COMMON COMPUTATIONAL STRATEGIES IN MACHINE AND BIOLOGICAL VISION

Maximilian Riesenhuber and Tomaso Poggio
Center for Biological and Computational Learning and Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, E25-201, Cambridge, MA 02139
{max,tp}@ai.mit.edu

ABSTRACT

Humans effortlessly perform visual tasks, like face detection and recognition, that have turned out to be extremely difficult for artificial vision systems. Research into the principles underlying biological object recognition therefore is of great potential interest to the machine vision community. Recent experiments — both psychophysical and physiological — have provided evidence that instead of using 3D models as in many "classical" computer vision systems, a biological strategy for object recognition is view-based, in the sense that objects are represented by a set of 2D views. Several models have presented plausible mechanisms explaining how this "snapshot"-based representation can develop and how it can be used to perform object recognition. Extending the view-based paradigm to artificial systems has proved to be an extremely successful approach to developing robust and flexible computer vision systems that are able to perform reliably in real-world situations. Among others, we present applications to object detection, face recognition, and pose estimation and synthesis.

1 INTRODUCTION

Vision is a crucial component in human-artifact interaction. Unfortunately, many vision task that we humans perform effortlessly have turned out to be extremely challenging for machine vision systems. Examples include face detection and recognition or the estimation of facial expressions (along with the inverse synthesis tasks of generating realistic images). We turn to the visual system for clues on how to solve these problems. Figure 1 shows a diagram of the macaque brain (that has a visual system very similar to that of man) and its ventral visual stream, the "where"-pathway which appears to mediate form perception [25]. Visual signals are converted to electric impulses in the retina and passed on to the lateral geniculate nucles (LGN) in the

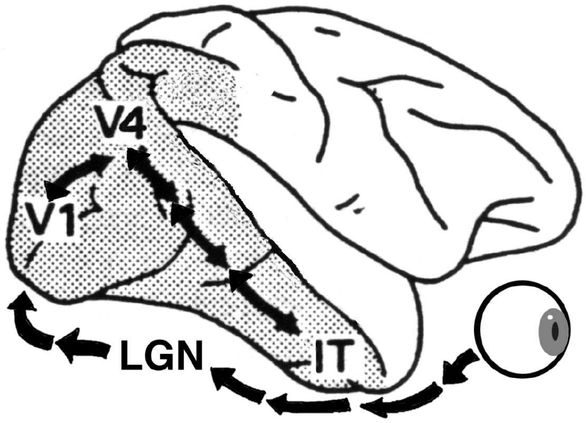

Figure 1: Sketch of the ventral stream of the macaque visual system. Adapted from [25].

thalamus and further to the cerebral cortex. The cortical ventral visual pathway exhibits cells responding selectively to increasingly complex objects while showing response invariance to an increasingly greater range of transformations such as translation, size changes or rotation in depth. While cells in the primary visual cortex, V1, respond preferably to simple stimuli like oriented edges and bars, in the inferotemporal cortex, IT, studies (see [19] for a review) have found cells that respond to complex objects [12], such as faces, even when the stimulus undergoes translation by several degree of visual angle, changes its size over several octaves or is rotated in depth by a certain amount [11]. These cells show a view-tuned response, *i.e.,* they respond maximally to a certain view of an object and fire progressively less as the viewpoint (and correspondigly the 2D image on the retina) changes.

Recently, Logothetis *et al.* [13] have shown that the stimulus-specificity of these neurons can be influenced through learning. Having trained monkeys to recognize paperclip-like objects using just one view of each object,

