Spatial Vision, Vol. 8, No. 1, pp. 57–76 (1994)

Computational approaches to human pattern recognition

TERRY CAELLI[1,*] and WALTER F. BISCHOF[2]

[1] *Department of Computer Science, Curtin University of Technology, GPO Box U1987, Perth, WA 6001, Australia*

[2] *Department of Psychology, University of Alberta, Edmonton, Alberta T6G 2E9, Canada*

Received for publication 11 October 1993

Abstract—This paper consolidates recent findings on how humans detect and recognize patterns and considers computational procedures which reflect observed performance. A multi-level correlation model for spatial information processing is proposed and used to interpret past results on human psychophysical performance.

1. INTRODUCTION

The aim of this paper is to consolidate research completed over the past decade into how humans detect and recognize forms or shapes, and to present an overview of the underlying processes which capture various aspects of observed behaviour. Different types of recognition problems are considered and it is shown that a common set of computational procedures seems to underpin the known performance of the human visual system in interpreting images. However, before dealing with these issues in some detail, some consideration of what is understood by 'form' and 'shape' is necessary.

Pattern, form or structure are quite difficult to define in a succinct way. But one necessary condition for the presence of a structure in a signal is the existence of, at least, some degree of correlation within the signal. A signal which is perfectly uncorrelated in all dimensions (that is, white noise) has, by definition, no structure. Here, 'uncorrelated' means that we cannot predict the intensity of any pixel from any set of other pixels and so structures in textures, scenes or patterns are determined by the types of correlations—or pixel dependencies—present. Indeed, the study of these correlations—and the techniques to represent them formally—has been, for a long time, the main focus of both texture and form analysis (see Julesz, 1962).

An important example of how the existence of spatial correlations define patterns is the notion of an 'edge': the locus of the points corresponding to the boundary of patterns. They are the pixels whose intensities cannot be predicted from their neighbours. This can be clearly seen in the error image resultant from linear predictive coding (LPC, see Rosenfeld and Kak, 1982) and can also be applied to spatio-chromatic domains (Caelli and Reye, 1993). One of the early papers which proposed this connection between correlation and human form perception was that of Dodwell (1970) who argued that eye movements provide the basis for the autocorrelation of image data. This theme of autocorrelation was further developed by Uttal (1985), and the aim of this paper is to show how

* To whom reprint requests should be addressed.

correlation has played a central role in the understanding of biological pattern recognition.

Correlation has played a critical role in even defining the processing of spatial information in the vertebrate visual system, in particular for the notion of a 'receptive field profile' (RFP) that is used so often to model biological information processing. Neurophysiology and psychophysics assume the 'principle of maximum signal response' to measure what is being processed. That is, the response of a cell to different parametric states of a signal (for example, orientation) defines the neurone's RFP and the maximum response determines the underlying feature detector (see, for example, Hubel and Wiesel, 1968). All this assumes that analogues to such detectors exist within the visual system and that the response is reasonably modelled by the system correlating the internal detector with the external signal. Without this assumption it would not be possible to conclude anything like what has been concluded from the electrophysiology of vision over the past 50 years.

The reverse is assumed to occur in masking studies. That is, the very notion of 'fatiguing' spatially tuned neurones in the human visual cortex by repeated exposure to a given signal also assumes that the maximum masking effect will be induced by the signal which 'best matches' the given detector's profile. Again, correlation defines this match. In fact, in a direct study of this, it was possible to show that we could not reject the hypothesis that the masking effect was determined by the cross-correlation between signal and mask (Caelli and Moraglia, 1987a). The experiments used Gabor signals (Gaussian modulated sinusoidal gratings) which could be decorrelated by changes in frequency, orientation or phase between the masking and test patterns. A forward-masking task was used where a test signal was presented after a masking signal and subjects were required to indicate whether the test was present or not. The percentage correct detection was clearly predictable from the inverse of the peak of the cross-correlation between signal and mask. Notice that this cross-correlation was not 'in-place': the masking could occur within a neighbourhood of the test centre. These results argued for an adaptive channel model where the channel centre and bandwidth are determined by the correlation between signal and mask.

Such results have also been duplicated with more natural scences (Caelli and Moraglia, 1987b) where the masking patterns had either the same or different amplitude or phase spectra as the test patterns. Here it was shown that similarity in power spectra was unrelated to the degree of masking (the power spectra of two images were identical or quite different) and that the correlation between the actual images (in the space domain) was the deciding factor.

This discussion demonstrates that there is some basis for the involvement of correlation in early visual encoding and broad support for channels—neuronal subsystems selectively sensitive to specific signal correlations—in biological vision. What is required, however, is an analysis of the extent to which such a computational procedure applies to higher-order or more natural visual tasks.

Image correlations—and lack thereof, as measured by edge coding—occur at many different scales and colour bands, as is illustrated in Fig. 1 using different isotropic filters (from Caelli and Reye, 1993). Here we have used standard zero-crossings of the $\nabla^2 G$ (Gaussian low-pass-filter followed by the Laplacian differential operator) with two different Gaussians to obtain the results over

