Qualitative Representations for Recognition

Pawan Sinha

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02142
sinha@ai.mit.edu

Abstract. The success of any object recognition system, whether biological or artificial, lies in using appropriate representation schemes. The schemes should efficiently encode object concepts while being tolerant to appearance variations induced by changes in viewing geometry and illumination. Here, we present a biologically plausible representation scheme wherein objects are encoded as sets of qualitative image measurements. Our emphasis on the use of qualitative measurements renders the representations stable in the presence of sensor noise and significant changes in object appearance. We develop our ideas in the context of the task of face-detection under varying illumination. Our approach uses qualitative photometric measurements to construct a face signature ('ratio-template') that is largely invariant to illumination changes.

1 Introduction

The appearance of a 3D object can change dramatically with variations in illumination conditions and viewing position. The challenge a recognition system faces is to classify all these different instances as arising from the same underlying object. The system's success depends on the nature of the internal object representations, against which the observed images are matched. Here we describe a candidate representation scheme that possesses several desirable characteristics, including tolerance to photometric variations, computational simplicity, low memory requirements and biological plausibility.

We develop the scheme in the context of a specific recognition task – detecting human faces under variable lighting conditions. Two key sources of difficulty in constructing face detection systems are (a) the variability of illumination conditions, and (b) differences in view-points. Experimental evidence [Bruce, 1994; Cabeza et al., 1998] suggests that the prototypes the visual system uses for detecting faces may be view-point specific. In other words, it is likely that distinct prototypes are used to encode facial appearance corresponding to different view-points. This hypothesis leaves open the question of how to collapse the illumination induced appearance variations for a given view-point into a compact prototype. The representation scheme we propose provides a candidate solution to this problem. Of course, many schemes for face detection have already been proposed in the computer vision literature [Govindaraju et al., 1989; Yang & Huang, 1993, 1994; Hunke, 1994; Sung & Poggio, 1994; Rowley et al, 1995; Viola & Jones, 2001]. What distinguishes our proposal from past work is that it is motivated primarily by psychophysical and physiological studies of the human visual system, as summarized in the next section. Our problem domain is shown in figure 1.

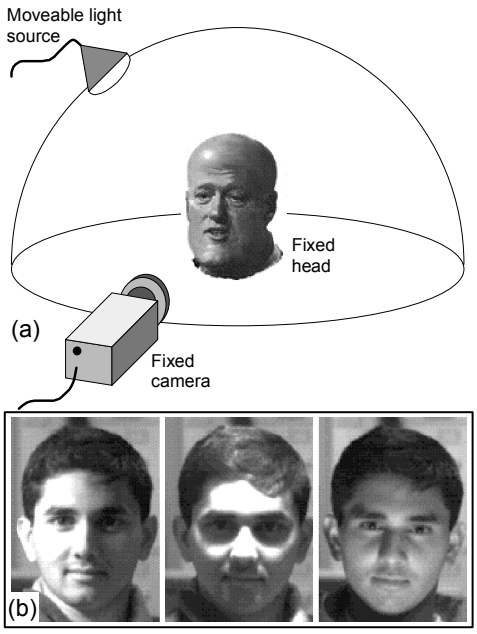


Fig. 1. Our problem domain. (a) Front facing upright heads being imaged under varying illumination conditions to yield images such as those shown in (b).

2 Face detection by human observers

In order to investigate the roles of different image attributes in determining face detection performance by human observers, we conducted a series of psychophysical experiments, details of which can be found in [Torralba and Sinha, 2001]. Here we focus on two experiments that are most relevant to the design of our representation scheme. The first seeks to determine the level of image detail needed for reliable face-detection and the second investigates the encoding of contrast relationships across a face.

2.1 Experiment 1: Face detection at low-resolution

What is the minimum resolution needed by human observers to reliably distinguish between face and non-face patterns? More generally, how does the accuracy of face classification by human observers change as a function of available image resolution? These are the questions our first experiment is designed to answer.

