Tina Memo No. 2000-002
Internal working document. Presented as external seminar at University of Surrey 2001.

Solving Shape Based Object Recognition from a Computational Standpoint - Practical and Physiological Constraints.

N.A.Thacker.

Last updated
25 / 2 / 2002

This document forms part of the **Recognition and Intelligence** Series available from www.tina-vision.net.

2007-001	Retinal Sampling, Feature Detection and Saccades: A Statistical Perspective.
2006-008	Statistical Principles for Selection of Computer Vision Algorithms as Modules for Visual Perception - Show Me the Errors.
1991-001	Designing a Layered Network for Context Sensitive Pattern Classification.
1997-002	Supervised Learning Extensions to the CLAM Network.
1996-003	Tutorial: Algorithms For 2-Dimensional Object Recognition.
1997-005	Speechreading Using Probabilistic Models.
2000-002	Solving Shape Based Object Recognition from a Computational Standpoint - Practical and Physiological Constraints.
1995-004	Assessing the Completeness Properties of Pairwise Geometric Histograms.
1996-004	Robust Recognition of Scaled Shapes Using Pairwise Geometric Histograms.
1996-005	Multiple Shape Recognition Using Pairwise Geometric Histogram Based Algorithms.
2007-007	Automatic Identification of Morphometric Landmarks in Digital Images.
1999-002	A Feature Representation for Map Building and Path Planning.
2001-015	Colour Image Segmentation by Non-Parametric Density Estimation in Colour Space.
2001-006	What is Intelligence?: Generalised Serial Problem Solving.
1994-002	A Correlation Chip for Stereo Vision.
1995-001	Specification and Design of a General Purpose Image Processing Chip.

TiNA
WWW.TINA-VISION.NET

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Solving Shape Based Object Recognition from a Computational Standpoint - Practical and Physiological Constraints.

Abstract

The approach to shape recognition based upon the use of 'Pairwise Geometric Histograms' (PGHs) [9] was originally developed for a project which aimed to make use of semi-realistic neural network models as the basis for a shape recognition system [18, 19]. In the course of publishing the PGH work however, it became clear that we could not continually relate the biological motivation in papers intended for a computer vision audience. This document attempts to explain the many constraints which were taken into account during the formulation of this approach, and how the method was intended to match realistic neuronal mechanisms while at the same time solve the computational complexities associated with the matching of boundary shapes in real images. Comments are included from Charles Leek, who criticises the ideas from the perspective of current views in psychophysics.

Introduction

The fundamental question in the subject of computer vision is; Can we understand the process of visual recognition and build systems which have similar capabilities?

The approach to object recognition in the vision literature can crudely be divided into two camps. The first is the standard computational vision approach which involves the building of object models and then using information on geometric constraints between features as the basis for object location and recognition ([4, 13] Guard, Pollard). The second is typified by the works of Poggio, Siebert and Adelman [15] where the concept of an object model is completely abandoned and replaced by a series of standard views. Recognition of non-standard views is achieved by a process of interpolation. This whole process can be designed and implemented in a neural network architecture, which being parallel should provide a scalable solution from small to large scale recognition tasks. The advantage that this second approach has is that it is much easier to construct standard views of an object from example 2D vision data than it is to construct geometric models. Though these methods do not provide quantitative location and orientation information the approach is much more suited to the development of a learning system. Although this advantage has been demonstrated, applications have been limited to unrealistic recognition problems such as stick objects and silhouettes. There have been no real attempts to deal with the problems of noise and segmentation encountered in the extraction of image features from real vision data. Largely this is because the input representations used have been very primitive and have not been designed to cope with these problems. A common approach is to model all possible variations of the image data, such as the Active Appearance Model (AAM). In doing so considerable effort must be expended in describing data which may be of no real use to the recognition process. There seems to be a need for a more sophisticated representation of shape designed to cope with the above problems. It is natural to extend the geometrical constraints used for model based matching for this task and this is the main subject of this paper.

Constraining the Problem

The human brain is complex and a model for visual recognition will require some constraints on potential designs if we are to make headway. Much is known regarding the microscopic bio-chemical structure of individual neurons (eg: Hubel and Weisel). Equally, quite a lot is known regarding the macroscopic modular structure of the brain (eg: Penfield et al.). However, neither of these levels of description allow us to really understand the computational processes involved in scene interpretation. Human cognition is an area which requires expertise from several disciplines, many branches of science could now be thought of as defining 'understanding the brain' as the main goal for research. In the situation where we cannot be expected to be experts in all of these areas, all we can do is to try and identify the main mechanisms which contribute to 'data processing' within the brain.

We start by assuming modularity, that separate modules exist for the purposes of representation of features in external images and for recognition. Second, that the model should not encode and transmit data in a way that could not be encoded and transmitted by neurons. Our third constraint is that we expect this modular system to make optimal use of data, we will explain how we define optimal below.

The human vision system is in the business of prediction. It must learn to make valid decisions from images in a way which minimises mistakes. We have an opportunity here to learn from the experiences of computer vision researchers. It has been recognised by many that the only way to reliably build computational analysis systems is to use quantitative statistics. The reason for this is that, the optimal performance will occur when this system correctly assesses the probability of a particular outcome given the data, this is statistical decision making. We might hope that the process of evolution has endowed us with near optimal performance in this respect.

Demanding a statistical interpretation of algorithms it is a very useful way of assessing the limits of what can be extracted from data. It provides a best case scenario for an object model hypothesis. Any model of the human visual system which cannot solve the task with the available data cannot be correct and the correct model must use additional information. This implies a natural direction for the consideration of top-down versus bottom-up models and for fusion of results from candidate modules. We should attempt to solve processing tasks with individual modules using specific feature types first and fuse results from multiple modules when this fails. We should attempt bottom-up approaches up to the point that we find top-down information is needed. Solutions which do not use top-down information are more generic, as they do not require prior experience of the data in order to extract information. The third constraint is therefore equivalent to requiring that workings of the system must have identifiable statistical principles.

Our basic constraints on any suggested model are therefore modularity, neuronal plausibility and statistical optimality, but we must also not forget the goal. We must seek a computation model with real capability with regard to the intended task, such capability can only be demonstrated by simulation. Ultimately, we would expect the model to display aspects of human performance and the modularity of the system should be consistent with results obtained from psychophysics experiments. However, we should not jump too quickly to demand identicality with human perception. Despite our everyday familiarity with the task, vision is clearly complex, whole books have been written on the subjects of vision, visual processing and computational vision. Even if we restrict ourselves to specific aspects of visual recognition capability, designing a computational process with useful capabilities is difficult.