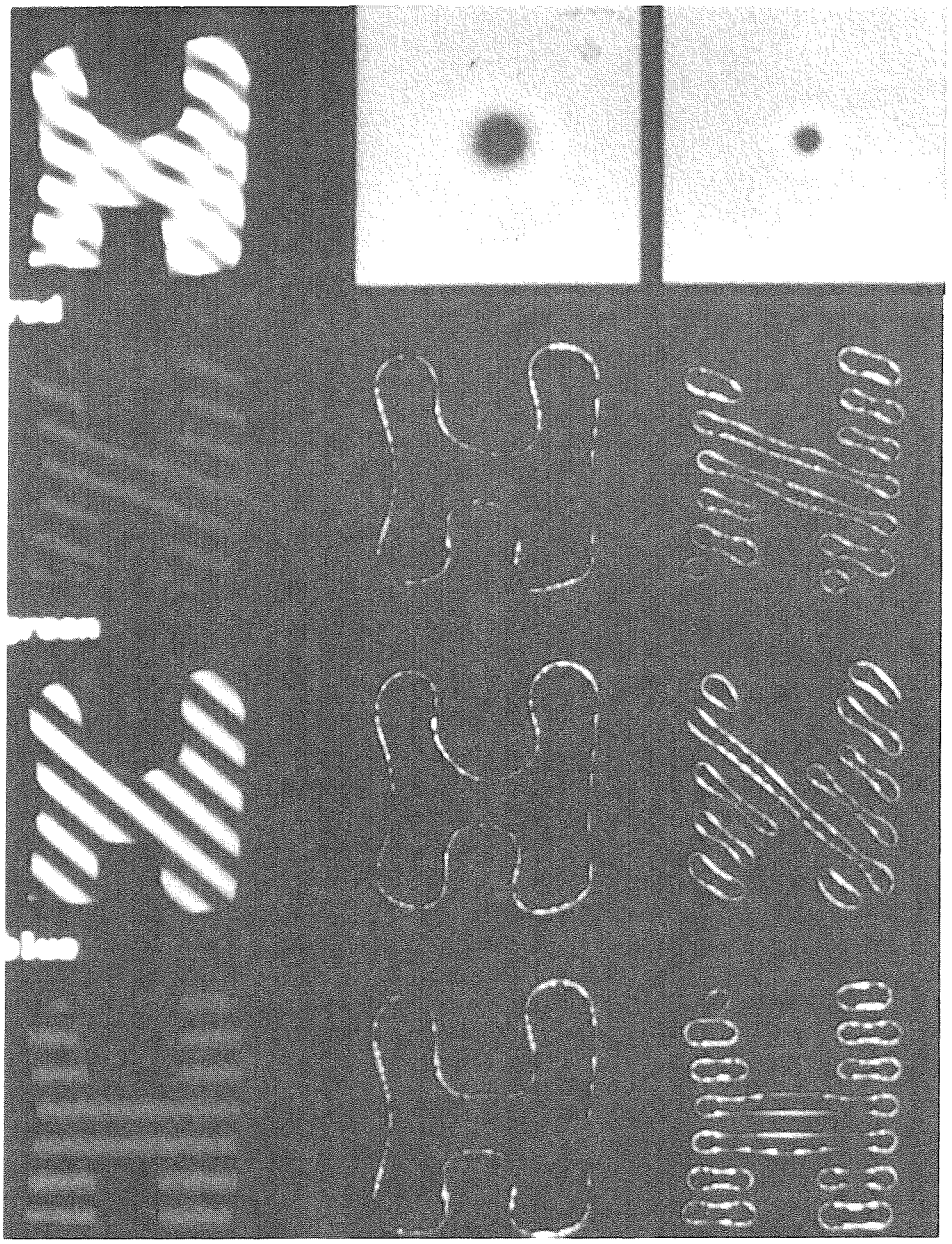


Figure 1. Different structures are captured by edge extraction (the lack of correlations 'marked' by zero-crossings of the $\nabla^2 G$ operator and in this case weighted by the pixel variance about each zero-crossing) at different scales. Here the scales were determined by different bandpass filters derived from different low-pass Gaussians ($G(\sigma)$ for $\sigma_1 = 16$ and $\sigma_2 = 8$ pixels, for the 128×128 pixel image) applied before the Laplacian (∇^2). The effective filter point spread functions are shown in the top row and the following rows correspond to red, green, and blue colour line responses to the input (top left) coloured and textured pattern. (From Caelli and Reye, 1993.)

different scales and colours. Though somewhat artificial, this example illustrates that the difference between the components of an image which define 'texture' and 'shape' is, to a large extent, one of scale. It also shows how different sized receptive fields are necessary to encode the broad range of image structures in an image which are needed to define a 'pattern' or 'shape' and that spatio-chromatic correlations must be also included in modelling spatial vision. For this reason, and others, the existence of multi-scale feature extraction is an important component of perceptual information processing.

Multi-scale analysis is particularly appropriate for pattern recognition problems which require the detection or recognition of pattern components and their relations. Although rather obvious, it is nonetheless important to emphasize that 'pattern' involves the definition of parts and relationships at different scales. This type of representation is standard in many computer vision tasks, as, for example, in OCR (optical character recognition) where individual characters are typically encoded by the relationships between more fundamental stroke patterns (see Suen, 1990). It has also been the operational model for more complex human 2D-pattern and 3D-object recognition (Biederman, 1985)—though, in both cases it must be clearly defined what constitutes parts and relations. Here, a 'part' typically refers to a set of contiguous locations—a region—where each location (pixel) shares common attributes. 'Relations' refers to those attributes (features) which depict measurable (binary) comparative properties such a distances and angles between parts.

Form, pattern and shape can be defined along a representational continuum, from the level of images to the level of symbolic or categorical descriptions. At the former end of the continuum, patterns correspond to signals defined exactly and, at the other end, they correspond to generic descriptions that assume the existence of segmented image parts and characterize such parts and their relationships. Patterns of the latter type are usually found as examples of object categories (for example, 'chair') and have been a major focus of attention in recent work in machine pattern recognition and classification. Algorithms for form, pattern or shape recognition must also contain processes which enable recognition which is invariant to imaging and geometric transformations to be representative of the types of problems that occur in both natural and man-made environments. The main types of invariances are: one, those related to image and object properties such as camera position, perspective, lighting conditions, and object material; and, two, those related to geometric transformations—translations and rotations in two and three dimensions and, in some cases, dilations and non-rigid motions.

Together then, processes for pattern recognition fall into two groups, a 'direct' representation where the signal is specified exactly, as an image or template; and an 'indirect' representation where the signal is specified by parts which can vary in their feature states. The former processes traditionally fall into the 'pre-attentive' or 'early-vision' routines. The latter are typically associated with more goal-directed or 'intentional' aspects of perception as, for example, in the formation of explicit descriptions by the analysis of specific image region types and their relationships. However, all such levels of representation involve parts and relations (determined by correlation-type processes) at one scale or another and the 'attentional distinction', again, is somewhat task dependent as it applies to spatial scale.

In all, then, our view of biological pattern recognition is summarized in Fig. 2 where the visual system is seen to contain three different types of correlational processes. The lowest level (pixel correlation level) operates on the raw retinal image to result in classical receptive field profiles, spatial feature detectors or channels leading to the extraction of significant image features. The second level (feature correlation) refers to the ability of the visual system to analyse, over space, the outputs of such feature analysers resulting in the extraction of image region types or 'parts' such as critical pattern features, texture-based segmented regions, etc. The final level of correlation refers to the ability of the visual system to extract similarities and differences between image or pattern parts and their relations and so form the basis of more symbolic levels of pattern description and recognition.

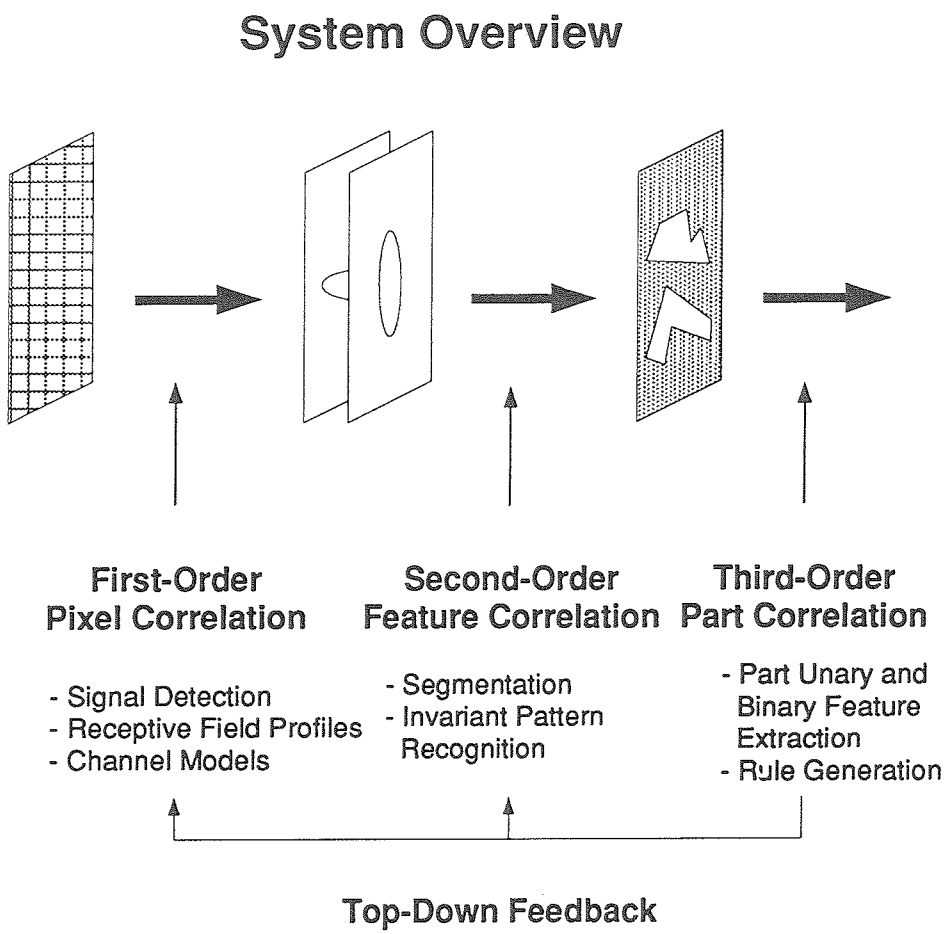


Figure 2. A multi-level perspective of pattern recognition. The processing of spatial information falls into three subsystems. The first corresponds to low-level pixel correlation processes for feature extraction, the second to low-level invariant pattern recognition and segmentation, and the third corresponds to goal-directed analysis of image characteristics for recognition and symbolic interpretation purposes.

