in an ATR system, it seems beneficial to exploit the information from recognition of objects or their parts for the disambiguation of the correspondence computation in the individual channels. We are planning to test this idea of a mutual interplay between recognition and correspondence computation using the HMAX model as basis of a definition of a "complexity pyramid". Such a pyramid is similar to classical "resolution pyramids", with the difference that instead of the spatial frequency range the complexity of the features on the different levels varies. For the realization of this idea we are planning to extend the already described system in two ways: (1) We want to add mechanisms for the computation of optic flow (or motion energy) on multiple hierarchy levels of the recognition system. (2) We want to couple the different complexity levels in a way that biases the correspondence calculations on lower levels in a way that are consistent with already recognized objects and their movements on higher levels. The resulting system is conceptually similar to "coarse-to-fine" algorithms for the calculation of optic flow. However, information would be propagated from levels with higher feature complexity to levels with lower feature complexity, rather than between levels with different spatial resolution.

For the realization of couplings between levels heuristic methods can be used, like initialization of the correspondence calculation on lower levels of the pyramid with the results from the correspondence calculation on higher levels. Theoretically more sound approaches would be Bayesian approaches in combination with stochastic models for the description of the dynamics on individual levels of the hierarchy, like Blake's CONDENSATION. Another intuitively simple approach for the realization of couplings, which is consistent with biologically inspired system architecture described before, is to use lateral and top-down connectivity in neural networks to realize the coupling.

There are also exciting synergies between the probably biological implementation of the HMAX model and novel neural-like circuit components. We have shown in preliminary work that winner-takes-all behavior, as required to calculate the optic flow in ambiguous situations, and the MAX operation can be realized within the same simple recurrent neural network structure (only varying slightly a parameter, like the lateral coupling strength). We also have shown that feature tracking in distributed neural representations for high-level vision can be achieved by exploiting traveling pulse solutions is such networks. It seems, therefore, appealing to try if maximum pooling and optic flow calculation with top-down influences can be integrated within a simple canonical neural circuit.

A.4.6 A Software Architecture for ATR and Surveillance

We plan to demonstrate extensions of the MIT and CMU systems on ATR and surveillance applications. In particular, we will develop a software system that is adaptive to local variation in the target appearance, target pose, clutter, occlusion, and image quality. In their current form, both the MIT and the CMU classifiers always use the same statistical models, regardless of the local situation. In the case of the CMU system this corresponds to use the same $P_i(pattern_i, x, y|Object)$ and $P_i(pattern_i, x, y|NonObject)$. To make the systems adaptive to contextual cues we will develop methods of modifying the classifier in response to various contextual measurements. Below we explain how we plan to achieve such adaptability.

Variation in Target Appearance: Both classifiers can be updated on-line. For instance, the CMU system can adapt to local variation in the target by modifying $P_i(pattern_i, x, y|Object)$. Both classifiers can be updated by feeding detections as additional training data. Alternatively, we can also "boot-strap" the systems with hand-labelled detections. By combining prior and on-line statistics this way we can adapt to many forms of local variation: lighting, camera behavior, environmental properties, weather, etc.

Clutter and attentional signals: Both classifiers can adapt to clutter in various ways. SVM classifiers can adapt on-line in the same way suggested for the object distribution (see above). In the case of the CMU approach, the model for clutter is $P_i(pattern_i, x, y|NonObject)$. Pre-trained multiple special-purpose clutter models, $P_i(pattern_i, x, y|NonObject_k)$ could be prepared. For example, we propose to have one