2.1.1 Methods

Subjects were presented with randomly interleaved face and non-face patterns and, in a 'yes-no' paradigm, were asked to classify them as such. The stimuli were grouped in blocks, each having the same set of patterns, but at different resolutions. The presentation order of the blocks proceeded from the lowest resolution to the highest. Ten subjects participated in the experiment. Presentations were self-timed. Our stimulus set comprised 200 monochrome patterns. Of these, 100 were faces of both genders under different lighting conditions (set 1), 75 were non-face patterns (set 2) derived from a well-known face-detection program (developed at the Carnegie Mellon University by Rowley et al

[1995]) and the remaining 25 were patterns selected from natural images that have similar power-spectra as the face patterns (set 3). The patterns included in set 2 were false alarms (FAs) of Rowley et al's computational system, corresponding to the most conservative acceptance criterion yielding 95% hit rate. Sample non-face images are shown in figure 2. Reduction in resolution was accomplished via convolution with Gaussians of different sizes (with standard deviations set to yield 2, 3, 4, and 6 cycles per face; these correspond to 1.3, 2, 2.5 and 3.9 cycles within the eye-to-eye distance.

Fig. 2. A few of the non-face patterns used in our experiments. The patterns comprise false alarms of a computational face-detection system and images with similar spectra as face images.

From the pooled responses of all subjects at each blur level, we computed the mean hit-rate for the true face stimuli and false alarm rates for each set of distractor patterns. These data indicated how subjects' face-classification performance changed as a function of image resolution.

2.1.2 Results of Experiment 1

Figure 3 shows data averaged across 10 subjects. Subjects achieved a high hit rate (96%) and a low false-alarm rate (6% with Rowley et al's FPs and 0% with the other distractors) with images having only 3.9 cycles between the eyes. Performance remained robust (90% hit-rate and 19% false-alarm rate with the Rowley et al's FA distractor set) at even higher degrees of blur (2 cycles/ete).

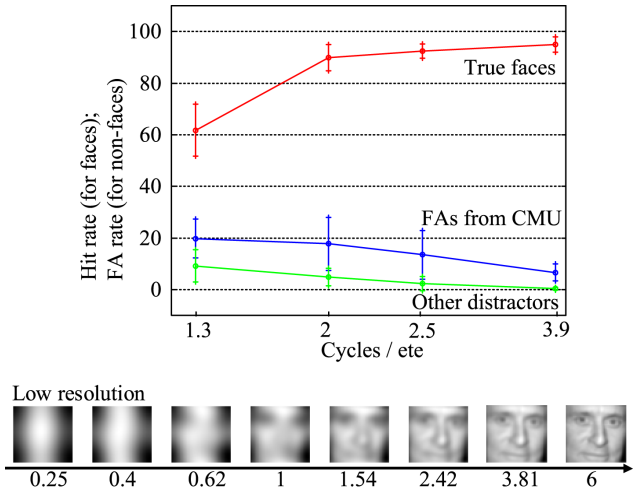


Fig. 3. Results from experiment 1.The units of resolution are the number of cycles eye to eye.

The data suggest that faces can be reliably distinguished from non-faces even at just 2 cycles eye-to-eye. Performance reaches an asymptote around 4 cycles/ete. Thus, even under highly degraded conditions, humans are correctly able to reject most non-face patterns that the artificial systems confuse for faces. To further underscore the differences in capabilities of current computational face detection systems and the HVS, it is instructive to consider the minimum image resolution needed by a few of the proposed machine-based systems: 19x19 pixels for Sung and Poggio [1994]; 20x20 for Rowley et al [1995]; 24x24 for Viola and Jones [2001] and 58x58 for Heisle et al. [2001]). Thus, computational systems not only require a much larger amount of facial detail for detecting faces in real scenes, but also yield false alarms that are correctly rejected by human observers even at resolutions much lower than what they were originally detected at.

2.2 Experiment 2: Role of contrast polarity in face detection

In studies of face identification, it has been found that contrast negation compromises performance [Galper, 1970; Bruce & Langton, 1994]. However, it is unknown how this affects the face-detection task. A priori, it is not clear whether this transformation should have any detrimental effects at all. For instance, it may well be the case that though it is difficult to identify people in photographic negatives, the ability to say whether a face is present may be unaffected since contrast negation preserves the basic geometry of a face. Experiment 2 is designed to test this issue. The basic experimental design follows from experiment 1. However the stimulus set of experiment 2 was augmented to include additional stimuli showing the faces and non-faces contrast negated.