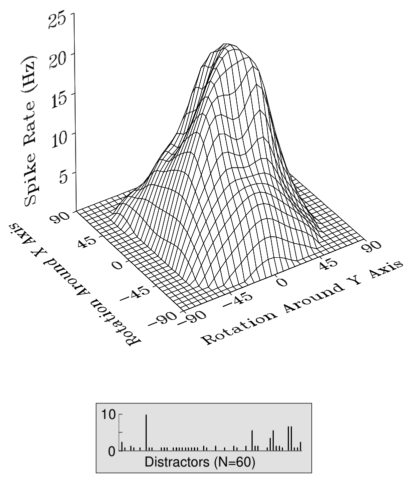

Figure 2: Response of a view-selective neuron for stimulus rotations around four axes. The x- and y-axes of the plot show the degree of rotation from the preferred view, the z-axis shows the response of the cell to the corresponding rotated view. The plot was generated by measuring the response of the cell as the stimulus was rotated first separately around the X and Y axes, resp., and then along the two diagonal axes. The response of the cell declines as the stimulus is rotated out of the preferred view. The inset shows the cell's response to 60 distractor objects, which were other randomly generated paperclip objects (from [13]).

they report neurons in IT that are sharply tuned to views of the training objects but not of similarly shaped distractor objects. Interestingly, the cells also exhibit tolerance to unseen transformations of the stimulus: Logothetis *et al.* report translation invariance (invariance being defined as producing a stronger response to the transformed object than to any of the distractor objects) over $\pm 2^\circ$ of visual angle, scale invariance over $\approx \pm 1$ octave and invariance to rotation in depth over 30°. Figure 2 shows the response of one such cells to stimulus rotations in depth. The response gradually falls off as the stimulus is rotated away from the preferred view.

This provides a strong constraint for any model investigating the view-tuning of IT neurons: Training the system with just one view leads to a finite, but not complete response invariance to image transformations such as translation, size changes and rotation in depth. Several models that have been proposed to explain biological view-invariant object recognition [8, 17] predict either no or complete response invariance, resp., from one training view.

In summary, while a possible approach to vision could be to employ hard-wired, task-specific systems, *e.g.,* for face recognition, physiological (and psychophysical, see, *e.g.,* [5]) experiments seem to indicate that this is not the solution chosen by the visual system. The discovery of view-tuned cells points to the visual system using a view-, *i.e.,* example-based representation that is shaped by visual experience. This is what endows the biological system with the robustness and flexibility needed to survive in a continuously changing environment.

The view-based hypothesis of biological object recognition is further supported by computational models, which we will describe in the next section, that show how view-invariant object recognition can be carried out in the neuronal framework. Finally, the view-based approach can be extended to problems in computer vision. Section 3 will present some examples.

2 MODELING THE VIEW-BASED REPRESENTATION IN THE VISUAL SYSTEM

The finding that training a monkey with just one object view leads to finite transformation tolerance was anticipated by a first model by Poggio and Edelman [20] that showed how a 2D view-based approach could be employed to perform recognition of 3D objects which were rotated in depth. Subsequently, a model by Bricolo and Poggio [4] moved away from the abstract feature-based stimulus representation used in [20] by working directly on gray-level images.

While the two previous models only investigated tolerance to rotation in depth, a recent model by Riesenhuber and Poggio [23] also investigated tolerance to translation and size changes, and further demonstrated how view-tuned neurons could self-organize in a stimulus-driven process. Figure 3 shows a cartoon of the model: A retinal input pattern leads to excitation of V1 cells, here abstracted as having derivative-of-Gaussian receptive field profiles. The V1 cells densely sample the retinal input pattern and have relatively small receptive fields [9]. In the present version of the model V1 cells have only one size. The V1 cells of the same type (*i.e.,* having the same preferred stimulus) feed into an intermediate layer of neurons. These intermediate neurons could be complex cells in V1 or V2 or V4 or even posterior IT: for simplicity we label them as "V4" cells (see figure). Thus, a V4 cell receives inputs from V1 cells over a large area ([12] reports an average receptive field size in V4 of 4.4° of visual angle). These "V4" cells in turn feed into a layer of "IT" neurons.