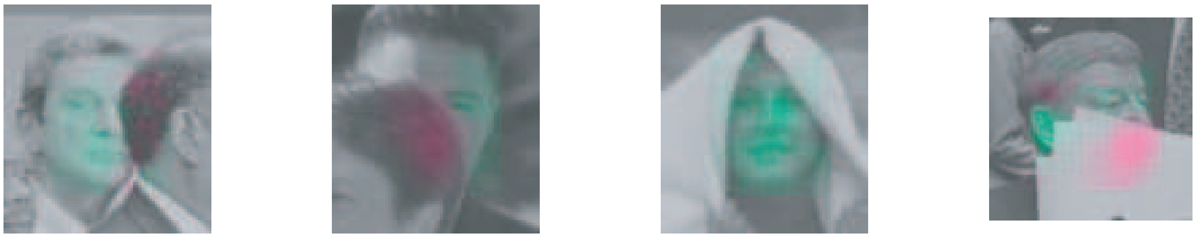


Figure 8: Parts-Based response to various faces

model for indoor (man-made) clutter and another for outdoor (natural) clutter. Before we search for the object, we will investigate whether to have a "meta-classifier" to the image. This meta-classifier would identify the form of clutter in the image. We would then automatically use this attentional signal to select the appropriate clutter model for the object classifier. A similar approach can be used for the SVM classifier with the advantage that density estimation is not needed and therefore a much smaller number of training examples will suffice.

Image Quality: In the same way we used a top-down signal to assess clutter, we could use one to assess image quality, focus, and other attributes. In the most simplistic model, the meta-classifier driving an attentional signal could control the threshold of the object classifier.

Occlusion: A components-based decomposition makes it possible to detect occlusion. Below we show how the CMU face classifier responds on a part-by-part basis to various input faces. Areas that are green indicate regions that are consistent with the presence of a face. Conversely, areas in red are not consistent with the presence of a face. In these examples, we can see a clear difference between visible parts of the face (contiguous green regions) and occluded areas (contiguous red regions). By taking into account this spatial distribution of the response, we can detect occluded regions and omit them from the classification decision. Similar results can be obtained by using the SVM component-based classifier and considering the margin – the real-valued output of the classifier – as a measure of confidence in the classification.

Object Pose: In any classifier there is a trade-off between generality and accuracy; that is, the narrower the range of variation in the object's appearance the easier it is to detect. This trade-off is particularly true with respect to generality in object pose. As we generalize a classifier to accommodate a greater degree of variation in pose, we will make a larger number of errors. A possible multistage strategy to cope with this trade-off will be tested at CMU. We will first apply a detector capable of accommodating significant variation in object size and pose. This detector will identify a set of candidates. We then apply an operator that estimates the pose of each candidate. Using the estimates we will normalize the pose of each candidate. (If the object is not present in the candidate normalization will have no meaningful effect.) Now we can apply a much more refined and accurate classifier designed to recognize the object over a much narrower range of variation. At CMU we have had success employing a similar strategy [53] to detecting objects with variation in in-plane rotation.

Testing: We plan to test such systems in a number of object detections problems. An example to be carried out at CMU is the detection of underwater mines using sonar imagery. This problem is clearly one that requires on-line adaptability. Mines can come in many shapes and sizes and the clutter can look very different depending on depth and other factors. Both at MIT and CMU we will also test these methods in vehicle detection problems using imagery of various types over a wide range of vantage points. Again, adaptability is essential. Vehicles can come in many shapes and sizes: passenger cars, buses, vans, trucks, tanks, etc. Lighting and weather influence their appearance. As well, the vantage point in which we view the vehicle affects its appearance. In addition, we plan (at MIT, CMU and USC) to test the systems on hyperspectral imaging data or available proxies of them, such as visual color image sequences. We hope to obtain data from the ONR/NRL HI-STAR project that is developing a high-resolution EO/IR (0.4-12 mm) hyperspectral imaging (10 to 40 nm spectral bands) airborne reconnaissance system. In addition to classification performance, the key challenges will be the ability to automatically detect in real time and to provide real-time adaptability to demands imposed by varying environmental conditions. Ultimately, at MIT we plan to test the systems on tasks of *scene interpretation*. We plan to try these tasks in the final stages of the project, exploiting the availability of multiple detectors, trained to many different types of objects. We conjecture that often a primitive parsing of the scene may actually be computationally easier than a binary detection problem!