2.2.1 Results of Experiment 2

Figure 4 shows that contrast negation causes significant decrements in face-detection performance. These results suggest that contrast reversal of face-patterns destroys the diagnostic information that allows their detection at low-resolution. Figure 5 also highlights the important role of contrast polarity in recognizing a pattern to be a face.

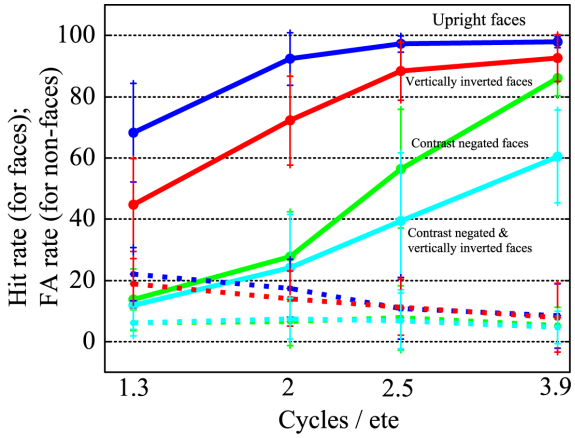


Fig. 4. Face detection performance following contrast negation.

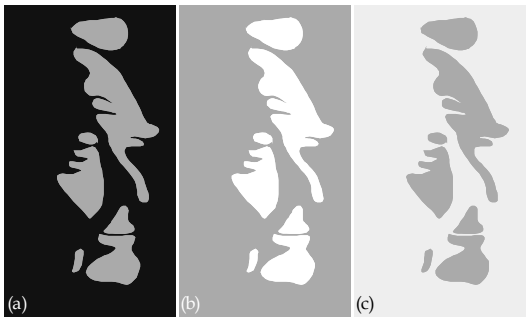


Fig. 5. Preserving absolute brightnesses of image regions is neither necessary nor sufficient for recognition. The patches in (a) and (b) have very different brightnesses and yet they are perceived as depicting the same object. The patches in (a) and (c), however, are perceived very differently even though they have identical absolute brightnesses. The direction of brightness contrast appears to have greater perceptual significance. (Mooney image courtesy: Patrick Cavanagh, Harvard University)

The significance of low-resolution image structure and contrast polarity rather than contrast magnitude per se is also reflected in the response properties of neurons in the early stages of the mammalian visual pathway. Beginning with the pioneering studies of Hubel and Wiesel, it has been established that many of these cells respond best to contrast edges. Many of these cells have large receptive fields and are, therefore, best suited to encoding coarse image structure. Additionally, studies exploring changes in response magnitude as a function of contrast strength have revealed that most neurons exhibit a rapidly saturating contrast response curve [DeAngelis et al, 1993]. In other words, the cell reaches its maximal response at very low levels of contrast so long as the contrast polarity is appropriate. For higher values of contrast, the cell-response does not change and is, therefore, uninformative regarding contrast magnitude. Such a cell thus serves as an ordinal comparator indicating whether the contrast polarity across the regions in its receptive field is correct and providing little quantitative information about contrast magnitude. The idealization of such a cell serves as the basic building block of our 'qualitative' image representation scheme.

3 'Ratio Templates': A qualitative scheme for encoding faces

The psychophysical results summarized above lead to two clear conclusions. First, human face detection performance is robust even at very low image resolutions and, second, it is sensitive to contrast polarity. The challenge is to devise a representation scheme that can take into account these results.