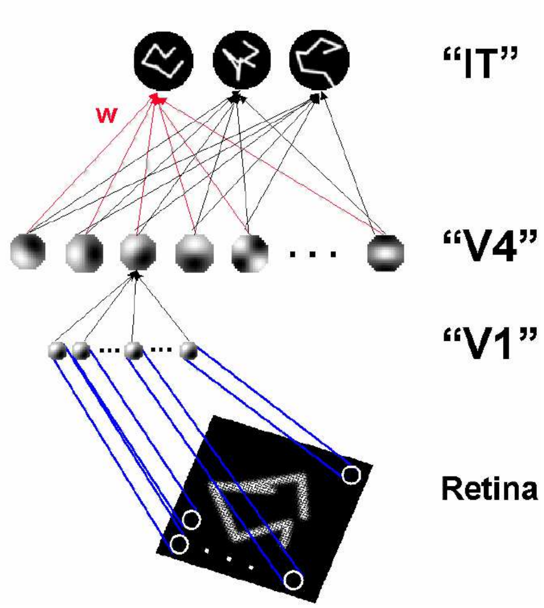

Figure 3: Cartoon of the model of Riesenhuber and Poggio. See text for explanation.

From the computational point of view the intermediate neurons should (i) measure the presence of specific features without being confused by clutter and context in the receptive field; (ii) maintain information about stimulus geometry, *i.e.,* position of the feature.

These requirements may be compatible with recent data. John Reynolds (Reynolds, personal communication, see also [16]) has gathered data that show that when two stimuli are introduced into the receptive field of a V4 cell, then the cell's response tends to be closer to the stronger one of the responses caused by each individual stimulus than to the sum of the two individual responses. This is in line with other results on interactions of multiple stimuli in V4 receptive fields [15]. In V4, the receptive fields are comparatively large, while the preferred stimulus is usually much smaller [6]. Interestingly, cells respond independently of the location of the stimulus within the receptive field, but in general response strength varies with stimulus location [6].

In the model, we implemented the computational requirements and the physiological constraints by having V4 neurons perform a "Winner-Take-All"-operation within their receptive fields, *i.e.,* having the response depend on the best match of preferred stimulus and the activaty in the neuron's receptive field. Moreover, V4 cell activation also depended on the location of the best match within the receptive field through a multiplicative modulation. Hence, the neuronal response is determined by the product of two components: degree of similarity to the preferred stimulus and stimulus position in the receptive field.

For every intermediate cell, the receptive field covered the whole 64×64 input patch, accounting for the fact that stimuli in the experiment covered between 0.5° and 3.6°[13], well below the average receptive field size of 4.4°[12].

The response function of top layer neurons was Gaussian, with the center being determined by the afferent weights $\mathbf{w}$ (cf. Fig. 3) of each cell from the middle layer. Riesenhuber and Poggio [23] demonstrate (using the EM algorithm for a mixture of Gaussians network [3]) that these weights can be learned in an unsupervised manner driven by stimuli consisting of views of different paperclips to yield view-tuned neurons.

The stimulus images were views of 21 randomly generated "paperclips" of the type used in the physiology experiment [13]. Distractors were 60 other paperclip images generated by the same method. To test size invariance, images containing paperclips of half the standard size (in both dimensions) were used.

The left plot in Fig. 4 shows the tuning curve for rotation in depth of one model neuron tuned to the 90°-view of a particular paperclip shown over the 90°position of the plot. The right panel shows the responses to the 60 distractor objects. We see that the model neuron shows tolerance to transformations comparable to that of IT cells. The model neuron also showed invariance to translations of ± 12 pixels from the center position. The averages for 21 model cells trained with one view of randomly selected paperclips were $38°$ of rotation invariance, size invariance with respect to the small stimulus in all but three cases, and translation invariance over 24×24 pixels.

2.1 Unsupervised Learning of View-Invariant Neurons

While the previous model showed how view-tuned cells can arise in a biological system, it left open the question of how these view-tuned neurons can be bound together to yield a unit that responds to the presence of an object in the visual field in a view-invariant manner (such cell have also been reported in IT [13]). Riesenhuber and Dayan [22], in the context of a model that learns to represent part-whole relationships of objects (as found in, *e.g.,* a face, which contains eyes, a mouth and a nose as parts), proposed a mechanism of how the hierarchical relationship between cells coding for individual objects in a view-independent manner and cells