A.5 Expected Outcome and Potential for DoD Applications

We expect:

- to design a software system for imagery interpretation building upon the MIT and CMU object detection systems, the HMAX model of object recognition in cortex and the CalTech model of bottom-up attention. The system will embed two significant extensions: on-line adaptability through top-down signals and task- and object-dependent learning of features;
- to design a real-time prototype of the SVM classifier with on-line incremental learning, capable of dealing with very high bit rates, exploiting a parallel, hierarchical architecture suggested by studies and models of the visual cortex;
- to demonstrate a software prototype of the system on one or more specific applications with data for assessment provided by ONR in the domain of surveillance or ATR, probably including hyperspectral imagery (1000 wavebands, 10 meter resolution) and underwater mine detection.

A.6 Performance Goals

Year I:

- We will implement hierarchical HMAX features in our SVM-based classifier and compare performance.
- 2. We will add bottom-up saliency-based mechanism to integrate information across spectral bands and focus processing on regions of interest. We will measure performance changes against a standard database.
- 3. We will benchmark at CMU and MIT our existing classification systems on proxies of hyperspectral imagery
- 4. We will compare different techniques of learning features from data.

Year II:

- 1. We will extend our saliency model to motion detection by implementing spatio-temporal energy model. We will implement it on parallel computer cluster and test it using artificial test sequences and standard video camera inputs.
- 2. We will experiment with and evaluate the addition of top-down attentional control signals in our classifiers.
- 3. We will begin development of an integrated system for information extraction from imagery data, starting to integrate saliency-based attention with HMAX object identification.
- 4. We will begin designing (at USC and JHU) and evaluating a software and hardware architecture for efficient implementation of on-line learning and for dealing with very high bit rate.

- 5. We will implement a multi-stage classifier designed to adapt to variation in the pose of the object.

Year III:

- 1. We will characterize at MIT and CMU the properties of the new trainable systems developed in the project in comparison with the more traditional ones that do not incorporate biological insights.
- 2. We will extend the saliency-based attention model to on-line adaptability, implementing algorithms to learn optimal interactions across feature channels, grouping, and task-dependent top-down biasing in a largely automatic manner (using the interplay between object recognition and attention).
- 3. We will fully implement a prototype system and test it on imagery into the visual spectrum. We will incorporate grouping into our saliency computation.
- 4. We will experiment with techniques for learning from partially labeled examples – which will be developed in a separate project – to improve the speed at which the system can adapt to new tasks and new types of sensor data.
- 5. We will develop methods to adapt to variation in clutter. We will implement a meta-classifier to categorize clutter and automatically select clutter models for object detection.

Year IV, V (optional):

- 1. We will design an integrated system comprising all the components developed in the first 3 years – attentional subsystem, feature learning, on-line adaptability, motion processing. The system will combine elements of the extended MIT and CMU object detection systems as well as the CalTech attentional module.
- 2. We will develop and test the system, implemented in software on a parallel cluster at USC, on hyperspectral imagery, on sonar images (for underwater mine detection) and on visual images (for vehicle detection). We will also prototype at JHU a real-time system, including parallel hardware.
- 3. We will experiment at MIT with *scene interpretation*, using detectors trained for different types of objects to parse a visual scene.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J Opt Soc Am [A]*, 2(2):284–299, Feb 1985.

[2] D. Beymer and T. Poggio. Image representations for visual learning. *Science*, 272(5270):1905–1909, June 1996.

[3] Heinrich H. Bulthoff and Shimon Edelman. Psychophysical support for a 2-d view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64, January 1992.

[4] C. Burges. A tutorial on support vector machines for pattern recognition. In *Data Mining and Knowledge Discovery*. Kluwer Academic Publishers, Boston, 1998. (Volume 2).

[5] G. Cauwenberghs and V. Pedroni. A low-power CMOS analog vector quantizer. *IEEE Journal of Solid-State Circuits*, 32(8):1278–1283, 1997.

[6] G. Cauwenberghs and T. Poggio. Incremental and decremental support vector machine learning. In *Adv. Neural Information Processing Systems (NIPS*2000)*, volume 14. Cambridge, MA: MIT Press, 2001.