We propose a representation that is a collection of several pair-wise ordinal contrast relationships across facial regions. Consider figure 6. It shows several pairs of average brightness values over localized patches for each of the three images included in figure 1(b). Certain regularities are apparent. For instance, the average brightness of the left eye is always less than that of the forehead, irrespective of the lighting conditions. The relative magnitudes of the two brightness values may change, but the sign of the inequality does not. In other words, the *ordinal* relationship between the average brightnesses of the <left-eye, forehead> pair is invariant under lighting changes. Figure 6

also shows several other such pair-wise invariances. By putting all of these pair-wise invariances together, we obtain a larger composite invariant (figure 7). We call this invariant a *'ratio template'*, given that it is comprised of a set of binarized ratios of image brightnesses. It is worth noting that dispensing with precise measurements of image brightnesses not only leads to immunity to illumination variations, but also renders the ratio-template robust in the face of sensor noise. It also reconciles the design of the invariant with known perceptual limitations - the human visual system is far better at making relative brightness judgments than absolute ones. The 'ratio-template' is not a strict invariant, in that there exist special cases where it breaks. One such situation arises when the face is strongly illuminated from below. However, for almost all 'normal' lighting conditions (light sources at or above the level of the head), the ratio-template serves as a robust invariant.

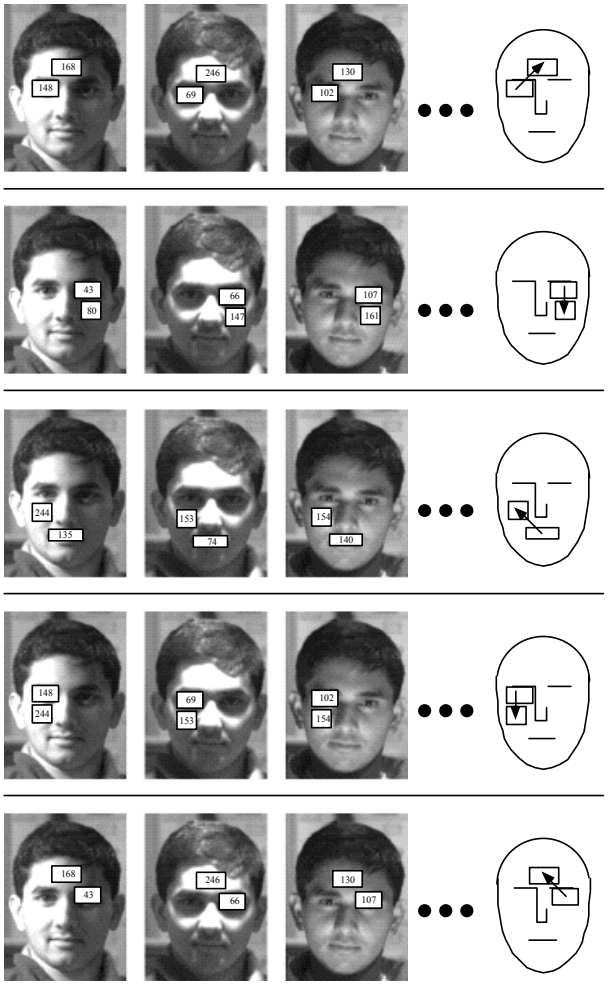


Fig. 6. The absolute brightnesses and even their relative magnitudes change under different lighting conditions but several pair-wise ordinal relationships are invariant.

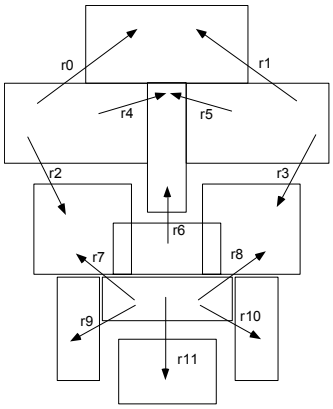


Figure 7. By putting together several pair-wise invariants, we obtain what we call a 'ratio-template'. This is a representation of the invariant ordinal structure of the image brightness on a human face under widely varying illumination conditions.