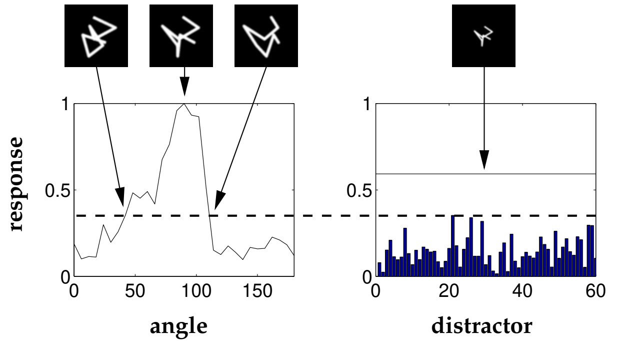

Figure 4: Responses of a sample top layer neuron to different views of the training stimulus and to distractors. The left plot shows the rotation tuning curve, with the training view (90° view) shown in the middle image over the plot. The neighboring images show the views of the paperclip at the borders of the rotation tuning curve, which are located where the response to the rotated clip falls below the response to the best distractor (shown in the plot on the right). The neuron exhibits broad rotation tuning over more than 60°. Moreover, the neuron is invariant to a decrease of the stimulus size by a factor of two, as the right plot shows. There, the thin horizontal line gives the response level to the small stimulus, which is well above the responses to any of the distractors (from [23]).

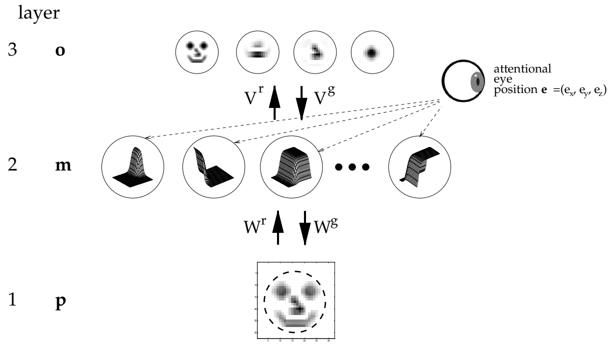

Figure 5: Cartoon of the model. In the top-down, generative, direction, the model generates images of faces, eyes, mouths or noses based on an attentional eye position and a selection of a single top-layer unit; the bottom-up, recognition, direction is the inverse of this map. The response of the neurons in the middle layer, which can be identified with the view-tuned neurons in IT, is modulated sigmoidally (as illustrated by the graphs shown inside the circles representing the neurons in the middle layer) by the attentional eye position. From [22].

coding for individual views of those objects can be learned in an unsupervised fashion. The overall scheme is based on an hierarchical top-down synthesis and bottom-up analysis model for visual processing, as in the Helmholtz machine [10]. The model ascribes a crucial role to the strong feedback connections that reciprocally connect modules in the ventral visual stream of the macaque [7], which form a synthesis pathway in the model (enabling a form of imagery by activating object-specific units in IT). The statistical inverse of the synthetic model is used to analyze images to determine what objects they depict. Figure 5 shows a diagram of the model with its two pathways. While the modulation by the attentional eye position shown there is vital to the part-whole representational mechanism suggested in [22], the crucial element for the unsupervised learning of view-invariant neurons is the combination of analysis and synthesis pathways that allows the system to find an efficient representation of the stimulus space by having the top layer units develop into object-specific but view-invariant neurons.

Starting from random initial connecting weights, the network was trained using a variant of the wake-sleep algorithm [10]. There, learning consists in repeated alternations of wake cycles, in which stimuli in the input layer lead to activations of higher-layer neurons via the bottom-up pathway, which are then used to train the top-down connections, and sleep cycles, in which the bottom-up model is learned through generated imagery via the top-down synthesis pathway. After training, neurons in the top layer become tuned to different objects (a face and its parts, resp., in the example in [22]) and respond to it independently of position (with position information given by the attentional eye position), even though middle layer units only show very limited translation invariance.

While these models are still at the level of a plausibility proof rather than a detailed physiological model, they establish that a 2D view-based representationas found in the experiment provides a possible substrate for view-invariant 3D object recognition.