[7] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing kernel parameters for support vector machines. *Machine Learning*, 2000. Submitted.

[8] V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. Wiley, New York, 1998.

[9] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

[10] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, 18:193–222, Rev 1995.

[11] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

[12] Shimon Edelman and Tomaso Poggio. Models of object recognition. *Current Opinion in Neurobiology*, 1:270–273, 1991.

[13] T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. A.i. memo, MIT Artificial Intelligence Lab., 1999.

[14] U. Franke, D. Gavrila, S. Goerzig, F. Lindner, F. Paetzold, and C. Woehler. Autonomous driving goes downtown. *IEEE Intelligent Systems*, pages 32–40, November/December 1998.

[15] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cyb.*, 36:193–202, 1980.

[16] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[17] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who? when? where? what? a real time system for detecting and tracking people. In *Face and Gesture Recognition*, pages 222–227, 1998.

[18] B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. A.I. memo 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2000.

[19] B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. Technical Report AI Memo 1687, CBCL and AI Lab, MIT, Cambridge, MA, 2000.

[20] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[21] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. In *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99), San Jose, CA*, volume 3644, pages 473–82, January 1999.

[22] L. Itti and C. Koch. Target detection using saliency-based attention. In *Proc. RTO/SCI-12 Workshop on Search and Target Acquisition (NATO Unclassified), Utrecht, The Netherlands, RTO-MP-45 AC/323(SCI)TP/19*, pages 3.1–3.10, June 1999.

[23] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[24] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual-attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–9, 1998.

[25] T. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proc. of the 7th International Workshop on Artificial Intelligence and Statistivcs*, 1998.

[26] R. M. Klein. Inhibition of return. *Trends Cogn Sci*, 4(4):138–147, Apr 2000. (ENG).

[27] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 916 1985.

[28] D. K. Lee, L. Itti, C. Koch, and J. Braun. Attention activates winner-take-all competition among visual filters. *Nat Neurosci*, 2(4):375–81, Apr 1999.

[29] B.W. Mel. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comp.*, 9:777–804, 1997.

[30] B.W. Mel and J. Fiser. Minimizing binding errors using learned conjunctive features. *Neural Comp.*, 12:247–278, 2000.

[31] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. Technical Report 326, mitmedia, 1995.

[32] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J.P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. AI Memo 1677, Massachusetts Institute of Technology, 1999.

[33] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virt ual examples. *Proceedings of the IEEE.*, 86(11):2196–2209, 1998.

[34] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc. Computer Vision and Pattern Recognition*, pages 193–199, Puerto Rico, June 16–20 1997.

[35] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *IEEE Workshop on Neural Networks and Signal Processing*, Amelia Island, FL, September 1997.

[36] C. Papageorgiou, T. Evgeniou, and T. Poggio. A trainable pedestrian detection system. In *Intelligent Vehicles*, pages 241–246, October 1998.

[37] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proceedings of the International Conference on Computer Vision*, Bombay, India, January 1998.

[38] C.P. Papageorgiou. Object and Pattern Detection in Video Sequences. Master's thesis, MIT, 1997.

[39] H.E. Pashler. *The Psychology of Attention*. MIT Press, Cambridge, MA, 1998.

[40] D.I. Perrett, J.K. Hietanen, M.W. Oram, and P.J. Benson. Organization and functions of cells responsive to faces in the temporal cortex. *Philos. Trans. Roy. Soc. B*, 335:23–30, 1992.

[41] J.C. Platt. Fast training of support vector machines using sequential minimum optimization. In Schölkopf, Burges, and Smola, editors, *Advances in Kernel Methods– Support Vector Learning*, pages 185–208. Cambridge, MA: MIT Press, 1998.

[42] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), September 1990.

[43] T. Poggio and F. Girosi. A Sparse Representation for Function Approximation. *Neural Computation*, 10(6), 1998.