3.1 The match metric

Having decided on the structure of the ratio-template (which, in essence, is our model for a face under different illumination setups), we now consider the problem of matching it against a given image fragment to determine whether or not that part of the image contains a face. The first step involves averaging the image intensities over the regions laid down in the ratio-template's design and then determining the prescribed pair-wise ratios. The next step is to determine whether the ratios measured in the image match the corresponding ones in the ratio-template. An intuitive way to think of this problem is to view it as an instance of the general graph matching problem. The patches over which the image intensities are averaged constitute the nodes of the graph and the inter-patch ratios constitute the edges. A directed edge exists between two nodes if the ratio-template has been designed to include the brightness ratio between the corresponding image patches. The direction of the edge is such that it points from the node corresponding to the brighter region to the node corresponding to the darker one. Each corresponding pair of edges in the two graphs is examined to determine whether the two edges have the same direction. If they do, a predetermined positive contribution is made to the overall match metric, and a negative one otherwise. The magnitude of the contribution is proportional to the 'significance' of the ratio. A ratio's significance, in turn, is dependent on its robustness. For instance, the eye-forehead ratio may be considered more significant than the nose-tip-cheek ratio since the latter is more susceptible to being affected by such factors as facial hair and is therefore less robust. The contributions to be associated with different ratios can be learned automatically from training examples although in the current implementation, they have been set manually. After all corresponding pairs of edges have been examined, the magnitude of the overall match metric can be used under a simple threshold based scheme to declare whether or not the given image fragment contains a face. Alternatively, the vector indicating which graph edges match could be the input to a statistical classifier.

3.1.1 First order analysis

It may seem that by discarding the brightness ratio magnitude information, we run the risk of rendering the ratio-template too permissive in terms of the patterns that it will accept as

faces; several false positives would be expected to result. In this section, we present a simple analysis showing that the probability of false positives is actually quite small. We proceed by computing how likely it is for an *arbitrary* distribution of image brightnesses to match a ratio-template. In the treatment below, we shall use the graph representation of the spatial distribution of brightnesses in the image and the template.

Let us suppose that the ratio-template is represented as a graph with n nodes, and e directed edges. Further suppose that if all the edges in this graph were to be replaced by undirected edges, it would have c simple cycles. We need to compute the cardinality of the set of all valid graphs defined on n nodes with e edges connecting the same pairs of nodes as in the template graph. A graph is 'valid' if it represents a physically possible spatial distribution of intensities. A directed graph with a cycle, for instance, is invalid since it violates the principle of transitivity of intensities. Each of the e edges connecting two nodes (say, A and B) can take on one of three directions:

1. if A has higher intensity value than B, the edge is directed from A to B, or
2. if B has higher intensity value than A, the edge is directed from B to A, or
3. if A and B have the same intensity values, the edge is undirected.

The total number of graphs on n nodes and e edges, therefore, is 3^e. This number, however, includes several invalid graphs. A set of m edges that constitute a simple cycle when undirected, introduce $2(2^m - 1)$ invalid graphs, as illustrated in figure 8. For c such sets, the total number of invalid graphs are

$$\sum_{i=1 \text{ to } c} 2(2^{m_i} - 1)$$

where m_i is the number of edges in the 'cycle set' i.

Therefore, the total number of valid graphs on n nodes, e edges and c cycles is

$$3^e - \sum_{i=1 \text{ to } c} 2(2^{m_i} - 1)$$

Of all these graphs, only one is acceptable as representing a human face. For most practical ratio-template parameters, the total number of valid graphs is quite large and the likelihood of an arbitrary distribution of image brightnesses accidentally being the same as that for a face is very small. For instance, for e = 10 and two cycle sets of sizes 6 and 3, the number of valid graphs is nearly 59,000. If all the corresponding intensity distributions are equally likely, the probability of a false positive is only $1.69 * 10^{-5}$.

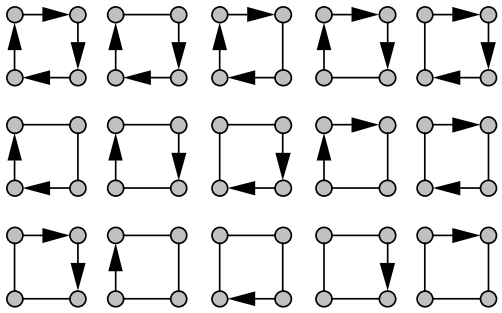


Fig. 8. A cycle set of m edges yields $2(2^m - 1)$ invalid graphs. A cycle set of 4 edges, for instance, yields 30 invalid graphs, 15 of which are shown above (the other 15 can be obtained by reversing the arrow directions). Each of these graphs leads to impossible relationships between intensity values (say, a and b) of the form a > b & b > a or a > b & a = b.