We must insist from the outset that the system deals appropriately with illumination changes and is invariant to position. Depending upon the competence modelled we may also need to require scale and rotation invariance. We must also remember that objects can be rotated within the plane of the image (so that information content of the image does not change) or rotated out of plane of the image, so that the available information can not only change, but change discontinuously due to qualitative changes in the visible features. Finally, there are confounding factors such as clutter and occlusion. Only a computational approach which can be shown to deal specifically with these issues can be accepted as a real candidate for a module in a model of visual perception. We must therefore seek representations of features which are automatically computable from image data in a way that such factors can be dealt with appropriately. It is too easy to start by defining object characteristics at the feature level, without considering first if a real system could produce these features reliably from an image.

In order to explain the motivation for our suggested model we will start by looking at the constraints imposed by neuronal function, how these might be related to a statistical decision process and finally how to take data present in an image and extract from it features in a suitable form for use in such a system.

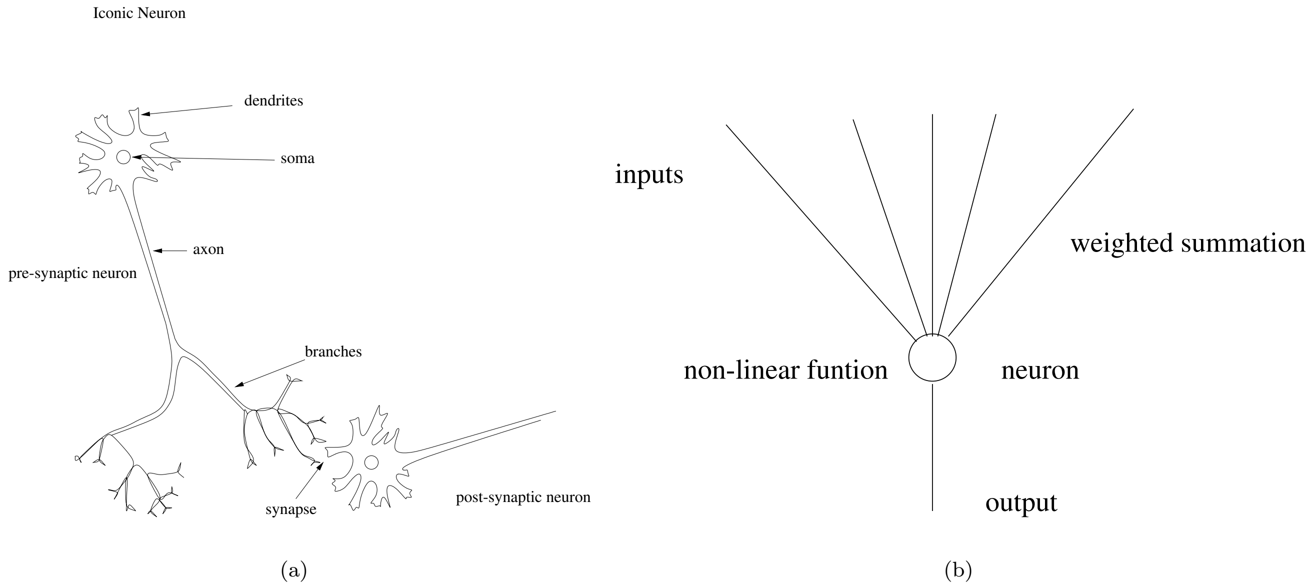
Neuronal Models

The aim of this section is to discuss the following question; If the brain is performing statistical decision making with frequency coded signals, what is the equivalent statistical function?

The prototypical neuron is often defined as a structure for generating and transmitting neuronal pulses (Figure 1(a)). The chemical mechanisms which underlie this are described in terms of a sodium/potassium imbalance which is maintained by a sodium 'pump'. The signals generated are transmitted along axons in the form of discrete pulses in the range of several to a few hundred hertz. These pulses are mediated between neurons at synapses via chemicals such as acetylcholine. The cumulative response to many such pulses impinging on the synapses of other neurons is yet more frequency coded signals. The details and structure of neurons varies greatly around the brain, but here we will assume that the important aspect of all neuronal firing is the representation of information as frequency pulses.

Work by Kohonen [11, 12] is a very popular model of neuronal function. It promoted the idea that the mapping between data and nodes in a neural network must depend upon a dot-product weighting function (Figure 1(b)). This is physiologically plausible, given the known anatomy of neurons, whereas difference based measures are not. A Kohonen-like algorithm assumes a particular property of the learned space. Specifically it assumes a limited dimensionality (typically 2). Topological mapping is observed in physiology, but may only be used to span an input vector space with available resources. It may not have much further computational value.

The artificial neural network ART (Adaptive Resonance Theory), designed by Grossberg [6, 7], reflected detailed physiological studies of firing rates in brain tissue. Grossberg says that local connectivity is mainly used for a competitive mechanism which performs the (winner-take-all) decision process. ART relies on frequency-coded information (such as neuronal firing rates), responding directly to a certain number of input impulses over a given period of time. Worth mentioning is the assumption that the results of the physiology studies reflect firing-rates in the brains memory modules, as this forms the basis of Grossberg's model. ART integrates the notion of a frequency-encoded representation with the concept of dot-product weighting as described by Kohonen.



(a)

(b)

Figure 1: A prototype 'real' neuron and its computational analogue.

The advantage of such a mechanism is that it allows a zero weighting to be ignored by the network's final output, a desirable characteristic if a system is to function in the real world where data is often cluttered or incomplete. Grossberg also developed the concept of direct competition between nodes in a network and the idea of 'winner-takes-all' as a plausible strategy for this kind of model.