[44] T. Poggio and T. Vetter. Recognition and structure from one 2D model view: observations on prototypes, object classes and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[45] Tomaso Poggio and Shimon Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[46] M. I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D. G. Bouwhuis, editors, *Attention and Performance X*, pages 531–556. L. Erlbaum, Hilldale, NJ, 1984.

[47] M. Riesenhuber and T. Poggio. Models of object recognition. In press, Nature Neuroscience.

[48] M. Riesenhuber and T. Poggio. Just one view: Invariances in inferotemporal cell tuning. In *Advances in Neural Information Processing Systems 10*. San Mateo, CA: Morgan Kaufmann Publishers, 1998.

[49] M. Riesenhuber and T. Poggio. Are cortical models really bound by the "Binding Problem"? *Neuron*, 24:87–93, 1999.

[50] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2:1019–1025, 1999.

[51] K. Rohr. Incremental recognition of pedestrians from image sequences. *Computer Vision and Pattern Recognition*, pages 8–13, 1993.

[52] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *pami*, 20(1), 1998.

[53] Henry Rowley, Shumeet Baluja, and Takeo Kanade. Rotation invariant neural network-based face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 1998.

[54] R. Sarpeshkar, J. Kramer, G. Indiveri, and C. Koch. Analog VLSI architectures for motion processing: From fundamental limits to system applications. *Proceedings of the IEEE*, 84(7):969–987, 1996.

[55] Henry Schneiderman. *A Statistical Approach to 3D Object Detection Applied to Faces and Cars*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2000.

[56] Henry Schneiderman and Takeo Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '98)*, pages 45–51, July 1998.

[57] Henry Schneiderman and Takeo Kanade. A statistical model for 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2000.

[58] T. Serre, B. Heisele, S. Mukherjee, and T. Poggio. Feature selection for face detection. Ai memo, Massachusetts Institute of Technology, 2000.

[59] A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378(6556):492–6, Nov 1995.

[60] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992.

[61] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE PAMI*, 20:39–51, 1998.

[62] K-K. Sung and T. Poggio. Example-based learning for view-based human face detection. In *Proceedings from Image Understanding Workshop*, Monterey, CA, November 1994.

[63] M.J. Tarr. News on views: pandemonium revisited. *Nat. Neurosci.*, 2:932–935, 1999.

[64] S.J. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

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

[66] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognit Psychol*, 12(1):97–136, Jan 1980.

[67] S. Treue and J. C. Martinez Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, Jun 1999.

[68] S. Ullman and E. Sali. Object classification using a fragment-based representation. In S.-W. Lee, H.H. Bülthoff, and T. Poggio, editors, *Proceedings of BMCV2000*, volume 1811 of *Lecture Notes in Computer Science*, pages 73–87, New York, 2000. Springer.

[69] R. Vaillant, C. Monrocq, and Y. Le Cun. Original approach for the localisation of objects in images. *IEEE Proc. Vis. Image Signal Process.*, 141(4), August 1994.

[70] V. Vapnik. *The nature of statistical learning.* Springer Verlag, 1995.

[71] V. Vapnik. *Statistical learning theory.* John Wiley and Sons, New York, 1998.

[72] V. Vapnik and O. Chapelle. Bounds on error expectation for SVM. In Smola, Bartlett, Schölkopf, and Schuurmans, editors, *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 2000.

[73] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

[74] G. Wahba. Smoothing and ill-posed problems. In M. Golberg, editor, *Solutions methods for integral equations and applications*, pages 183–194. Plenum Press, New York, 1979.

[75] G. Wahba. Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. In J. Ward and E. Cheney, editors, *Proceedings of the International Conference on Approximation theory in honour of George Lorenz*, Austin, TX, January 8–10 1980. Academic Press.

[76] G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized splines smoothing problem. *The Annals of Statistics*, 13:1378–1402, 1985.

[77] G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.

[78] M. Weber, W. Welling, and P. Perona. Towards automatic discovery of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000. In press.

[79] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *to appear NIPS'00*, 2000.

[80] W. Wickelgren. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psych. Rev.*, 76:1–15, 1969.