2. FIRST- AND SECOND-ORDER CORRELATIONS

As discussed earlier, at the lowest level of processing, the essential model behind the very definition of receptive field profiles has always been cross-correlation or template matching. In this formulation of 'structure detection' the problem is to design a filter to optimally find a signal $S(x, y)$ in (zero-mean) white noise $N(x, y)$ where the image (G) model is:

$$G(x, y) = S(x, y) + N(x, y). \tag{1}$$

The solution is given by the 'matched filter theorem' (Rosenfeld and Kak, 1982), where the filter is the signal itself. The match is determined by the cross-correlation between the signal (S) and the image (G):

$$C(x, y) = S * G = \sum_{u, v} S(u, v) G(x + u, y + v), \tag{2}$$

where, by the Cauchy–Schwarz inequality,

$$\sum_{u, v} S(u, v) G(x + u, y + v) \leq \Big(\sum_{u, v} S^2(u, v) \Big)^{1/2} \Big(\sum_{u, v} G^2(x + u, y + v) \Big)^{1/2}, \tag{3}$$

with equality if and only if $G(x + u, y + v) = \lambda S(u, v)$ where λ corresponds to an amplification constant. The 'energy detector' (Van Trees, 1968) is the matched filter of the form:

$$\frac{C(x, y)}{\sqrt{E_G(x, y)}} = \sqrt{E_s}, \tag{4}$$

where the signal and local image energy are, respectively:

$$E_s = \sum_{u, v} S^2(u, v) \tag{5}$$

and

$$E_G(x, y) = \sum_{u, v} G^2(x + u, y + v). \tag{6}$$

A number of experiments have been done to evaluate whether human observers conform with this model in the detection of known signals in white noise (Burgess and Ghandeharian, 1984) and non-white noise (coloured noise) and even natural images (Caelli and Moraglia, 1986; Caelli and Nawrot, 1987). These experiments involve a forced-choice task where the known signal is embedded in one of a number of specified regions of an image and the observer, on any given trial, is asked to indicate in each trial where it is. A trial consisted of presenting the signal for a brief exposure (200 ms) to the observer, and after 500 ms (to avoid the perception of apparent motion) the full image was presented containing the embedded signal. The observer was required to choose the position of the signal from a given set of possible positions.

Performance in such experiments, as measured by percentage correct detection rates show a number of characteristics of human recognition. Humans are quite inefficient at this task (Burgess and Ghandeharian, 1984). That is, when one compares Eqn (4) with detection rates it is quite clear that humans are much less efficient than the ideal detector would be in the same task (ibid.). This could be due

to inefficiences in later decision making processes, but it might also be that the detection model is inappropriate, or that the signal was encoded in a different way.

Indeed, Caelli and Moraglia (1986) and Caelli *et al.* (1988) have shown that, if multi-scaled edge versions of signals are used, performance is predicted much more accurately for non-stationary images such as natural scenes, faces, etc. For example, Caelli *et al.* (1988) have investigated how different correlation models can predict the recognition of faces using a supervised learning paradigm. In this case, examples of three different classes were generated by (digitally) mixing three fundamental faces with various degrees of each basic face to produce the types of images shown in the top row of Fig. 3.

Subjects were then trained to correctly classify these examples in a 'learning phase' where the examples were briefly exposed (200 ms) followed by a display of

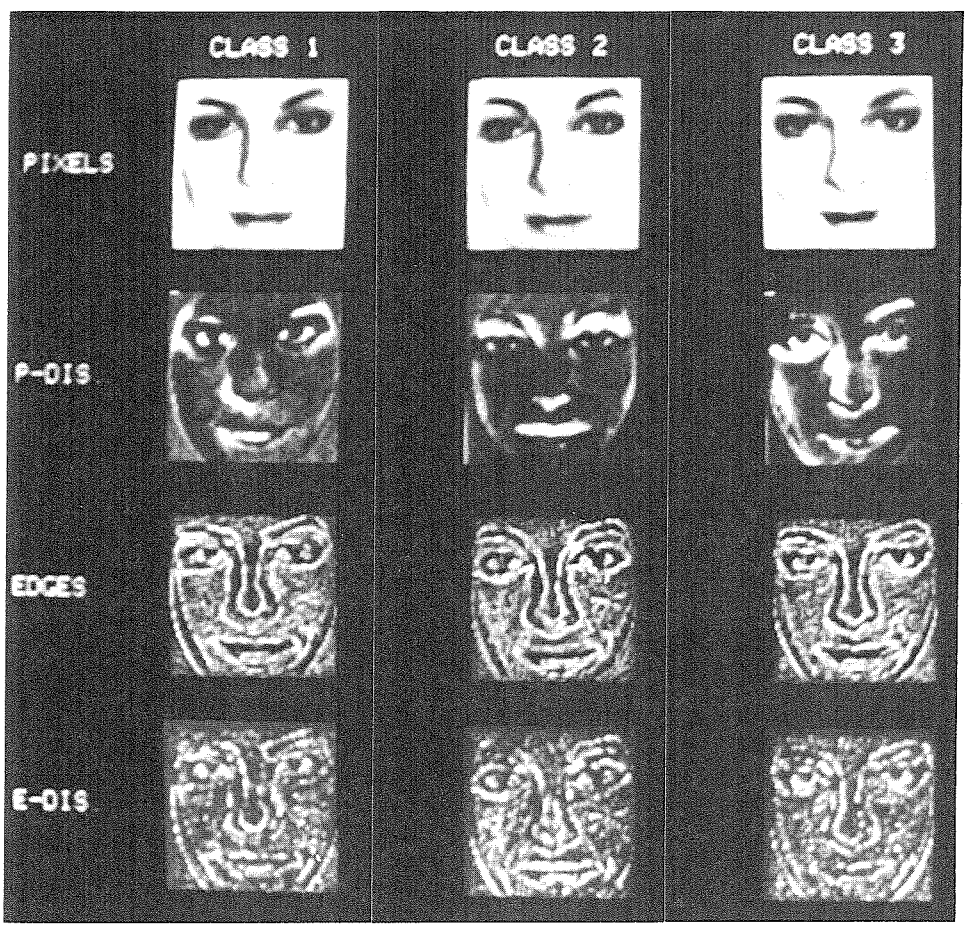


Figure 3. The face classification problem studied by Caelli *et al.* (1988). Here, subjects were trained to criterion performance on examples of faces and then asked to classify new composite faces. Top row shows composite (average) faces for three classes. Second, third and fourth rows correspond to intensity-disjunctive, multi-scale edge and edge-disjunctive versions of the class prototypes. Here, the disjunction operation was based on detecting features which were not common over classes—enhancing the discrimination of the class prototypes. Again, performance was best predicted by cross-correlations between edge-disjunctive prototypes of the multi-scaled edge versions of the faces.

a number indicating the class (1, 2, or 3). These training trials were interleaved with test trials where observers were required to identify the class membership without feedback. The learning phase was halted when observers made no errors in classification of all training examples. Following this, the 'test phase' consisted of displaying new examples (new mixed versions of the fundamental faces) which observers were required to classify into the three learned classes. The aim, then, was to investigate what type of correlation model is most parsimonious with human classification performance. Examples of these different types of prototypes are shown in Fig. 3.

Four different correlation models were studied and are defined below. In each case the class 'prototype' was defined by the aggregate of the (aligned) samples.

1. Direct correlation model. The predicted similarity between a sample and the aggregate class prototype is determined by the peak value of the cross-correlation between the signal and prototype.

2. Intensity-disjunctive correlation model. The predicted similarity between a sample and the aggregate class prototype is determined by the peak value of the cross-correlation between the signal and discriminating regions of the class prototypes.

3. Edge correlation model. The predicted similarity between a sample and the aggregate class prototype is determined by the peak value of the cross-correlation between a multi-scale edge version of the signal and the equivalent version of the class prototypes.

4. Edge-disjunctive correlation model. The predicted similarity between a sample and the aggregate class prototype is determined by the peak value of the cross-correlation between a multi-scale edge version of the signal and the discriminating components of the equivalent edge versions of the class prototypes.

Classification performance was consistent with an internal encoding of the prototypes and signals in terms of the multi-scale edge information and a representation of class prototypes which emphasized the discriminating features of each class.

These types of results support a number of other claims in the literature related to second-order feature correlations in spatial vision. For example, the spatial gradients of feature correlations (how the correlations between feature detector outputs vary over space) has been the predominant procedure for texture segmentation over the past decade (see Caelli, 1988; Gurnsey and Browse, 1989). In all, then, such experiments show that, in recognition processes, the visual system seems to emphasize regions (parts) which lack correlations (that is, edges) and which provide a basis for differentiating between different classes. In the above examples these 'parts' corresponded to regions which evidenced edges over a variety of scales (see Fig. 3). Similarly, we have argued for the encoding of patterns as 'signed blobs', at multiple scales, for texture (Caelli *et al.*,1986) and for images in general (Watt, 1987). Such processes, again, form the basis for recognition-by-parts (Biederman, 1985) as proposed for 3D object recognition.