Artificial neural networks such as ART are already functioning in a way which makes implicit decisions regarding the processing of input data. For example pattern recognition, or the requirement to add a new node to the network. In order to make these processes more explicit Thacker et al [18] devised an algorithm, CLAM (Contextual Layered Associative Memory), as a statistical re-engineering of ART. We may have naively expected the brain to implement a chi-square measure for the comparison of Poisson distributed data, as this is the most frequently recommended statistical measure for this task.

$$\chi^2 \; = \; \sum_i \frac{(h_i \; - \; t_i)^2}{t_i}$$

where h_i are frequency measures (eg: histogram bin values) and t_i is the theoretical prediction. Unfortunately, this difference measure (as explained above) would not be physiologically plausible. This measure is however, only a comparison of Poisson distributions as an approximation to a Gaussian distribution with variance t_i. In fact we have been able to show [21] that a better approximation to the probability of getting the histogram data given the model is given by;

$$log(P) \; \approx \; \sum (\sqrt{h_i} \; - \; \sqrt{t_i})^2 \; = \; |\mathbf{h}| \; + \; |\mathbf{t}| \; - \; 2 \sum_i \sqrt{h_i t_i}$$

The final term here (which does all of the discrimination for sets of dissimilar patterns $\mathbf{t}$) can easily fit within the dot-product restrictions of neuronal tissue. In addition, this form permits the storage of all template histograms $\mathbf{t}$ in normalised form;

$$\sum_i t_i \; = \; 1$$

as required by these network architectures.

In this network architecture the dot-product metric can be directly interpreted as a maximum likelihood measure for the comparison of frequency coded signals via the above relationship. The winner-take-all interpretation of neuronal behaviour can then be determined by computing the probability that a particular pattern ($\mathbf{t}$) would have the largest response to the dot-product measure in competition with its neighbours $P(\mathbf{t}_j = max|\mathbf{h})$. In addition, the statistical method of probability recoding can be used to compute the conditional probability that a given piece of data is consistent with a particular hypothesis ($P(C|\mathbf{h})$) by summing over the layer of neurons representing histogram templates $\mathbf{t}_j$ [19].

$$P(C|\mathbf{h}) \;=\; \sum_j P(C|\mathbf{t}_j)P(\mathbf{t}_j|\mathbf{h})$$

where $P(C|\mathbf{t}_j)$ are links to a classification layer, established during the training process. Once again this calculation is in the form of a dot-product and the output $P(C|\mathbf{t}_j)$ can be represented as a firing frequency, and therefore be computable on a neuronal analogue. Using the above calculations CLAM is able to compute the reliability of the answer to the statistical question being asked based on the available data. CLAM is designed to be self-generating - it is capable of growing its own nodes if the assessed need for these exceeds some pre-determined statistical threshold. The CLAM network is also designed to support hierarchal layering of classifiers, for example classification of 'sets of .. sets of' features.

The main feature of this approach is that the natural relationship between signal and variance inherent in Poisson statistics can be used to construct a measure which takes account of the accuracy of the encoded data. Such information cannot be ignored if neuronal function is to make meaningful statistical decisions. The advantages of frequency-coded neural networks are clear. Each impulse is as statistically valid as any other and thus carries the same level of significance. Taking a statistical view of the workings of neuronal function thus forces us to form rather strong opinions of the way that data is encoded and what it should represent. In addition, as the approach is based upon sound statistical principles, and there is no limit to the amount of resource available (due to node recruitment) the learning process is not restricted by a priori choice of architecture and the network is **guaranteed to learn** whatever is presented to it.

The above is the simplest possible analogy to statistical decision making in a frequency coding system. We should also be prepared to ask the question; Is this all that the brain does? Other aspects might involve relative timing of pulses and inhibition. Timing in particular might allow these ideas to be extended into the time domain so that the same neuronal decision processes could be applied to temporal patterns as well as static ones. If this is the main expected role for timing then we could simply ignore this extension for now. Aspects involving making use of particular non-zero phase relationships between correlated signals would present much more of a challenge to interpretation. The possibility of inhibition can also be finessed if we simply say that inhibitory connections only play a role in the 'winner-takes-all' mechanism, but are not used in general pattern representation.

Finally, from the purely pragmatic standpoint, it would probably be prudent to start by assuming that simple comparison of frequency samples in a pattern matching task is the main computational role of brain tissue, and only seek more complicated interpretations on the basis of need.

Design Features

Now that we have defined a biologically plausible statistical approach which is capable of learning we must decide what data to provide it with.

An important set of constraints on any model of visual object recognition concern extraction of consistent object-related information from varying data. The field of machine vision can tell us something about the practical difficulties of trying to construct such a representation. As we have already mentioned above, practical models must solve the problem of object feature segmentation from input data that varies with changes in the object's orientation, proximity, relative illumination, etc.. Characteristics of the object that remain extractable despite these changes are known as invariances. An ideal object representation would be invariant to all possible transformations and distortions that may affect the input image and complicate the recognition process. The more variations there are on a particular object representation, the more computational resource (brain tissue or memory) will be needed to solve the visual interpretation task.

To solve the problems raised by the mutual occlusion of objects in a scene, something frequently found in real-world situations, a strategy is required that deals with missing input data; one which focuses on the available data and is not compromised by any omissions. A related and very important factor is the global or local nature of the representation. Any system having to deal with data that is occluded or cluttered or in some way missing cannot utilise a global representation: if parts of a global feature such as a major axis of symmetry are missing, that entire feature is non-computable. The use of multiple redundant local representations has long been accepted in the machine vision literature as the most robust computational solution to this problem.

One of the greatest obstacles in trying to construct local representations for object recognition is the 'aperture problem'. If one of these object fragments is a straight line, then all positions along the line will be indistinguishable when viewed through a restricted aperture. The aperture problem prevents unambiguous point-to-point matching of locations along continuous linear features for both stereo correspondence and object matching tasks. Obviously, for the system to be of practical use, it must withstand such effects of stimulus ambiguity; therefore, taking account of the aperture problem is essential for the development of a practical object recognition system.

Changes in light levels and/or sources will affect the solution of a visual object recognition problem for a frequency-coded neural network in just the same way as they do for the human brain. Shape must be acquired from the available information, whether that be systematic changes in brightness (denoting edges), texture, shading, etc; and this information must be as near as possible to illumination-invariant. Clearly a system that can only extract features from a well-lit environment with unvarying light sources is unlikely to be of much practical use.