3 APPLICATIONS TO MACHINE VISION

In this section, we will briefly illustrate by several examples from recent work of our group how the view-based approach to object recognition can be extended to machine vision. While the examples are from quite different areas of computer vision and graphics, they all have in common that they use 2D example-based object representations without any explicit 3D models.

3.1 Object Recognition

An obvious application of our learning-from-examples framework is object recognition. The idea is simply to train the learning module with a few views of the object to be recognized and the corresponding label (as output). One interesting class of objects for recognition are human faces. Over the last several years several computer vision groups have used this view-based paradigm to develop and demonstrate successful face recongition systems. In our own group at MIT we have developed systems for face verification for secure login and a pose-invariant recognition system. The latter [1] uses 15 views per person, views that cover different out-of-plane rotations. Translation, scale, and image-plane rotation are factored out by first detecting eyes and nose features and then using these features to register the input with model views. The step of localizing eyes and nose features uses a bank of eyes-nose templates covering a variety of "exemplar" people and poses. A new input is matched against all model views of all people and the best match is reported. The matching step consists of a geometrical registration step driven by the eyes and nose features followed by correlation using eyes, nose, and mouth templates. On a small data base of 620 images (62 people, 10 test views per person), the recognition rate was 98%. The test images were at a variety of image rotations, both in and out of the image plane.

3.2 Object Detection

Even more difficult than recognizing an isolated specific object is detecting an object of a certain class in a cluttered image. Compare, for instance, the human task of recognizing a specific face with the task of noticing whether a cluttered image contains a face at all, and then locating it. Sung and Poggio [24] have presented a system that can be successfully trained to find faces, eyes, and other objects.

Figure 6 shows an example of the performance of the system after it has been trained to locate frontal images of faces. Key to the system is the large number of virtual examples, which are rotated versions of real face examples, used to train the system: there are 3000 virtual examples of face images in addition to the 1000 real examples and to the 50,000 negative examples of non-face patterns.

More recently, Oren *et al.* [18] have developed a system for the more difficult problem of pedestrian detection using a similar example-based approach.

3.3 Estimating Pose and Expression

If a learning network is to be trained to estimate arbitrary poses and expressions of, say, a person, a set of example

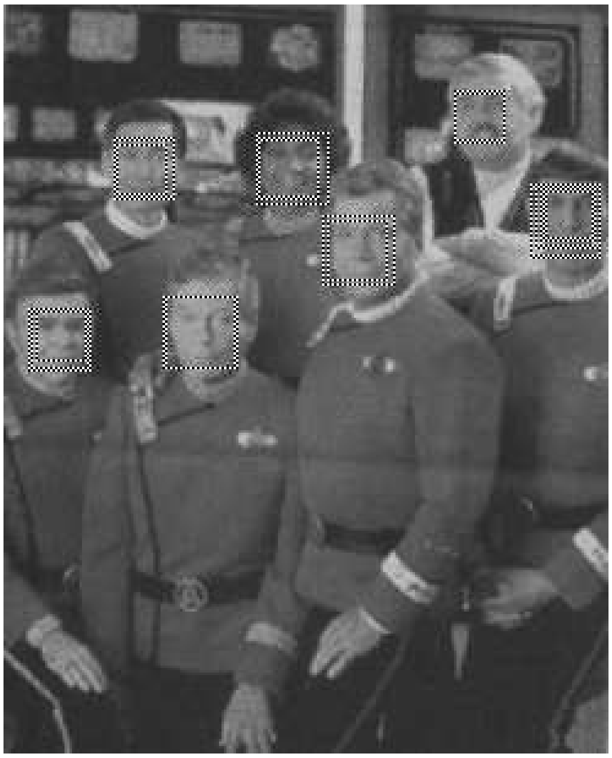

Figure 6: Some face detection results by Sung-Poggio system. From [24].