[81] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. Technical report, MIT Media Laboratory, 1995.

B Facilities and Management

We plan to use available equipment in our five institutions, where we have a number of reasonably usable workstations and other computing facilities that are sufficient for the proposed project. To optimize our work we apply for the equipment option with the goal of acquiring state of the art Linux workstations and one parallel machine for each team configured in exactly the same way so to allow the exchange of compatible parallel software – of which we describe the development in the proposal – with minimal effort. Because using Beowulf clusters of separate computers requires significant maintenance and system administration expertise, we will opt for a more manageable solution consisting of deploying single high-end multi-CPU machines in all institutions. The focus of the proposed project at MIT will be housed in the existing Center for Biological and Computational Learning, building E25-201. CBCL can also make use of the resources of the Artificial Intelligence Laboratory (located in building NE43). As an example of the computational facilities available in our five institutions here is partial list of the CBCL facilities: (10) Dell Precision III 610 Workstations, (2) Dell Precision III 620 Workstations, (3) Dell Precision II Workstations, (6) Dell Dimension XPS MiniTowers, (1) Dell Precision 410, (2) Dell OptiPlex GXI, (1) Dell Dimension XPS Pro200 and (3) Pentium PC) running LINUX and/or NT operating systems, (2) Sun-SparcUltras, (4) Sun-Sparc-Sta-20 running the UNIX operating system, (1) 16-cpu Paralogic Xtreme machines and (2) 4-cpu Xtreme machines.

For research oversight, we will use the model for the Center for Biological and Computational Learning (CBCL), successfully time-tested over the past several years. We will manage the research effort through a steering committee of the PIs and co-PIs. We will have, in addition to normal meetings between faculty and students:

- Exchange of students between MIT, CMU, Caltech, USC and Hopkins to help integrate the different software components of the systems;
- Quarterly meetings of the project leaders (Poggio, Koch, Kanade, Cauwenberghs, Itti).
- Weekly meetings with discussions and informal talks among the faculties and graduate students.
- One internal workshop summarizing every six months (once a year also for the sponsors) the progress of the project and plan the next steps.
- The "Brains and Machines" seminar series at MIT, focusing on applications of learning technologies. The Center for Biological and Computational Learning (at BCS) and the Artificial Intelligence Lab will jointly organize this seminar series.

C Impact on the University, Education, Students and Training

The research funding will allow us to undertake a project which otherwise might not have been possible. Without this funding we would not be able to focus on the applicability of the object classification model to scenarios of interest to DoD. Over the last few years, the center of gravity of our research – for instance in our lab at MIT – has shifted away from DoD projects mainly because of the funding environment. Thus, this project would not only contribute substantially to graduate education—since the bulk of our expenses will go towards supporting students—but will also contribute substantially to educate them on broad problem areas of long-term interest to the DoD. The funding would especially enhance our ability to attract graduate students interested in exploring both the purely scientific and also the applied aspects of research in computational vision. Our project includes a significant educational component. We plan to involve 16 graduate students in the project from 5 institutions and several Departments, ranging from Computer Science to Neuroscience. For them this will be the main project in the course of their PhD. Some of the graduate and undergraduate courses in Machine Learning, Computer Vision, and Neuromorphic VLSI that we teach at MIT, CMU, Caltech, USC and Hopkins will also benefit and be directly influenced by the research proposed here and in turn influence students beyond the ones involved directly. In addition, we will disseminate results with technical reports, journal papers and conference participation by PIs and students.

C.1 Training of Students

The award will support and help training 1 postdoc to supervise the project (equivalent to 2 graduate students), 6 graduate students at MIT, 2 students at CMU, 2 students at CalTech, 2 students at USC and 2 students at JHU. Eight faculty members throughout the five institutions will supervise these students.