A further invariance characteristic that should be extractable by a practical object-recognition system concerns object transformation, or, more specifically, rotation, both within and out of the image plane. Imagined spatial transformation of objects or features in two and three dimensions is essential to the matching and recognition processes if the system is not to re-learn an object each time it is encountered from a new viewpoint. As such, this is as much a problem of practical storage and retrieval as it is of generalisation. Out-of-plane rotation invariance is especially difficult to achieve, because (as we have mentioned above) the available information actually changes after such a translation in a way that it does not during an in-plane rotation (where information may be encoded at a different point in the image but is never lost). This leads to a lack of consistency in the information available for the recognition task. Psychology studies of response times [Leek 1998] give clear evidence that the strategy used by humans is to interpolate between learned vantage-points, which is efficient both in terms of processing time and storage space.

The approach adopted here is to make use of an artificial neural network to perform multiple view based recognition similar the work of Seibert(91). The set of views gathered into a node are usually defined on the basis of some difference metric between the view representations, Chakravarty & Freeman [8]. The resultant network constitutes an appearance based representation of the object, similar to an aspect graph, Koenderink & Van Doorn [25], and has as its nodes representations of characteristic views or aspects of the object. The links between the nodes of the graph define the allowable view transformations that may occur as the particular object rotates or as the view angle changes. These will typically correspond to 'catastrophic events' where an object feature becomes occluded or uncovered, though this need not be the case.

Lastly, recognising the real size of an object is also very important. A practical system must allow a certain flexibility in the interpretation of scale if the same object at varying distances is not to be stored as multiple differently-sized objects. This does not just concern efficiency of storage but is a practical essential if we are not to view objects naively upon successive presentation.

Further Observations

Taking the grey level image as input one requires a representational scheme which makes explicit the relevant features of the object's shape but which is invariant to image properties that are not used in recognition. This involves extracting information regarding the defining contours of the object and transforming it into a compact, general shape description. In order to be of use as input to a recognition system, the descriptions produced using the representational scheme should ideally possess a number of desirable characteristics. We will take our cue here from several decades of experience in the computer vision community.

The representation should be robust to the types of noise expected in vision data from any real, unconstrained environment. This includes the addition and loss of visible features due to changes in lighting conditions and extraneous data from other objects. Some object detection methods (eg: the face recognition system of Sung and Poggio [17]) explicitly remove planar variation in grey level in order to remove the effects of strong directional lighting. For shape based recognition (of man-made objects, for example), insensitivity to such problems is more easily provided by representations based on local shape features, as global features tend to be adversely affected by fluctuations in the available data.

In addition to these constraints, the scheme should also conform to a number of criteria detailed by Marr [14].

Firstly the representation must be accessible in the sense that descriptions can be quickly and easily computed from the information available in the image. This is partly the justification for using a view-based recognition strategy as representations based on an object centred coordinate system are notoriously difficult to construct. Secondly the representation must exhibit versatility in being applicable to arbitrary shapes if it is to be considered a useful step towards the goal of generic vision. Thirdly the representation must have sufficient descriptive power to allow discrimination between all dissimilar objects, theoretically up to the differences due to the required invariance properties. Fourthly the representational scheme should be capable of producing descriptions which capture the large scale similarities between shapes. If small perturbations in projected shape result in greatly differing descriptions then the recognition based on these descriptions will not be stable. Finally, in contrast to the stability requirement is the need for the representation to be sensitive to small scale details of the shape so that fine resolution recognition can be performed.

The requirements of later processes always impose constraints on the kinds of information that should be represented in the shape description. The fact that the shape description required here is to form the input to a multiple view based recognition system provides additional constraints to the more general ones listed above. The smooth, exclusively quantitative changes that occur in the visible feature set as an object rotates between characteristic views should produce smooth, relatively small changes in the shape description. If these views are to be partitioned as aspects, along the lines suggested by Koenderink & van Doorn [25], then the shape descriptions for views either side of a qualitative change should be sufficiently different to signal that a catastrophic event has occurred. If on the other hand the views are to be partitioned purely on the basis of some difference metric, Seibert [16], then the shape description can afford to be relatively insensitive to qualitative changes in the set of visible features.

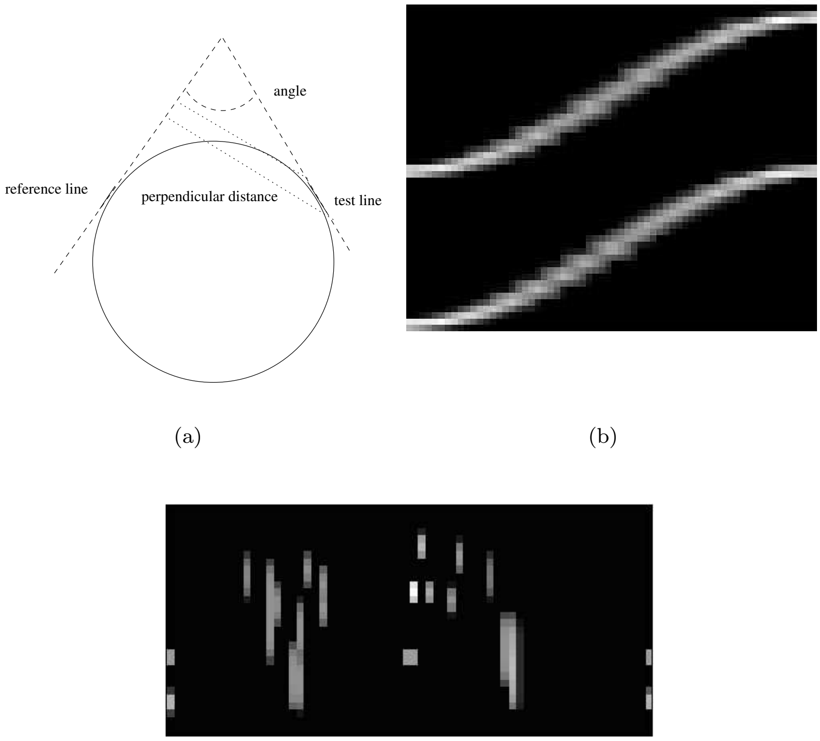
There are a large number of possible constraints that can be used to define the geometric relation that holds between two oriented line segments. Consequently there are many pairwise representations that can be constructed on the basis of these constraints. Bray & Hlavac [5] provide a formal analysis of the properties of a number of local geometric constraints, together with a list of qualitative criteria for their assessment. Various combinations of these constraints can be used to form representations of shape with differing invariance characteristics. However, it should be remembered that it is the statistical characteristics and not the invariances which are the most important feature of the data once we start passing the data into a pattern recognition system.