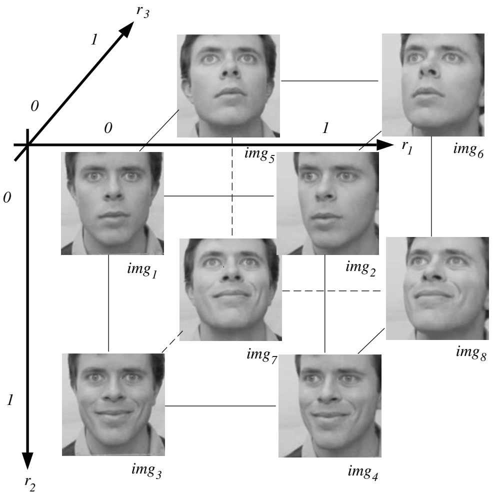

Figure 7: In our demonstrations of the example-based approach to image analysis and synthesis, the example images img_1 through img_8 are placed in a 3D rotation-expression parameter space (r_1, r_2, r_3), here at the corners of the unit cube. For analysis, the network learns the mapping from images (inputs) to parameter space (output). For synthesis, we synthesize a network that learns the inverse mapping, that is the mapping from the parameter space to images. The three axes of the parameter space correspond to smile, rotation left-to-right and rotation up-down. From Beymer and Poggio (unpublished).

images of that same person must be labeled in pose and expression space (Fig. 7). The input to the network, or our image representation, is not the image itself but the associated *shape vector* that is computed from the image. The shape vector can be thought of as a representation of (x, y) locations of facial features. The output in this case are three parameters, representing the amount of smile and rotation (left-right and up-down) of the face. A regularization network is trained from eight example images, shown in Fig. 7. After "training", the network can "generalize" to new images of the same person, estimating the associated rotation and smile parameters. Examples of input images and their estimated parameters are shown in Fig. 8.

3.4 A Computer Graphics Task: Synthesizing Novel Images

Beymer, Shashua and Poggio [2] used an example-based technique for generating images of non-rigid 3D objects as a function of input parameters. This image synthesis task – a computer graphics task – is the inverse of the problem of image analysis discussed in the previous section.

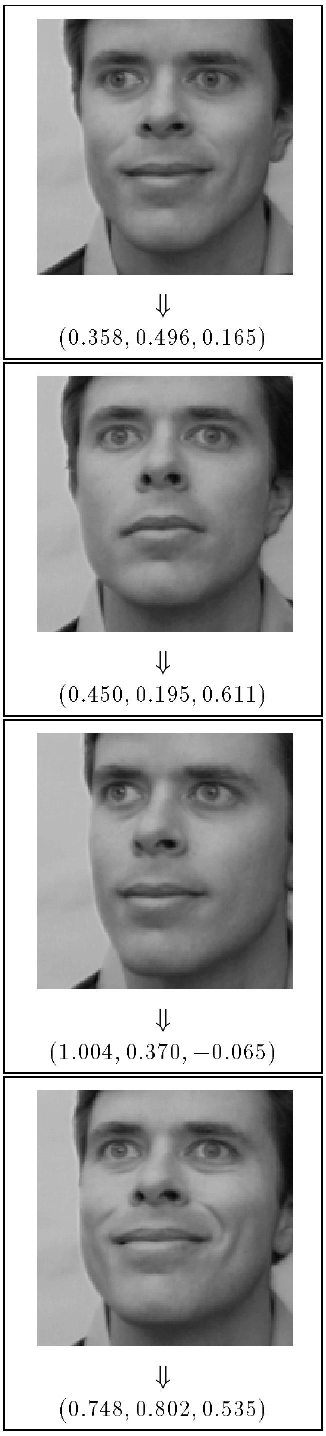

Figure 8: The analysis network trained on the eight examples of Fig. 7 estimates the two rotation and expression parameters for a novel image. The figure shows the network output (r_1, r_2, r_3) for four novel images that are different from the examples in the training set.

Here we will demonstrate the technique on real images of faces with input parameters such as pose and facial expression, but other parameters may also be used. The first step is to find pixelwise correspondence in order to compute the shape vector representation of each example image. This process of vectorization is done automatically using so-called optical flow algorithms that provide for each pixel in one image its (x, y) displacement relative to another image, chosen as the reference image. Once the example images are vectorized, the learning stage can take place. To train the synthesis network, the user associates each example image to an appropriate position in the parameter space, uses a multidimensional regularization network to synthesize the mapping between inputs and outputs, and displays (renders) the vectorized image as a grey-level or color image.