In fact, earlier, Caelli and Dodwell (1984) and Dodwell and Caelli (1985) argued for such a form of tokens—encoding position, orientation and local invariance structures—and implemented a method for predicting the recognition of patterns by these measures. These experiments involved observers discriminating between pairs of 'vectorgraphs', distributions of oriented line

segments over the image plane, where one differed from the other in terms of jitters (random rotations) introduced in the orientations of the line elements or parts. Figure 4 shows examples of patterns used in these studies. The columns show different positional configurations while the rows show different orientation distributions and, together, different spatial gradients of oriented line segments occur. From such configurations it was possible to more clearly study the role of orientation encoding as a function of the spatial arrangement of the patterns. The results showed that discriminability was not predictable from simple orientation

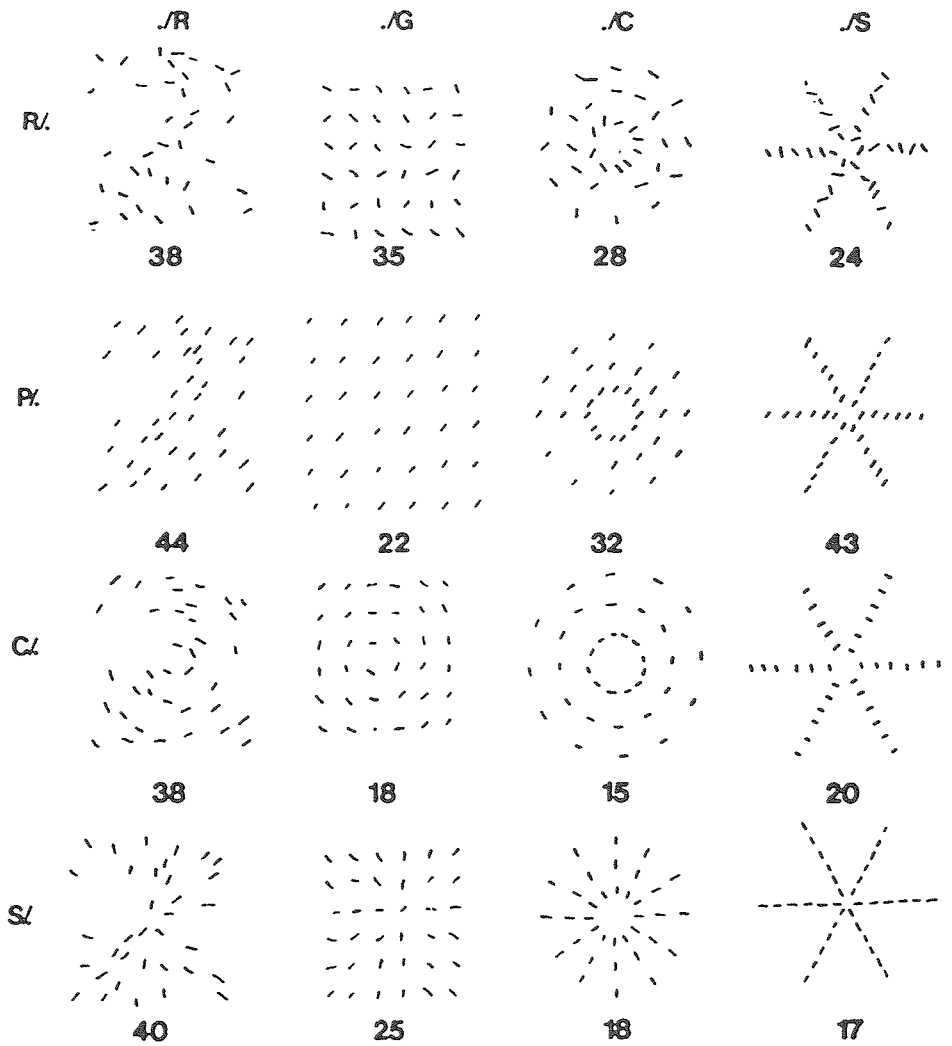


Figure 4. Examples of 'vectorgraphs' used by Caelli and Dodwell (1982) and Dodwell and Caelli (1985) to study how such configurations are processed in human vision. The columns show different positional configurations: random (R), grid (G), circle (C) and star (S), and the rows show different orientation distributions: random (R), parallel (P), circle (C) and star (S). Pattern discrimination was determined as a function of orientation jitters in each pattern (labels merely indicate pattern number). Discrimination was found to vary not only as a function of orientation differences alone, but also as a function of the relationship between orientation and position of the elements.

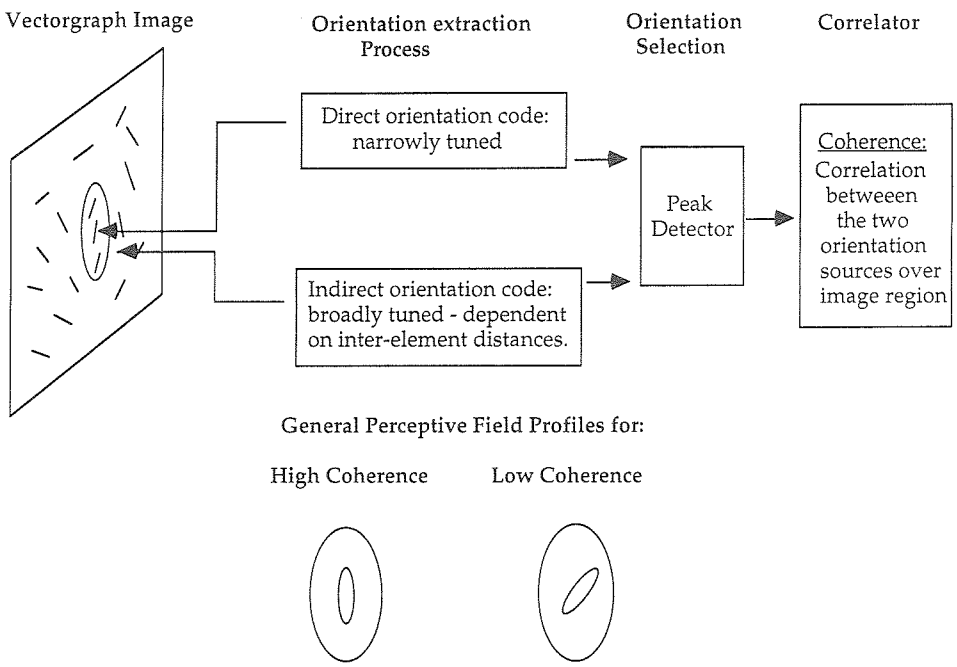


Figure 5. Coherence extraction algorithm used by Caelli and Dodwell (1984) to predict discrimination performance. Orientation discrimination was found to be enhanced when the spatial gradients of the positional information was locally consistent with the given orientations.

differences, but by the combination of positional and orientation information in the signals—a form of second-order feature correlation. It was possible to represent this type of information via two components: coherence and invariance. The former referred to the local consistencies between the rate of change of position with respect to rate of change of orientation (second-order feature correlations). The latter referred to the degree to which the pattern was actually invariant to rigid motions. Both components enhanced sensitivity to orientation discrimination. Figure 4 shows the types of patterns studied by us involving patterns with the same orientation histograms but having different spatial arrangements. Figure 5 shows the coherence computation which was most consistent with the discrimination performance where two types of orientation components were extracted and correlated.

That is, support was found for at least two levels of correlations: one which extracts local orientation information; and the other, a second-order feature correlator, which enables perceptual grouping via correlating different features extracted from the image. In this case, the concern was with 'coherence'. This measure defined the correlation between the actual orientations of line elements within an image region and orientations extracted from the rate of change of orientation with respect to position over that region. As would be expected by summation, when both sources of orientation were similar (highly correlated over the image: high coherence) individual line orientation sensitivity increases. This process is depicted in Fig. 5. Similar findings (without a computational model) have been recently reproduced by Ullman (1990) and Field *et al.* (1993).