3.2 Implementation issues

The face-invariant we have described above requires the computation of the average intensities over image regions of different sizes. An implementation that attempted to compute these averages over the image patches at each search location would be computationally wasteful of the results from previous search locations. A far more efficient implementation can be obtained by adopting a multi-resolution framework. In such a framework, the input image is repeatedly filtered and subsampled to create different levels of the image pyramid. The value of a single pixel in any of these images corresponds to the average intensity of an image patch in the original image if the filter used during the construction of the pyramid is a diffusing one (like a Gaussian). The deeper the pyramid level is, the larger the patch. Thus, a pyramid construction equips us with a collection of values that correspond to precomputed averages over patches of different sizes in the original image. The process of determining the average value for any image patch is thus reduced to picking out the appropriate pixel value from the bank of precomputed pyramid levels, leading to a tremendous saving in computation. The appropriate scale of operation for a given ratio-template depends on the chosen spatial parameters such as the patch sizes and the distances between them. By varying these parameters systematically, the face detection operation can be performed at multiple scales. Such a parameter variation is easily accomplished in the pyramid based implementation described above. By tapping different sets of the levels constituting the image pyramid, the presence of faces of different sizes can be determined. To have a denser sampling of the scale space, the inter-patch distances can be systematically varied while working with one set of levels of the pyramid. The natural tolerance of the approach to minor changes in scale takes care of handling the scales between the sample points. It is worth noting that the higher the pyramid levels, the lesser is the computational effort required to scan the whole image for faces. Therefore, the total amount of computational overhead involved in handling multiple scales is not excessive.

3.3 Tests

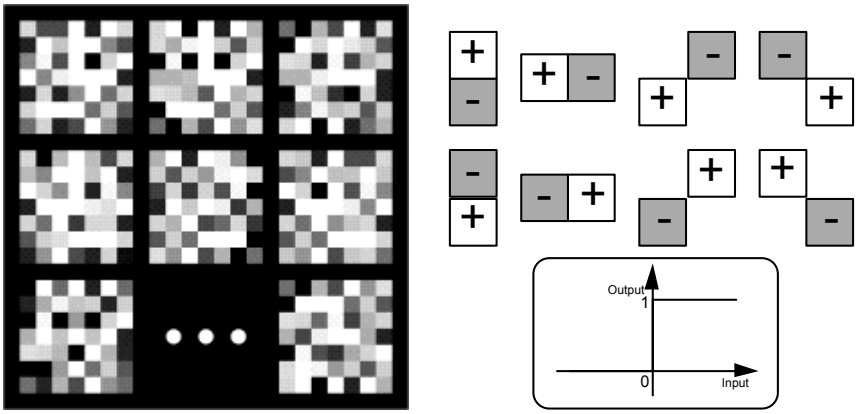
Figure 9 shows some of the results obtained on real images by using a ratio-template for face detection. Whenever it detects a face, the program pinpoints the location of the center of the head with a little white patch or a rectangle. The results are quite encouraging with a correct detection rate of about 80% and very few false positives. The 'errors' can likely be reduced even further by appropriately setting the threshold of acceptance. The results demonstrate the efficacy of the ratio-template as a face detector capable of handling changes in illumination, face identity, scale, facial expressions, skin tone and degradations in image resolution.

3.4 Learning the Signature

Our construction of the ratio-template in the preceding sections relied on a manual examination of several differently illuminated face images to determine whether there existed any regularities in their brightness distributions. A natural question to ask is whether we can design a learning system that can automatically extract a ratio-template from example streams. In principle, to accomplish this task, a learning system needs to determine which members of a potentially large set of image measurements are highly correlated with the presence of a face in the example images. We have tested this conceptually simple idea by extracting a ratio-template from a set of synthetic face images. Figure 10(a) shows some of the input images we used. These were generated using a program that embedded certain face-like invariances in variable random

backgrounds. The task of the learning system was to recover these invariances from labeled (face/non-face) examples.