After the learning stage the "trained" network can synthesize a new image for any reasonable pose input vector given to the network. The synthesis network has typically a small number of inputs – the pose parameters – and a large number of outputs, each corresponding to one attribute of either shape or texture of a pixel in the image.

Figure 7 shows eight examples of a face and their assigned pose parameters in a 3-dimensional input space. Figure 9 shows some pairs of (r_1, r_2, r_3) parameters and the image produced by the synthesis module.

Image synthesis as done by the network is equivalent to performing multidimensional interpolation or approximation, in this case a form of multidimensional morphing.

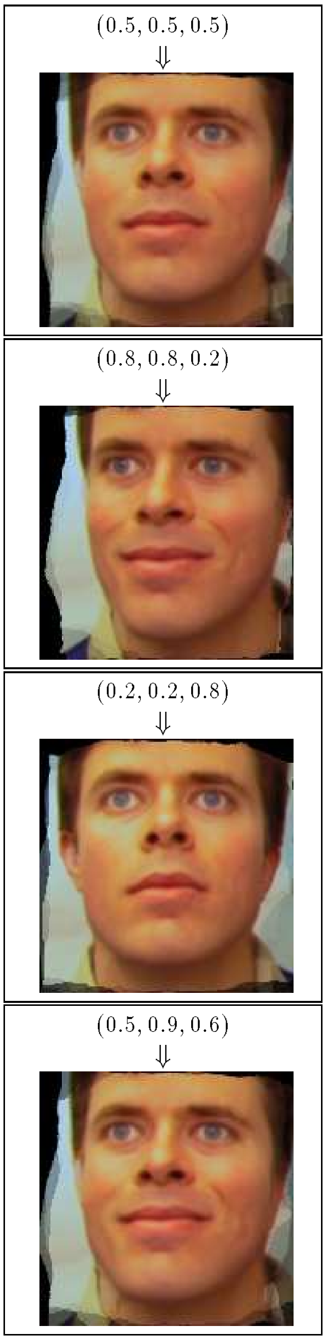

Figure 9: The synthesis network trained on the eight examples of Fig. 7 synthesizes new color images from the input rotation and expression parameters (r_1, r_2, r_3).

4 CONCLUSION

In this paper, we started out by presenting evidence that the visual system is using a view-based object representation. We presented models that establish the feasibility of the view-based approach to object recognition in the neural framework. Extending the example-based paradigm to applications in machine vision entails formulating the problem at hand in terms of training a learning box with appropriate sets of examples, *i.e.,* input-output pairs, with the underlying theory making use of function approximation techniques, neural network architectures, and statistical methods. We presented systems that learn to recognize objects, in particular faces, systems that learn to find specific objects in cluttered scenes, and algorithms that learn to synthesize novel image sequences from a few real pictures and thereby promise to achieve extremely high compression in applications such as video conference and video e-mail. In all of these applications, the notion of *learning* instead of *programming* the task is key.

There are challenging problems still to be solved. In the biological realm, we have presented models that perform recognition of *individual* objects. A key problem is how population codes for *classes* of objects can be learned: IT seems to code for different classes of objects using a sparse code over different cell populations [28, 29]. We currently investigate possible models of learning such a representation. In machine vision, we are faced with a related problem: In representing a class of objects by a linear combination of prototypes [27], how can the prototypes be automatically selected from the data to yield a compact yet complete representation? Another question concerns finding a good partitioning of the input space for classification. Here, stimuli at the class boundaries are most important. Algorithms that exploit this insight, like the support vector machine [26], have been used with great success in applications [18, 26]. Possibly, these algorithms might help us to understand how the brain performs pattern classification.

References

[1] Beymer, DJ (1994). Face Recognition Under Varying Pose. In *Proceedings of CVPR 1994*, 756-761.

[2] Beymer, DJ, Shashua, A & Poggio, T (1993). Example Based Image Analysis and Synthesis. AI Memo 1431. Cambridge, MA: MIT.