This study provides an example of the strong evidence that human recognition processes do not actually operate on the image, *per se*, but rather on critical features, their relations, and the way these features index world structures. Further, it turns out that the very types of coding principles proposed to exist within the vertebrate visual cortex (namely, the significant involvement of spatial differentiation throughout the visual system as illustrated by excitatory and inhibitory regions of receptive fields right through the visual system) are a crude way of optimizing such operations as cross-correlations, segmentation and compression of image information. That is, such methods essentially reduce the redundancy in the signal by decreasing the perceptual salience of the correlated (non-edge) components. This type of process apparently occurs over different 'scales' of analysis, involving different sizes of receptive fields.

Results like those discussed above support a common theme emerging in the recent physiological and psychophysical literature. That is, although the visual system has available a relatively complete (retinal) representation of the image it attends primarily to less redundant features (less correlated regions over many scales) of an image and particularly those region which optimize performance on a given vision problem. In the face-classification experiment (see above) such critical features corresponded to edge-disjunctive information which encoded the most discriminating edge-specific facial features. These 'conjunctive' and 'disjunctive' images, again, delimit correlations common within a pattern class and not common between classes, respectively. From a cross-correlation perspective, this makes sense for non-stationary images since direct cross-correlations produce less than ideal performance, a result already noted for white noise by Burgess and Ghandeharian (1984) with human observers.

2.1. Geometric invariance

Such explorations do not accommodate two other important aspects of form which fall under the generic problem of 'stimulus equivalence' (Dodwell, 1970): geometric invariance and form invariance. The former refers to the problem of recognition under geometric transformations while the latter refers to recognition under shape distortions. In the following two sections these questions are briefly dealt with in this order.

A number of claims have been made as to just how the human visual system—and associated cognitive processes—copes with the problem of recognizing patterns independent of their geometric transformations. Some have argued for complete four-parameter invariance (translations, rotations and scale (dilations)) while others argue for essentially none. Psychophysical studies have repeatedly demonstrated a lack of complete rotation invariance and, in some tasks, the importance of absolute position of the stimulus in the visual field even for recognition processes. However, the actual visual task needs to be carefully considered before general conclusions can be made. It should be remembered that, normally, the human observer is quite capable of 3D object recognition with familiar objects. This is not only invariant to 2D but 3D (6-parameter) rigid motions. Consequently, task specificity and adaptation processes cannot be overemphasized in this area as many experimenters have clearly demonstrated how easy it is to train subjects to mentally rotate and consistently recognize

rotated patterns (Kolers and Perkins, 1975). On the other hand, it has been claimed that the existence of multiple orientation and scale-specific detectors at each retinal position permits, in principle, a 4-parameter invariant architecture. However, only recently have we (Zetzsche and Caelli, 1989) shown what extra normalizations are required to attain full invariance using this type of pattern encoding.

It should be noted that there are two types of invariances: weak ('blind') invariance and 'strong' (parametric) invariance. The former refers to the development of encoding schemes which simply do not code the transformations. A pixel histogram, for example, does not register image rotations and the (Cartesian) Fourier power spectrum does not encode translations of the full image (see Ferraro and Caelli, 1988, for more details). By parametric invariance is meant an encoding scheme (R) which is linear with respect to the transformations. For translations, rotations and scale transforms, the associated 4D representation is:

$$T(R(\bar{x}, \bar{y}, \bar{\sigma}, \bar{\theta})) = R(\bar{x} + \Delta\bar{x}, \bar{y} + \Delta\bar{y}, \bar{\sigma} + \Delta\bar{\sigma}, \bar{\theta} + \Delta\bar{\theta}), \tag{7}$$

where the transformation T for translations, rotations and dilations (by amounts Δ) is, in general, defined by:

$$x' = a^{\Delta\sigma}(x \cos \Delta\theta + y \sin \Delta\theta) + \Delta x \tag{8}$$

and

$$y' = a^{\Delta\sigma}(-x \sin \Delta\theta + y \cos \Delta\theta) + \Delta y. \tag{9}$$

Here, $\bar{x}$, $\bar{y}$, $\bar{\sigma}$, $\bar{\theta}$ correspond to normalized image coordinates which satisfy the 4D shift-invariant property. The computational procedure is illustrated in Figs 6 and 7 and consists of the following steps:

- *Create a set of filters indexed by orientation and scale.* In this implementation we used Gabor filters: filters whose point-spread functions corresponded to Gaussian-modulated grating of the spectral form:

$$g(u, v) = \exp\left(-\frac{(u - u_0)^2 + (v - v_0)^2}{f^{3/2}}\right), \tag{10}$$

where (u, v) define spatial-frequency coordinates, (u_0, v_0) the filter centre (spatial frequencies of the cosine gratings), and $f^{3/2}$ the frequency bandwidth, where $f = \sqrt{u_0^2 + v_0^2}$.

- *Implement correlations* between detector outputs in normalized coordinates.
- *Invert transforms* for determining match and transformation states:

$$x = a^{-\Delta\sigma}((x' - \Delta x) \cos \Delta\theta - (y' - \Delta y) \sin \Delta\theta) \tag{11}$$

and

$$y = a^{-\Delta\sigma}((x' - \Delta x) \sin \Delta\theta + (y' - \Delta y) \cos \Delta\theta). \tag{12}$$

The process is illustrated in Figs 6 and 7 below.

Up to the normalization factor, this system is perfectly consistent with the type of organization of the visual cortex proposed by Hubel and Wiesel (1968) as long as additional second-order correlations occur between orientation and size-tuned

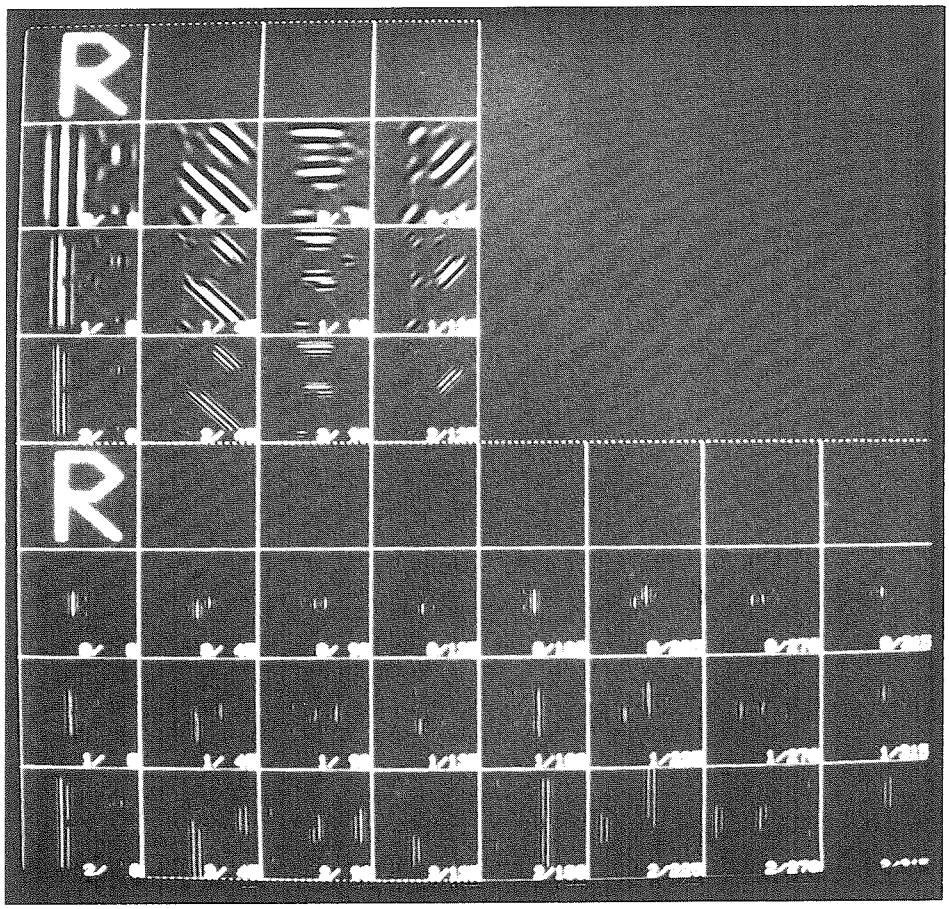
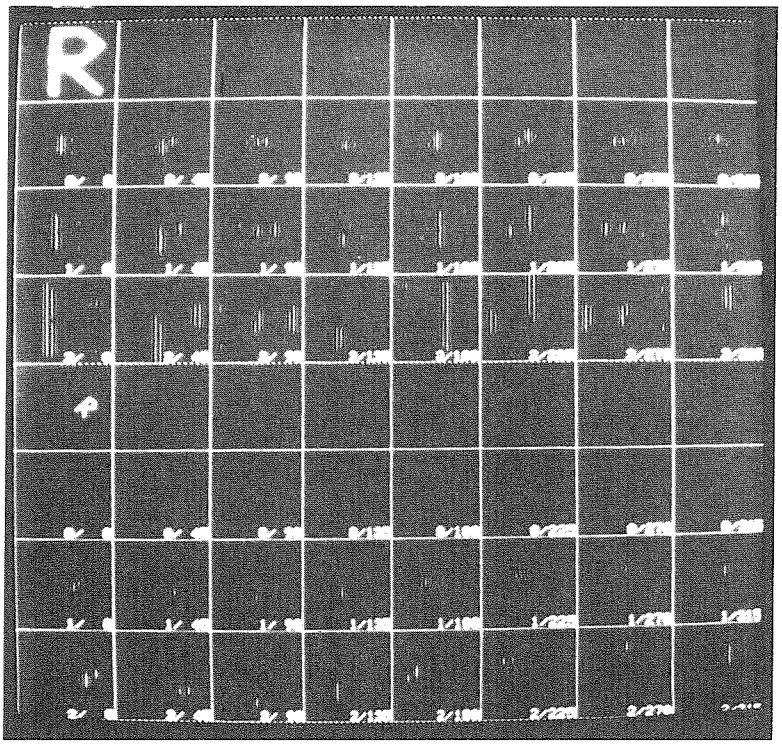