Descriptions constructed within the proposed representational scheme should be based on a set of local image features that somehow define the shape of the object and which can be reconstructed at all rotations, translations and scales. A sensible choice is 2D oriented line segments, for several reasons. Firstly lines are reliably reconstructed at the same position and orientation relative to the object regardless of its occurrence within the image. Line segments are also robust to details of illumination, as evidenced by the success of feature based stereo matching algorithms. Secondly there is good evidence that biological vision systems utilise such features at early stages of visual processing, Hubel & Weisel [10]. Thirdly, and somewhat conveniently, this is what our computer vision system delivers, Porrill et al.(89).

Pairwise Geometric Histograms

We start by assuming that in a modular system we will at least need a module which can extract edge features and use them to represent and recognise configurations of edges as shapes. PGHs are an attempt to satisfy all of the design features discussed in the previous section, to form part of a workable model for visual object recognition. Their construction is also consistent with the statistical constraints generated by the CLAM work. They encode, for any feature (used as a **reference**), the perpendicular distance (in the image plane, measured in arbitrary units) to all other features and the angle of orientation of those other features relative to the first feature (Figure 2 (a).) This is done in a way which allows the approximation of edge data up to the limit of a specified accuracy (Figure 2(b)), which is implicitly encoded in the representation via a blurring. The fact that any arbitrary shape can be approximated by a set of straight line segments means that the application of the representational scheme is not restricted to polygonal shapes but can also be used on linearised smooth curves, as illustrated in Figure 2(b).

Any recognition system must define a significance measure for the presence of a specific feature in an object. In this case the similarity measure used in networks can be analysed to understand the statistical nature of an input component. To make correct use of the network's internal similarity criterion, this entry must take the form of a probability, defined as the relative likelihood of there being an entry in that particular region of the representation space given the noise on the measurement system. Exact calculation of the probabilities would take an excessive amount of computation but an adequate approximation can be made in what follows using simple blurring. Since the values in the histogram are evidence for the presence of features in the image at a certain geometric relationship,



(a)

(b)

(c)

Figure 2: Construction of a Pairwise Geometric Histogram (PGH) (b) for a line fragment from a circle (a). A more typical histogram from a real scene (c).

the value of the entries should be related to the importance of the lines in forming the shape of the object to be represented. This is achieved if the entries are distributed such that the total entry made in the histogram is equal to the product of the lengths of the two lines. As the reference line is of fixed length, this regenerates a process which is directly analogous to Poisson sampling. This approach is therefore suitable for use with a likelihood based similarity measure which compares histogram entries on the basis of a square-root dot-product.

Encoding the measurement error is the first stage in achieving robustness to image variation, by explicitly matching the statistical matching metric to the expected noise mechanism we automatically achieve robustness to measurement noise (repeatability error). However, as we have mentioned above, we also need robustness to more general object imaging artefacts, such as occlusion (or shadow). No representation of shape can be completely invariant to loss of data (due to for example occlusions), but what can be done is to ensure that the statistical similarity measure used during matching is not invalidated by this process. The inclusion of the angular dot-product element in PGHs is sufficient to overcome the problems of occlusion and scene clutter: if the comparison of input image and stored representation yields a zero value for a given feature or line fragment, the dot product is zero and that particular feature is ignored for the purposes of comparison in that instance. Further, the resulting similarity function is equivalent to what would have been constructed if we had intended to only compare the visible features. By utilising only local line fragments and by not reconstructing global features such as major axes at this initial stage, PGHs bypass the possibility of occlusion or image fragmentation as impediments to recognition. Further, by encoding perpendicular distance **within the image plane** in arbitrary units, PGHs present a solution to the aperture problem. In particular, histograms constructed from fragments of the same extended line will give identical histograms up to a normalisation factor (which plays no role during selection of the best match). Normalisation of histogram patterns during learning further ensures that fragmentation does not generate multiple instances of a stored pattern. The angular aspect of the representation solves the problem of in-plane rotation as, for any feature, this is computable with respect to all other features of the object, but without the need for establishing a fixed point of reference.

Information encoded by PGHs is given in frequency-encoded space (as explained above) and is thus suitable for use with an algorithm such as CLAM. In addition it satisfies almost all of the design criteria specified and is consistent with the representational requirements suggested by previous researchers in computer vision. The main criteria which are not satisfied are out-of-plane rotation invariance and scale invariance. This is a consequence of requiring well behaved statistical behaviour, and would appear to be impossible for any representation. The possible space of patterns has been shown to be consistent with the numbers required to emulate human performance [2, 3] We have also shown how sets of histograms can be optimised for the recognition of multiple objects [1].

Comparison with Preliminary Psychophysics Results

Neural networks complete all of the above processes in parallel. It is not clear whether the human brain works in a similar fashion although the well-documented linear relationship between 'mental rotation' and subject response times suggests the existence of non-rotation-invariant representation [23] as well as an orientation-invariant representation such as that encoded by PGHs.

PGHs do not encode edge data in an out-of-plane rotation invariant or scale invariant manner. To do so would mean that measurement accuracy varied with viewpoint, which would undermine the statistical stability. Despite these limitations, PGHs can be shown to encode a significant proportion of the required invariances in a way which is complete [20]. By this we mean that the original object can be reconstructed from the set of histograms which describe it. This has several important consequences. Firstly, as no information is lost in the representation, use of the correct statistical recognition methodology (such as CLAM) should result in an optimal recognition process. Secondly, invariance characteristics which would appear to destroy this completeness could not be encoded in an optimal system. Such an encoding would result in potential ambiguities in shape representation and recognition. We currently believe that scale invariance would fall into this category. If this is true, the PGH would appear to be the best representation achievable for arbitrary rigid shape. Specifically which shapes to encode is still a research question, but one which may have some answers from psychophysics experiments. Any model of the human visual recognition process will likely require a computable feature grouping process. For example, systems designed to fuse information from bounded regions or surfaces will still require robust extraction of these features. PGHs may be considered a suitable candidate for this process. CLAM can also be considered as a valid computational approach for the hierarchal classification and recognition of features, feature groupings and surfaces.