[3] Bishop, CM (1995). *Neural networks for pattern recognition.* Oxford University Press.

[4] Bricolo, E, Poggio, T & Logothetis, N (1997). 3D object recognition: A model of view-tuned neurons. In *Advances In Neural Information Processing* **9**, 41-47. MIT Press.

[5] Bülthoff, H & Edelman, S (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Nat. Acad. Sci. USA* **89**, 60-64.

[6] Desimone, R & Schein, S (1987). Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *J. Neurophys.* **57**, 835-868.

[7] Felleman, DJ & van Essen, DC (1991). Distributed Hierarchical Processing in the Primate Visual Cortex. *Cerebral Cortex* **1**, 1-47.

[8] Földiák, P (1991). Learning invariance from transformation sequences. *Neural Computation* **3**, 194-200.

[9] Foster, KH, Gaska, JP, Nagler, M & Pollen, DA (1985). Spatial and temporal selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *J. Phy.* **365**, 331-363.

[10] Hinton, GE, Dayan, P, Frey, BJ & Neal, RM (1995). The wake-sleep algorithm for unsupervised neural networks. *Science* **268**, 1158-1160.

[11] Ito, M, Tamura, H, Fujita, I & Tanaka, K (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophys.* **73**, 218–226.

[12] Kobatake, E & Tanaka, K (1995). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex *J. Neurophys.* **71**, 856-867.

[13] Logothetis, NK, Pauls, J & Poggio, T (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Opinion in Neurobiology* **5**, 552-563.

[14] Logothetis, NK, & Pauls, J (1995). Psychophysical and Physiological Evidence for Viewer-centered Object Representations in the Primate. *Cerebral Cortex* **3**, 270-288.

[15] Luck, SJ, Chelazzi, L, Hillyard, SA & Desimone, R (1997). Neural mechanisms of spatial selective attention in areas V1, V2, V4 of macaque visual cortex. *J. Neurophys.* **77**, 24-42.

[16] Nicholas, J, Reynolds, J & Desimone, R (1996). Texture segmentation with attention in area V4 of the macaque. *Soc. Neurosc. Abstr.* **22**, 1614.

[17] Olshausen, BA, Anderson, CH & Van Essen, DC (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700-4719.

[18] Oren, M, Papageorgiou, C, Sinha, P, Osuna, E & Poggio, T (1997). Pedestrian Detection Using Wavelet Templates. In *Proceedings of CVPR 1997*, 193-199.

[19] Perret, D & Oram, M (1993). Neurophysiology of shape processing. *Image Vision Comput.* **11**, 317-333.

[20] Poggio, T & Edelman, S (1990). A Network that learns to recognize 3D objects. *Nature* **343**, 263-266.

[21] Poggio, T & Beymer, D (1996). Learning to see. *IEEE Spectrum*, May, 60-69.

[22] Riesenhuber, M & Dayan, P (1997). Neural models for part-whole hierarchies. In *Advances In Neural Information Processing* **9**, 17-23. MIT Press.

[23] Riesenhuber, M, and Poggio, T: Just One View: Invariances in Inferotemporal Cell Tuning. Submitted to NIPS*97.

[24] Sung, K & Poggio, T (1995). *Example based learning for view-based human face detection.* AI Memo 1521, CBCL paper 112, Cambridge, MA: MIT.

[25] Ungerleider, LG & Haxby, JV (1994). 'What' and 'Where' in the Human Brain. *Current Opinion in Neurobiology* **4**, 157-165.

[26] Vapnik, V (1995). *The Nature of Statistical Learning Theory.* Springer.

[27] Vetter, T & Poggio, T (1995). Linear Object Classes and Image Synthesis from a Single Example Image. AI Memo 1531. Cambridge, MA: MIT.

[28] Wang, G, Tanaka, K & Tanifuji, M (1996). Optical Imaging of Functional Organization in the Monkey Inferotemporal Cortex. *Science* **272**, 1665-1668.

[29] Young, MP & Yamane, S (1992). Sparse Population Coding of Faces in the Inferotemporal Cortex. *Science* **256**, 1327-1331.