Fig. 9. Testing the face-detection scheme on real images. The program places a small white square at the center of, or a rectangle around, each face it detects. The results demonstrate the scheme's robustness to varying identity, facial hair, skin tone, eye-glasses and scale.



(a) (b)

Fig. 10. (a) These images are representative of the inputs to the learning system. The images are synthetic and are meant to represent differently illuminated faces on varying random backgrounds. (b) The receptive fields of the pre-processors and their output function.

The 'receptive fields' of the pre-processor units are shown in figure 10(b). These units can be thought of as detecting inequality relations between adjacent image patches. The learning system needs to estimate the correlation of each measurement with the presence of a face. As is to be expected, only the measurements that are part of the invariant survive through all the examples while others weaken. This is shown in figure 11. It is

important to notice that by the end of the computation, we have not only constructed the object concept (the 'ratio-template' in this case) but have also implicitly learned to segment it from the background. This approach, therefore, simultaneously addresses two important issues in recognition: 1. What defines an object?, and 2. How can one segment a scene into different objects? In recent work (Thoresz and Sinha, 2002), we have successfully tested this learning approach on real images besides the synthetic ones shown here.

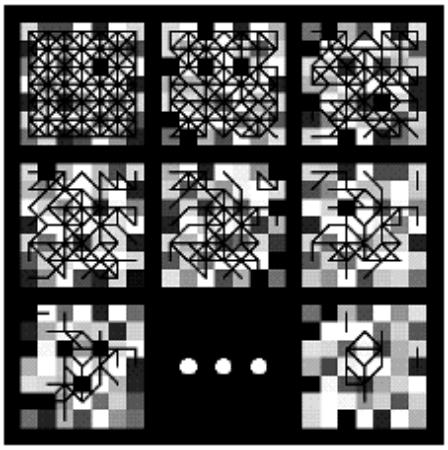


Fig. 11. Detecting relevant features via correlational learning over a set of examples. The output is the object concept. Segmentation is a side-effect.

4 Endnote

We have suggested the use of a qualitative face signature, that we call a ratio-template, as a candidate scheme for detecting faces under significant illumination variations. One can think of this specific scheme as an instance of a more general object recognition strategy that uses qualitative object signatures. Such a strategy would be attractive for the significant invariance to imaging conditions that it can potentially confer. However, it also has a potential drawback. Intuitively, it seems that the 'coarseness' of the measurements they use would render qualitative invariants quite useless at tasks requiring fine discriminations. How might one obtain *precise* model indexing using qualitative invariants that are, by definition, comprised of *imprecise* measurements? Depicting this problem schematically, figure 12(a) shows a collection of object models positioned in a space defined by three attribute axes. To precisely index into this model set, we can adopt one of two approaches: 1. we can either be absolutely right in measuring at least one attribute value (figure 12(b)), or 2. we can be 'approximately right' in measuring *all three* attributes (figure 12(e)). Being approximately right in just one or two attributes is not expected to yield unique indexing (figures 12(c) and (d)).

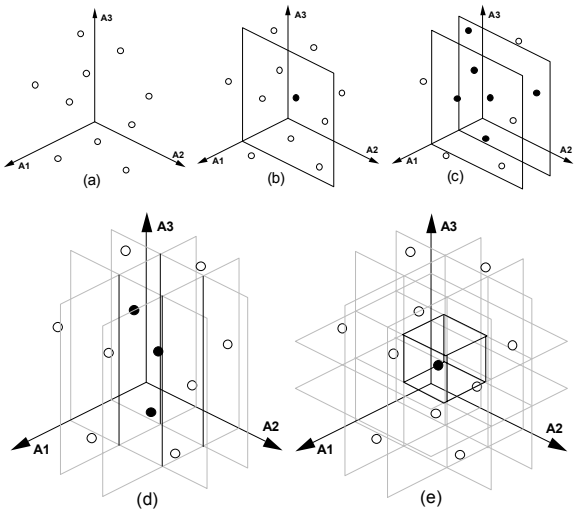


Figure 12. (a) A schematic depiction of a collection of object models positioned in a space defined by three attribute axes. To precisely index into this model set, we can adopt one of two approaches: 1. we can either be absolutely right in measuring at least one attribute value (figure (b)), or 2. we can be 'approximately right' in measuring all three attributes (figure (e)). Being approximately right in just one or two attributes is not expected to yield unique indexing (figures (c) and (d)).