The computational approach taken by the recognition of rigid shape using pairwise geometric histograms permits the construction of various alternative histograms with different sets of invariance characteristics. These include mirror symmetry, contrast invariance, in-plane rotation invariance and translation invariance. The existence of a method for the reconstruction of an object for some forms of histogram proves the completeness (and therefore optimality) of the approach, but only for some of these forms.

Psychophysical experiments with mental rotation of simple 2D line drawings can be interpreted as providing evidence for the hypothesis that the brain has two alternative pathways for recognition of such objects [24]. One of which appears to require a mental rotation process which can be inferred by both the reaction time and the varying area of activation. The existence of a mental rotation process infers that whatever mechanism is used for rotation invariant recognition cannot be the optimal way of recognising objects. At first sight this may seem to be at odds with our contention that it is possible to construct an optimal recognition system using a rotation invariant representation.

The key to understanding this contradiction lies in the proof of completeness. This requires that an edge density distribution profile can be constructed from individual line segments which can be oriented around the centroid of the object. In fact due to the fact that we are unable to specify the orientation of a simple line to better than a two fold (180 degree) ambiguity, we are unable to correctly specify the angular relationships between the set of histograms for an object without picking a fixed reference direction in the scene. Ie: an optimal representation requires specification of the co-ordinate system such as defining a direction for 'up'. This direction would be used at all times in the construction of the histogram so that a consistent orientation for the reference line could be selected from the two alternatives available. Representations based on lesser forms of histograms with orientation invariance, such as would be suitable for rotation independent recognition, will have representational ambiguities for some classes of object.

This model predicts all of the main characteristics of our observations so far. It implies that a rotation mechanism (which scans possible 'up' directions) would be required to optimally recognise objects which could well result in a time dependency. However, we might also reasonably expect that we might also have a rotation invariant recognition process. This can be constructed trivially by the additive folding of the full representation, which would serve to recognise objects which were sufficiently unique so as not to require a complete representation or alternatively identify familiar objects at novel orientations. As all objects could be represented and recognised in a rotation dependent manner, it would also predict that the selection of which approach to use would be based upon experience (ie: learning).

This computational model also makes some interesting predictions. For example, any other method which can uniquely be used to specify the orientation of a line (such as contrast or colour) may break the two fold ambiguity and allow the brain to once again recognise objects optimally without recourse to rotation. In other words, Tarr and Pinker's reaction time results for rotation dependent objects [22] may have only occurred because of the simplistic nature of the image data. Rotation effects may not be observable for anything other than line drawings.

Conclusions

This paper has introduced the main motivations for using shape descriptions based on pairwise geometric relationships as input to a multiple-view-based neural network recognition system. Experimental results suggest that such descriptions exhibit many of the characteristics desirable of a compact, invariant 2D shape representation. In particular the representation has the capacity to deal robustly with the kind of problems encountered in real vision systems. The approach is capable of representing smooth arbitrary non-complete curves of the kind extracted from grey level images. The representation is invariant to translation and rotation. It is robust to illumination changes, and copes well with occlusion and clutter via the use of a principled similarity metric which takes due account of data reproducibility and varying quantities of evidence. All of these properties are achieved by design and result in a scheme which has potentially far more scope than previous representations used for input to adaptive recognition systems [1] . Though we have published techniques designed to deal with scale changes [3], based upon interpolation between stored templates, we have not yet addressed out-of-plane rotation. It has long been envisaged that extension of the approach for use with arbitrary projected views of 3D objects could be achieved via use of an eigenvector decomposition of allowable deformations of individual histograms, without loss of the key design features of this representation. This same approach could also be used to deal with object deformation, in a manner analogous to the techniques used for Active Appearance Modelling (AAM).

Comments from Charles Leek

You are discussing two issues here;

The first is the neuro-biological question about frequency coding and neural implementation (and statistical optimality). Obviously, this makes a lot of sense to anyone working in cognitive neuroscience (although the references to the neuroscience work need updating (Penfield is a bit old hat and I can help with that). It is worth, though, bearing in mind that your description of frequency coding is based largely on the likely functional properties of the signal neuron (given the all-or-none nature of neural firing etc, and summation of input activation). At the level of systems neuroscience there is currently a lot of interest in providing a functional characterization of brain activity across networks of neurons, and interactions among them. Such as, for example, phase synchronization (e.g., in the gamma band around 30-80hz) and its possible role in feature integration, that is, the key issue of how different features of objects are 'bound' together [how shape is bound with colour, texture, location, scale etc]. One other issue that springs to mind (which you mention on p. 5) concerns the excitatory and inhibitory nature of neural activity among networks of inter-connected cells, and groups of cells. There is also, I suppose, the issues of thresholds and neural sensitivity related to synaptic transmission and refractoriness of the cell in relation to action potentials. I guess the question is to clarify the relevance of these other factors to the implementation of a PGH representation in the way you are suggesting.

The neurons modelled in the CLAM network attempt to predict frequency coded responses to particular sets of frequency coded inputs. Inherently such a model has a kind of scale invariance, in that sums of Poisson samples are themselves Poisson samples, so that the output could just as well have been computed from multiple neurons as a single one. Therefore although it looks like I am only considering one neuron I might just as easily be describing groups. From a computational standpoint however, it is easier to implement simulations with fewer computational units. Therefore the tendency is to take the minimalist approach when describing the computations and algorithms. I have heard some argue that the reaction times for some decision processes are so fast that there is no time available to form meaningful estimates of frequency on the synapses of a neuron. This has even been given as a justification for the need for phase sensitive processing (your later comments below) Such an observation can only be answered by saying that groups of neurons, perhaps even hundreds, when taken on mass, would be able to represent and utilise frequency encoded signals in much shorter time frames.

[1] It has recently been noticed that research in the area of image retrieval has begun to re-invent the use of histogram based geometry descriptions of shape. This work has yet to regenerate either the statistical insights presented here, or the proof of completeness needed to justify methods which embody invariance characteristics as a 'solution' to the shape recognition problem.

Although issues such as phase synchronisation would certainly add computational possibilities to the performance of network architectures my approach has always been to try to see what can be computed without such extensions in order to find if they are strictly necessary to solve the computational problems. To my knowledge the simple approach has not been rejected as a possibility, so there is no need to yet assume greater complexity.