Figure 6. Upper left images: Input pattern 'R' and 12 versions filtered by Gaussian-modulated gratings of four different orientations (0, 45, 90 and 135 deg) and three different scales. Lower half: Input pattern and 24 versions as shown above but plotted in normalized filter coordinates. (From Zetzsche and Caelli, 1989.)

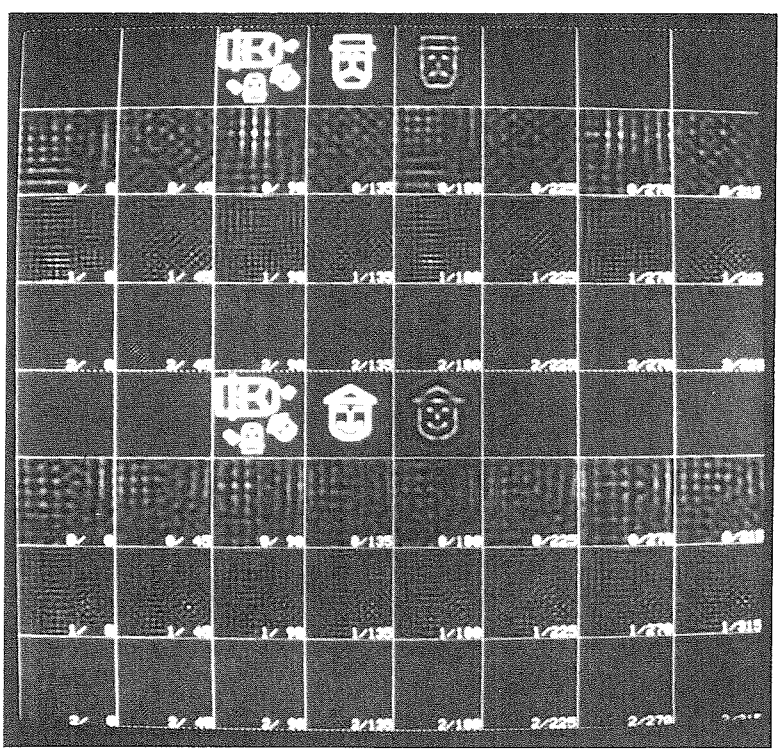
detectors. However, although useful for machine vision, the system is not consistent with many perceptual recognition tasks which show that pattern recognition is not invariant to rotations (see, for example, Foster, 1978). This does not imply that invariance encoding is not present. Rather, it simply shows that the pattern recognition required of many psychophysical experiments does not 'occur' at precisely this level of processing. That is, although this type of invariance may be present in the visual system, this does not imply that it is used in processing and learning spatial descriptions of patterns. Indeed, there is good evidence that many patterns, reflecting 3D objects, do have preferred orientation, positions and even size in the visual field. These expected states of familiar patterns seem to have been adapted to the extent that other positions are not so easily recognized.

Finally, some comments about Lie transformation groups, Lie algebras, Fourier transforms and invariance seem appropriate. It should first be noted that the visual system is particularly sensitive to specific patterns like concentric

(a)



(b)



circles, gratings and star-like patterns. This has been shown in recognition (Caelli and Dodwell, 1984; Dodwell and Caelli, 1985), adaptation and masking paradigms (Simas and Dodwell, 1990). However, this does not imply that the visual system has, explicitly, the differential operators corresponding to the equivalent Lie operators nor does it imply that these exhibit the internalized 'annulling' action of such operators to result in these invariant path curves being of specific importance as Hoffman (1968) originally claimed.

Indeed, as recently shown by Ferraro and Caelli (1988), these patterns also correspond to different kinds of Fourier transform basis functions and there is a direct relation between the corresponding Fourier power spectra invariants and invariants of corresponding Lie operators. In other words, it cannot uniquely be concluded from these observations that the result can only be 'explained' in terms of Lie operators or Fourier analysis 'in the head'. What can be concluded is that forms of spatial differentiation occur in human vision, and that there seems to be some sensitivity to patterns which correspond to invariant path curves of retinally (perspectively) projected rigid motions: affine projections of object motions onto the retinal surface.

2.2. Third-order correlations: Explicit structural descriptions

So far concentration has been on pattern representations where the signal and image are encoded in terms of another image. This type of representation, however, cannot readily accommodate the variations of form that the visual system can easily cope with as in the recognition of various types of 'cups', 'chairs' or facial expressions. This is because 'form' or 'patterns' also reference real-world objects or structures and so techniques which solely rely on definitions, qua images, are bound to be inadequate. Secondly, such types of representations are quite expensive on storage—a particularly important facet of 3D object recognition where objects have to be represented from many views. For these reasons, at least in the computational vision literature, it is common to extract a symbolic representation. That is, once the image has been segmented into regions based on position, orientation, size, colour and texture characteristics, systems typically extract the unary (part) and binary (relational) features (see Ballard and Brown, 1982). The 'internal' representation and matching processes are then enacted at the symbolic, or explicit feature-based, levels as outlined in theories of 'recognition-by-parts' (RBP).

This form of representation has been proposed in biological vision by Watt (1987) in terms of the extraction of 'blob' features at multiple scales and by Biederman (1985) for the representation of 3D objects by 'geons', parts represented by superquadrics. Unfortunately, no full recognition model has been produced in either case. In a similar way, Caelli *et al.* (1993) have recently shown

---

Figure 7. (a) Comparison of responses for two versions of 'R'. Notice how the 4D representation translates rotations, scale and translations all into translations in the 4D normalized coordinate system. (b) Results of cross-correlating the normalized forms of an image (top row, column 3; row 5, column 3) and a signal (top row, column 4; row 5, column 4) but only using one frequency band (top row, column 5; row 5, column 5). The output cross-correlation images are shown in rows 2–4 and 6–8 where intensity indicates the likelihood of the signal (of a given size and orientation relative to the original signal) being present at each position. (From Zetzsche and Caelli, 1989.)

how the perceptual matching of parts of sequentially displayed images are matched in apparent motion by the similarities of their part properties and their relations. In this case it was possible to model the correspondence process (the process of matching each pattern in Frame 1 with patterns in Frame 2, see Fig. 8) as a constraint satisfaction process defined by a parallel relaxation-type process of the form:

$$p_i^{t+1}(j) = \vartheta\Big((p_i^t(j) + \sum_{k,l} C(i,j;k,l)p_k^t(l)\Big), \tag{13}$$

where i, k correspond to part pairs in Frame 1 and j, l correspond to part pairs in Frame 2 as shown in Fig. 8; $p_i^t(j)$ corresponds to the probability that part i in Frame 1 is mapped to part j in Frame 2 at iteration t; C corresponds to the compatibility function between these two mappings; and ϑ corresponds to a non-linear function. This dynamical relaxation process converges in a few steps to the mapping between two patterns in terms of their part similarities and the compatibilities between the part relations (see Rosenfeld and Kak, 1982). The results demonstrated here that the perceptual apparent motion between patterns is defined by the perceived similarity between the individual shapes and the associated compatibilities between the mappings. In such cases the compatibilities are defined in terms of the similarities between part relations in each configuration as terms of the specific binary feature states.