D Subawards, Role of Collaborating Institutions and Technology Transfer to DoD

The composition of our team was dictated by the structure and the components of the project. Some members of our team are experts in applications of statistical learning theory (Poggio, Cauwenberghs), in image detection and classification (Schneiderman, Kanade, Heisele, Grimson, Itti), in computational neuroscience (Koch, Itti, Riesenhuber, Poggio, Giese, Sinha). In addition, we have experts in new models of computation to develop algorithms that could deal with very high speed and thruput, using neural-like hardware (Cauwenberghs, Koch, Itti). More specifically, the original HMAX model was developed at MIT-BCS (Riesenhuber and Poggio) and the SVM applications to object detection were done in AI+BCS (Papageorgiou and Poggio). The model of visual attention that suggests how to include top-down signals was developed at CalTech and also applied to image processing systems (Koch, Itti). The other successful system for object detection was developed at CMU (by Schneiderman and Kanade). For the extension to learning features in a task dependent way we need a psychophysicist and computer vision expert who has contributed novel insights in exactly this problem area (Sinha). We added expertise in computer vision, reconnaissance and especially ATR (Grimson). Expertise in physiology is available to us through an independent NSF funded project (E. Miller, with whom we collaborate on physiological experiments). The team members have previously collaborated and also competed with each other successfully on several occasions. They are highly committed to working together on this project and accomplishing the stated objectives.

The division of tasks is as follows. MIT and CMU will extend their object detection systems as described in the proposal, following some of the biological insights. CalTech and USC will extend their attentional module and provide software (year II and III) to be integrated with the MIT and CMU systems. USC will develop parallel implementations on clusters of processors of some of the modules and make software available to the MIT and CMU clusters. JHU will design hardware system and prototype them in the final two optional years, depending on the results of the tests. Throughout the project, components of the overall system will tested on real data and will be made ready for applications and technology transfer. The fourth and fifth year will see an integration effort and tests of the integrated system.

Our project will focus on basic science and engineering research. Technology transfer for defense and commercial purposes will take place through the educational component of our project and the various mechanisms described earlier for disseminating the results of the research. In our experience, this broad process is very effective in the long term, especially in its educational component: the best technology transfer is through people. However, we plan to put in place a more specific mechanism for transferring directly our most relevant results to DoD applications, such as Hyperspectral imaging for Surveillance. The plan is to establish a linkage with an organization that transitions research findings to DoD application. Our first choice is Alphatech because of its record of achievements, because is located close to MIT and because several of us have worked with Alphatech in the past. Our main technical contact there is Dr. Gil Ettinger, who has stated their readiness to cooperate as needed. In addition, we will be able to transfer relevant technology to robotics application through interactions and collaborations with Rod Brook's group and his Cog project at MIT.

E Involvement of PI in Other Research Projects

The PI Tomaso Poggio and his group (total about 20 people of which 12 are graduate students) are involved in several projects listed in the section on Personnel. The main ones are: an NSF project entitled, "Learning and Representations of Objects and Object Classes in Visual Cortex" which has very useful synergies but no overlap with the present project; another NSF project titled "From Bits to Information: Statistical Learning Technologies for Digital Information" which does not have overlap in application domain or goals but some possible synergies in terms of the statistical learning theory underlying the algorithms (for instance we will try to develop techniques to train with partially labeled examples in this NSF project; if successful such algorithms will be used in the present MURI proposal in one of the last years). The PI is also involved (not as PI; PI is Dr. Eric Grimson) in a DARPA funded project titled "The iEyes: An Integrated Vision Network for Identification, Indexing and Intruder": there is no overlap here. There is also no overlap with all other on-going projects. Pending proposals include a proposal submitted to NSF titled "Trainable Visual Aids for Object Detection and Identification" (not as PI) where there is some overlap in the goals and there are synergies in the statistical learning algorithms; a project submitted to NASA titled "Neural Computing for Classification of Temporal Patterns" where again there is overlap in goals and there are synergies in the underlying algorithms (if both grants are awarded we will work out with the funding agencies how to organize the parts of the projects and their funding; the MURI project has for us the top priority). For all other pending proposal there is no overlap at all.

F Other Parties Receiving this Proposal

None.