The qualitative invariant approach constructs unique signatures for objects using several approximate measurements. The ratio-template is a case in point. It achieves its fine discriminability between face and non-face images by compositing several very imprecise binary comparisons of image brightnesses. In several real world situations, there might in fact be no alternative to using approximate measurements. This could be either because precise invariants just might not exist or because of noise in the measurement process itself. The only recourse in these situations would be to exploit several attribute dimensions and be 'approximately good' in measuring all of them. This is what qualitative invariants are designed to do. The 'recognition by qualitative invariants' approach is eminently suited to a complex visual world such as ours. Most objects vary along several different attribute dimensions such as shape, color, texture, and motion, to name a few. The qualitative invariant approach can exploit this complexity by constructing unique object signatures from approximate measurements along all of these dimensions. Evidence for the generality of this approach is provided by some of our recent work. We have implemented recognition schemes based on qualitative templates (that use not only qualitative photometric measurements but also spatial ones) for robustly recognizing a diversity of objects and scenes including natural landscapes, graphic symbols and cars.

A related observation is that the ratio-template representation is a 'holistic' encoding of object structure. Since each ordinal relation by itself is too coarse to provide a good discriminant function to distinguish between members and non-members of an object-class, we need to consider many of the relations together (implicitly processing object structure holistically) to obtain the desired performance. At least in the context of face-detection, this holistic strategy appears to be supported by our recent studies of concept

acquisition by children learning to see after treatment for congenital blindness [Sinha, 2002, in preparation].

Acknowledgements

The author wishes to thank Prof. Tomaso Poggio, Dr. Peter Burt, and Dr. Shmuel Peleg for several insightful discussions regarding this work.

References

Bruce, V. (1994) Stability from variation: the case of face recognition. *Quarterly Journal of Experimental Psychology*, 47A, 5-28.

Bruce, V. & Langton, S. (1994). The use of pigmentation and shading information in recognizing the sex and identities of faces. *Perception,* 23, 803-822.

Cabeza, R., Bruce, V., Kato, T. & Oda, M. (1998). The prototype effect in face recognition: Extension and limits. *Memory and Cognition.*

Galper, R. E. (1970). Recognition of faces in photographic negative. *Psychonomic Science*, 19, 207-208.

Gregory C. DeAngelis, Izumi Ohzawa, and Ralph D. Freeman (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J. Neurophysiol.* 69: 1091-1117.

Heisle, B., T. Serre, S. Mukherjee and T. Poggio. (2001) Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images. In: *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, IEEE Computer Society Press, Jauai, Hawaii, December 8-14, 2001.

Hunke, H. M. (1994) Locating and tracking of human faces with neural networks. Master's thesis, University of Karlsruhe.

Govindaraju, V., Sher, D. B., Srihari, R. K. and Srihari, S. N. (1989) Locating human faces in newspaper photographs. *Procs. of the Conf. on Comp.Vision and Pattern Recog.*, pp. 549-554.

Rowley, H. A., Baluja, S., Kanade, T. (1995) Human face detection in visual scenes, CMUtechnical-report,#CS-95-158R.

Sinha, P. (2002). Face detection following extended visual deprivation. (In preparation).

Sung, K. K., and Poggio, T. (1994) Example based learning for view-based human face detection, AI Laboratory memo # 1521, MIT.

Thoresz, K. and Sinha, P. (2002) Common representations for objects and scenes. *Proceedings of the Annual Meeting of the Vision Sciences Society,* Florida.

Torralba, A. and Sinha, P. (2001). Detecting faces in impoverished images. MIT AI Laboratory Memo, Cambridge, MA.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In: *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, IEEE Computer Society Press, Jauai, Hawaii, December 8-14, 2001.

Yang, G. and Huang, T. S. (1993) Human face detection in a scene. *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 453-458.

Yang, G. and Huang, T. S. (1994) Human face-detection in a complex background. *Pattern Recognition*, 27(1) pp. 53-63.