Certainly, this approach has now become standard in many machine vision problems from optical character recognition (see Suen, 1990) to 3D object recognition systems (see Jain and Hoffman, 1988). In all such cases there are typically five common processes in the learning/run time phases:

- *Encoding:* the processes which define how image intensity or depth (range) data are encoded relative to specific task demands. This includes, for example, the determination of pixel feature vectors which encode colour, textural information in intensity images or the encoding of local surface curvatures in range images.
- *Segmentation:* the process of breaking the image up into parts based on differences in colour, texture or, for depth information, shape or surface curvature differences.
- *Feature extraction:* the processes of generating a description of the resultant parts (unary predicates) and their relations (binary predicates). Unary features include part intensities, areas, positions and orientations. Binary features include interpart angles, distances and contrasts.
- *Rule generation:* the extracted features have to be summarized and generalized to form rules which maximally evidence difference patterns or classes.
- *Matching:* new patterns are evaluated and classified using the class discrimination rules.

Using the RBP approach, the perception of shape identity invariant to geometric transformations is determined by the types of unary and binary features encoded by the observer. For example, interpart distances are invariant to rotations and translations, while part contrasts are invariant to translations, rotations and dilations. However, at this stage, no formal experiments have been reported (in 2D or 3D) which systematically investigate the types of unary and binary features which are probably used to solve such recognition problems. Such

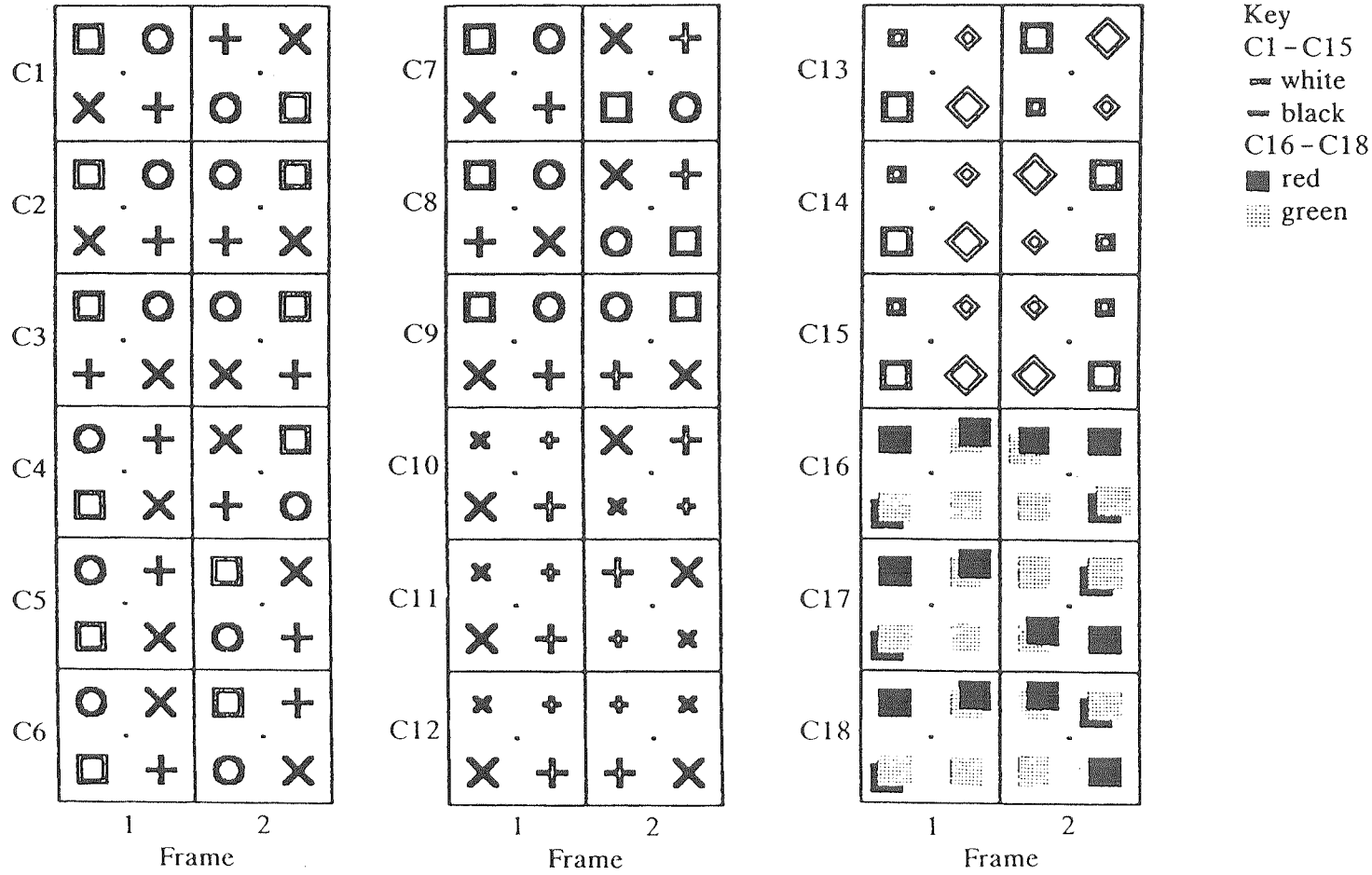


Figure 8. Frame 1 and Frame 2 for each of the 18 experimental conditions (C1–C18) for apparent motion. Grey scales stand for the different colours used on each pattern. For frame 1 (left) in C16–C18 the squares correspond to: red (R), red on green (R/G), green (G) and green on red (G/R) in clockwise order from top left. Frame 2 (right) squares were (in same order): C16: R/G, R, G/R; C17: G, G/R, R, R/G; C18: R/G, G/R, R, G. In all cases apparent motion between all possible pairs (12 possible motion paths) was recorded and analysed in terms of feature similarities and other constraints. (From Caelli *et al.* 1993.)

experiments are necessary to determine the perceptual-cognitive bases for the types of processes described above.

3. CONCLUSIONS

In this paper a number of issues have been argued. First, there is no single subsystem in visual information processing for form recognition. That is, the extraction, processing and identification of form covers a large number of visual routines from low-level pixel correlators to the generation of high-level symbolic representations of image parts and their relations. Results and interpretations of visual function have been combined into the multi-level system shown in Fig. 2. This description argues, as in this paper, for different levels of correlation occurring in vision. This view is particularly topical in the texture-processing literature where the first two of these correlation levels (see Fig. 2) have been proposed to be necessary to explain many texture discrimination results (see Caelli, 1988; Gurnsey and Browse, 1989). Another way of viewing these subsystems is that early processing is focused on 'producing chunks' while latter stages are concerned with 'interpreting and using chunks'.

This work has focused on using filter theory and standard correlation techniques to represent what may be processed in images by human observers. However, similar and possibly more salient representations for encoding, segmenting and defining structural information in images can be derived by the use of differential geometry. Recently, this has been investigated by Barth *et al.* (1993). In this form the intensity image is treated as a geometric surface and the more salient regions correspond to those which have non-zero Gaussian curvatures. Such regions correspond to corners, etc. and depict the regions of low spatial correlation as discussed earlier. This alternative view is mentioned to emphasize that there are many different ways of formally representing the visual processing claims proposed here.