Issues of details of specifically how the brain manages to arrive at a principled statistical computation are of course important. But first we need to understand what that computation is and why it is required. The argument I am advancing is that there is a simple computational process which has some statistical validity which can be derived immediately we decide to assume that the brain encodes information as pulses (frequency codes). (NAT)

A couple of specific questions:

a. I think you need to be more specific about how you define a module (page 3) do you mean a functionally encapsulated computation (which may or may not be distributed across adjacent or non-adjacent brain tissue)

Yes i think so. I am not in a position to be able to suggest where it may be located or distributed across tissue. In fact the rather general nature of the computation would suggest that there is little real advantage to be gained by keeping the neurons together at all. The computational form of the algorithms does not require a fixed structure. They could be placed anywhere within the limits of available connectivity, thereby simlifying the construction of a working system, but equally creating havoc for anyone trying to understand the brain by attributing behaviour to ever smaller regions. (NAT)

b. It wasn't clear to me why a dot-product weighting function is more physiologically plausible, given what is known about the anatomy of neurons. This might be because I don't have the background on dot product functions, but neither will the intended psychology audience. All I think I know is that dot product functions return a single value (but the relevance of that I found hard to follow).

It's just that if you accumulate charge pulses at multiple synapses and sum them on the neuron the computational form looks more like a weighted sum than a sum of differences. This is the computational form that the early work of Grossberg and Kohonnen argued for. I don't think this observation is in any way contentious. (NAT)

c. Could you provide more detail about the CLAM algorithm and how it works. Without reading the earlier papers this bit is hard to follow and the more detailed description of the PGH could be incorporated from your earlier work in the two Ashbrook, Thacker papers –

Yes certainly. CLAM is a self generating look-up system, which learns coded templates using the statistics of freqeuncy coding. Self-generation is supported using statistical decisions regarding the novelty of a pattern in comparison to those already stored. The dot-product similarity function makes this process robust to outlier data. Probabilities of firing rate are computed in accordance with Grossberg's winner take all system. These are therefore, in effect, predictions of relative firing rate. These probabilties are then used as components of the learning process and can be used, via some conventional probability theory, to make classification decisions.

The Ashbrook, Thacker papers give only evaluations (practical demostrations) of the method. However, there are only three things you need to know about PGH's: they are designed to meet the list of invariance criteria as outlined in this document (which are logically the maximal possible set); they provide a complete and flexible representation of feature based shape; and they are suitable for input to a pattern recognition system which takes frequency coded (ie: histogram) data. (NAT)

The second key issue in the paper is about the representational format for encoding shape (the PGH) in a manner that is consistent with your assumptions about frequency coding in the brain (and optimality).

A few questions about that section:

a. You argue on page 6 that out-of-plane invariance is difficult to achieve because of the changes that occur in the image features. Some discussion of so-called Non-accidental Properties (NAPs) is probably needed Biederman uses NAPs (parallelism of edges, co-termination of edges at endpoints etc) as a basis for the invariant recovery of his basic volumetric primitives (geons).

Yes good point. NAPs can be seen as a surrogate for defining those features which maximise our ability to discriminate objects. If we directly optimise the ability to recognise using a limited set of features, these features may in some respects appear to score highly in terms of their NAP-like qualities. However, this does not mean that this is the principle which leads to their selection. On the contrary, NAPs as a representation scheme have the disadvantage that there can be no guarantee that the information encoded is enough to uniquely reconstruct the object. In turn, this leads to the problem of potential recognition ambiguity and an inability to deal with the variety and number of objects present in the real world. Feature selection on the basis of NAPs cannot therefore lead to optimal recognition. Representational ambiguity is common in many cue based recognition schemes found in the computer vision literature. While utility can often be demonstrated on a small number of carefully selected objects (or data sets), they effectively throw the baby out with the bath water in the way they achieve invariance when it comes to general scenes. Such behaviour is often overlooked in computer vision on the basis of computation sophistication (eg: Zisseman's projective invariance) or computational utility (eg: Lowe's SIFT). NAPs could only be used in combination with other recognition schemes (say PGH's!) in order to fill in for its limitations. (NAT)

b. Regarding the psychophysical evidence. The story is pretty mixed regarding viewpoint-dependence and viewpoint-invariance. As you know, whether or not viewpoint effects are found seems to depend on a variety of stimulus and task variables in both *the 3D and 2D situation. At least for the 3D case, there is pretty strong evidence that viewpoint matters, with frequency of exposure being a key factor. Here there is some strong supporting evidence for the kind of model you seem to propose - that the viewing sphere is represented by a set of 2D 'aspects'. For the 2D (image plane) case the story is more complicated, whether or stimulus orientation has an effect on performance depends, not only on prior exposure, but also on object geometry. Contrast the Tarr & Pinker (1990) demonstration of orientation-dependence vs. orientation-invariance that you will already be familiar with. According to some people, orientation invariance is found when objects can be discriminated from each other by contrasting 1D ordering of features alone (as might be possible with 2D forms with at least one axis of symmetry).*

I would be very surprised if the brain had only one way of recognising objects. If you are looking for one simple way of describing all psychophysics effects I don't think you will ever find one. These experiments seem to be consistent with there being several methods designed for different characteristic recognition tasks. The overall human perception experience would therefore be rather difficult to predict unless you had models for all of them and their potential interactions. One thing is for sure, whatever the other modules are, no sensible computational theory should be based upon axes of symmetry or elongation. As this property is uncomputable for partially occluded objects and is therefore of no practical use in real scenes. If nothing else this paper needs to be able to make this point. The other point I'm trying to make is that we should be able to make arguments which tell us what the limits of performance are based upon an understanding of quantitative data analysis. This would give us the ability to determine when data fusion is necessary and some indication of the necessary components in the full system. (NAT)

c. The paper implies that the representational structure is based solely on locally specified relations among (pairs of) edges. And this is my understanding also from the other PGH papers.However, the empirical evidence highlights other shape primitives as having functional significance in human shape representations. Biederman has argued that volumetric primitives play a role (although, as you know, this evidence is not at all convincing). In contrast, we have recently found evidence that human recognition does seem to care about segmenting images into constituent 2D edge bounded (closed) polygons. I'm wanting to argue for human vision that these polygons might be used as candidate primitives to approximate surface layout (see also, in the machine vision domain, (Olivia) Camps, Huang & Kanungo (1998) Hierarchical organization of appearance-based parts and relations for object recognition. Proc IEEE Conference on Computer Vision and Pattern Recognition — which you might well already be familiar with.