One of the major pardigms of neurophysiology and psychophysics over the past century has been the idea that specific tasks such as detection, discrimination and recognition are actually *localized* along this 'pixels-to-predicates' pathway. The position put forward here is that this is not the appropriate conceptualization of vision. Most psychophysical experiments involve a conscious observer with prior knowledge of some sort or another and trained to solve, in the most general sense, an image interpretation problem. In this sense, then, such tasks use all the processes described in this paper and illustrated in Fig. 2. The channel models, neuron doctrine and signal detection paradigms have been included all within the low-level pixel-correlation models. This is because all these approaches to understanding spatial encoding are based on the notion that there exist, within the visual system, analogues to specific intensity profiles which are more important to biological vision than others and that their presence is detected by correlation-type mechanisms. It can be argued that, although psychophysical tasks have been constructed to investigate and determine the types of profiles that may be involved in such spatial encoding, this is not sufficient for a full model of spatial pattern recognition—particularly for complex patterns, as, for example, occur in the recognition of handwriting.

Further, it is proposed that the very notion of pixel correlation has analogous

processes in higher-level vision. Examples of this include the 'feature-correlator' stage referred to in Fig. 2 which has been investigated by a number of workers over the past decade. Further along the processing pathway, our perceptual ability to correlate part and part-relational features, and to generate evidence for patterns from image part data is simply another form of correlation, and the 'parts-correspondence' approach to apparent motion discussed in the previous section is an example of this.

To further illustrate this point, techniques for invariant pattern recognition have been briefly reviewed and results have been pointed out on the limits of human invariant pattern recognition. Such results suggest that human pattern recognition is more appropriately characterized as a recognition-by-parts process which is fed by the low-level processing systems. The point is, however, that in first- and second-order correlation domains there is the basis for fully invariant pattern recognition using orientation and size-specific feature extractors, as shown in Figs 6 and 7. However, it seems not to be used directly by the visual system. Rather, it is more likely that it subserves the more conscious recognition-by-part matching procedures which are more powerful for complex recognition and interpretation problems but do not necessarily encode pattern parts and relations by invariant features.

Some comments should be made concerning 3D object recognition. In the machine vision literature, most approaches to this problem involve all of the processes defined above but these usually apply to range (depth) image data. What is not clear, in human vision, is where along the visual pathways intensity information is converted into depth information and what features, feature correlation and segmentation procedures are really used to solve object recognition problems. For example, claims by Biederman (1985) that observers decompose objects into superquadrics (geons), and by Hoffman and Richards (1986) that such segmentation at least involves breaks at lines of minimum curvature are, at present, insightful speculations.

Finally, in Fig. 2 the existence of top-down feedback for the selection and tuning of correlation processes is suggested. By this, it is simply meant that in most recognition tasks observers have prior knowledge of the visual environment, the range of feature states and even image regions of importance. Such knowledge clearly affects what is attended to in the image and what types of applications of the various correlation processes are to be used.

Acknowledgements

This project was funded by grants from the Australian Research Committee and the Natural Science and Engineering Research Council of Canada.

REFERENCES

Ballard, D. and Brown, C. (1982). *Computer Vision.* Prentice-Hall, Englewoods Cliffs, NJ.

Barth, E., Caelli, T. and Zetzsche, C. (1993). Image encoding, labeling and reconstruction from Differential Geometry. *Computer Vision, Graphics Image Proc.* **55,** 428–466.

Biederman, I. (1985). Human image understanding: recent research and a theory. *Computer Vision, Graphics Image Proc.* **32,** 29–73.

Burgess, A. and Ghandeharian, H. (1984). Visual signal detection. II. Signal-location identification. *J. Opt. Soc. Am.* **A1,** 906–910.

Caelli, T. (1988). An adaptive computational model for texture segmentation. *IEEE Trans. Systems, Man Cybernet.* **18,** 9–17.
Caelli, T. and Dodwell, P. (1982). The discrimination of structure in vectorgraphs: Local and global effects. *Percept. Psychophys.* **32,** 314–326.
Caelli, T. and Dodwell, P. (1984). Orientation-position coding and invariance characteristics of pattern discrimination. *Percept. Psychophys.* **36,** 159–168.
Caelli, T. and Moraglia, G. (1986). On the detection of signals embedded in natural scenes. *Percept. Psychophys.* **39,** 87–95.
Caelli, T. and Moraglia, G. (1987a). Is pattern masking predicted by the cross-correlation between signal and mask? *Vision Res.* **27,** 1319–1326.
Caelli, T. and Moraglia, G. (1987b). The concept of spatial frequency channels cannot explain some visual masking effects. *Human Neurobiol.* **6,** 63–65.
Caelli, T. and Nawrot, M. (1987). Localization of signals in images. *J. Opt. Soc. Am.* **A4,** 2274–2280.
Caelli, T. and Reye, D. (1993). On the classification of image regions by colour, texture and shape. *Pattern Recognition* **26,** 461–470.
Caelli, T., Hübner, M. and Rentschler, I. (1986). On the discrimination of micropatterns and textures. *Human Neurobiol.* **5,** 129–136.
Caelli, T., Liu, Z. Q. and Bischof, W. (1988). Filter-based models for pattern classification. *Pattern Recognition* **21,** 639–650.
Caelli, T., Manning, M. and Finlay, D. (1993). A general correspondence approach to apparent motion. *Perception* **22,** 185–192.
Dodwell, P. (1970). *Visual Pattern Recognition.* Holt, Rinehart and Winston, New York.
Dodwell, P. and Caelli, T. (1985). Recognition of vectorpatterns under transformations: Local and global determinants. *Q. J. Exp. Psychol.* **37,** 1–23.
Ferraro, M. and Caelli, T. (1988). Relationship between integral transform invariances and Lie group theory. *J. Opt. Soc. Am.* **A7,** 738–742.
Field, D., Hayes, A. and Hess, R. (1993). Contour integration by the human visual system: Evidence for a local 'association field'. *Vision Res.* **33,** 1–22.
Foster, D. H. (1978). Visual comparison of random-dot patterns: evidence concerning a fixed visual association between features and feature-relations. *Q. J. Exp. Psychol.* **30,** 637–654.
Gurnsey, R. and Browse, R. (1989). Asymmetries in visual texture discrimination. *Spatial Vision* **4,** 31–44.
Hoffman, D. and Richards, W. (1986). Parts of recognition. In: *From Pixels to Predicates.* A. Pentland (Ed.). Ablex, New Jersey, pp. 268–294.
Hoffman, W. (1968). The Lie algebra of visual perception. *J. Math. Psychol.* **3,** 65–98.
Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195,** 215–243.
Jain, A. and Hoffman, D. (1988). Evidenced-based recognition of objects. *IEEE Trans. Pattern Anal. Machine Intelligence* **10,** 783–802.
Julesz, B. (1962). Visual pattern discrimination. *IRE Trans. Information Theory* **IT-8,** 84–92.
Kolers, P. and Perkins, D. (1975). Spatial and ordinal components of form perception and literacy, *Cognitive Psychol.* **7,** 228–267.
Rosenfeld, A. and Kak, A. (1982). *Digital Picture Processing.* Academic Press, New York.
Simas, M. and Dodwell, P. (1990). Angular frequency filtering: A basis for pattern decomposition, *Spatial Vision* **5,** 59–74.
Suen, C. (1990). *Frontiers in Handwriting Recognition.* CENPARMI Press, Concordia University.
Ullman, S. (1990). Three-dimensional object recognition. In: *Cold Spring Harbor Symposia on Quantitative Biology, Vol LV.* Laboratory Press, Cold Spring Harbor, NY.
Uttal, W. (1985). *An Autocorrelation Model of Form Detection.* Erlbaum, Hillsdale, NJ.
Van Trees, H. (1968). *Detection, Estimation and Modulation Theory: Part I.* Wiley, New York.
Watt, R. (1987). An outline of a primal sketch of human vision. *Pattern Recognition Lett.* **5,** 139–150.
Zetzsche, C. and Caelli, T. (1989). Invariant pattern recognition using multiple filter image representations. *Computer Vision, Graphics Image Proc.* **45,** 251–262.