It would be very interesting for me if you could demonstrate how a PGH representation could be used to encode such higher-order primitives and how that might be useful in representing local spatial configuration (through the perpendicular distance measure, presumably i.e., the perpendicular distance between spatially adjacent pairs of polygon. And how this might be incorporated into a CLAM-type algorithm. I've taken a lot of inspiration from the work of Camps and Kanungo (see above) on so-called 'appearance-based' parts. They use a form of minimum description length 'VMDL' solution to the problem of segmentation, which in their case yields essentially a set of aspects of individual polygons parsed under a range of illumination and viewing conditions. I have only recently come across the MDL idea from reading their stuff. But it occurs to me that it can be likened to the classic Gestalt ideas about good form (that is, we tend to perceive and segment images into the simplest and most regular forms possible).

Just because PGH is based soley on relationships between edges doesn't mean that I think this is all human vision system is capable of. This is intended to be one part of a reductionist model.

As for hierarchal recognition, PGH recognise patters of edges across an image, there is no restriction of what these need to be and I would not want to restrict this to volumetric primitives. I can't see any reason why the brain might not be dealing with any reliable projected grouping of features as an object part, even if they do not constitute an obvious volumetric entity. The reason I would like to keep this distinction is that a volumetric interpretation of the scene requires accurate segmentation of volumetric primitives prior to object recognition. Those who advocate such an approach should expect to be required to demonstrate a workable computational model, something which appears to me impossible given the available data. I would prefer to take another interpretation. That familiar 2D groupings infer surfaces. The recognition process can then proceed by combining 2D groupings in a view based manner. Surfaces, or at least their hypotheses, would not be generated prior to recognition but as part of final the interpretation process.

Though MDL may superficially appear to provide solutions with required behaviours it can have no foundation in quantitative use of probability as it is not invariant to the way the data is represented. It's origins as a theory are in data compression (where data is trivially always in binary form) and I think the ideas have been mis-applied in the context of recognition. Non-the-less it has become a hot topic in computer vision over the last few years. I would hesitate before setting off to try to explain this one in this paper. (NAT)

References

[1] F.J.Aherne, N.A.Thacker and P.I.Rockett,"Optimal Pairwise Geometric Histograms." Proc. BMVC 97, 480-490, Essex, 1997.

[2] A.P.Ashbrook, N.A.Thacker and P.I.Rockett, 'Pairwise Geometric Histograms. A Scaling Solution for the Recognition of 2D Rigid Shape.' Proc. for SCIA95, Uppsala, Sweden, pp271, 1995.

[3] A.P.Ashbrook, P.I.Rockett and N.A.Thacker, 'Multiple Shape Recognition using Pairwise Geometric Histogram Based Algorithms.' Proc. IEE Image Processing, Edinburgh, July 1995.

[4] A.Bray, Recognising and Tracking Polyhedral Objects, PhD Thesis, Sussex University, 1991.

[5] A.Bray and V.Hlavec, 'Properties of Local Geometric Constraints.', BMVC 1991.

[6] G.A.Carpenter, S.Grossberg,D.D.Rosen. ART 2: ART-2A: An Adaptive Resonance Algorithm for Rapid Category Learning and Recognition, Neural Networks, Vol.4, No.4, pp. 493-504. 1991.

[7] Carpenter, G.A., & Grossberg, S. (1987). A Massively Parallel Architecture for a Self Organising Neural Pattern Recognition Machine. *Computer Vision Graphics and Image Processing*, **37**, 54-115.

[8] I. Chakravarty and H.Freeman, Characteristic Views as the Basis for Three-Dimensional Object Recognition, SPIE, Robot Vision, 336, 37-45, 1982.

[9] A.Evans, N.A Thacker and J.E.W.Mayhew. 'Pairwise Representations of Shape.' Proc(I). International Conference on Computer Vision and Pattern Recognition (ICPR). The Hague August 1992.

[10] D.Hubel and T. Wiesel, Receptive Fields of Single Neurons in the Cats Striate Cortex, J.Physiology, 148, 574-591, 1959.

[11] T.Kohonen. The Self Organising Map, proc. IEEE, Vol. 78, No.9, pp 1464-1480. 1990

[12] Kohonen, T. (1988). The "Neural" Phonetic Typewriter. *IEEE Computer*, **11**, 22.

[13] D.G.Lowe, Perceptual Organisation and Visual Recognition, PhD Thesis, Stanford University, 1984.

[14] D.Marr, "Vision", W.H.Freeman and co. pubs. San Francisco, 1981.

[15] T.Poggio and S.Edelman, A network that Learns to Recognise Three-Dimensional Objects, Nature, 343, 263, 266, 1990.

[16] M.Seibert, Neural Networks for Machine Vision: Learning Three-Dimensional Objects Represenations, PhD Thesis, MIT, 1991.

[17] K.K.Sung and T.Poggio, Example-based learning for View-based Human Face Detection., IEEE Trans. PAMI, 20 (1), 39-51, 1998.

[18] N.A.Thacker and J.E.W.Mayhew, 'Designing a Network for Context Sensitive Pattern Classification.' Neural Networks 3,3, 291-299, 1990.

[19] N.A.Thacker, I.A.Abraham and P.Courtney, 'Supervised Learning Extensions to the CLAM Network.' Neural Networks Journal, 10, 2, pp.315-326, 1997.

4 Leek reference (1998?)

[20] N.A.Thacker, P.A.Riocreux, and R.B.Yates, 'Assessing the Completeness Properties of Pairwise Geometric Histograms", Image and Vision Computing, 13, 5, 423-429, 1995.

[21] N.A.Thacker, F.Aherne and P.I.Rockett, 'The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data.' Kybernetika, 34, 4, 363-368, 1997.

[22] M.J. Tarr and S.Pinker, Mental Rotation and Orientation-dependence in Shape Recognition. Cog. Physiol.,21, 233-282, 1989.

[23] R.N.Shepard and J. Metzler, Mental Rotation of Three-dimensional Objects. Science, 171, 701-703, 1971.

[24] M.A.Goodale and A.D.Milner, Separate Visual Pathways for Perception and Action. Trends in Neuroscience, 15, 20-25, 1992.

[25] J.J.Koenderink and A.J.Van Doorn, Internal representation of Solid Shape with respect to Vision, Biological Cybernetics, 32, 211-216, 1979.