Chapter 2: Similarity and Object Recognition

This section critically surveys historical and recent work related to the topic, with the aim of situating the present theory in various ongoing dialogs related to effective pattern recognition and the nature of biological perception. Given the heavy interactions between these disciplines, the assignments are perhaps somewhat arbitrary, but are simply intended to reflect the organization of the literature. Similarly, occasional references to other disciplines or technical topics covered in depth in later sections seem to be inevitable; it is hoped that the reader will exercise patience, and expect that such material will be approached again in a later section.

PERSPECTIVES FROM PSYCHOLOGY

The wealth of data from functional brain imaging has presented challenges to cognitive psychology and artificial intelligence, which for their first decade or so could presume that there might be universally effective algorithms in a given domain of perception or cognition. Given the distributed nature of processing revealed by imaging, a rough consensus exists across psychology and neuroscience that, for any given domain of cognitive and neural processing, diverse mechanisms come into play depending on the exact nature of the task faced by the organism. Cognitive processing is attributable to moment to moment shifting between large-scale brain states that interact with and control diverse linkages of distributed and local processing networks, not simply a sequential flow of information through static modules. Thus there is no *universal* processing flow for perception which is independent of the time course of the information presented, nor from the task context in which a perceptual act is embedded. This is stated only to reinforce the appreciation that such changes in the nature of the task may induce a different cognitive or neural network architecture, subtle changes to the task may change the picture considerably.

To be concrete, much of the literature on visual recognition discussed below is focused on rather contrived conditions of matching objects presented briefly in sequence. This task differs from a natural ecological embedding condition, involving access of short or long-term memory representations of objects with some meaning to the organism, and searching for them in a natural scene. Thus observations at the levels of psychophysics and neural activity may give only limited insight into other modes of object recognition. For match-no match tasks on rapidly presented objects, *preattentive* neural dynamics, which can rapidly make discriminations with minimal involvement from more complex representations, are likely to dominate. The psychological meaning of similarity in such a task setting and in the context of object recognition, will differ from other classic work, such as Gestalt era studies of similarity (Goldmeier 1972). In the latter, subjects make choices at their leisure between various drawings; their judgements shed light on perceptual issues of the interaction and dominance of aspects of form such as orientation, size, and spacing on grouping processes. Whether the same processes come into play in very rapid recognition process is an open question, but most of the computational

procedures to be examined here do not focus on issues of grouping, or other areas of figural processing which includes phenomena such as length distortions, illusory contours and size distortion illusions.

Similarity and Metric Spaces

Similarity plays a fundamental role in theories of knowledge and behavior. It serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations. Indeed, the concept of similarity is ubiquitous in psychological theory. It underlies the accounts of stimulus and response generalization in learning, it is employed to explain errors in memory and pattern recognition, and it is central to the analysis of connotative meaning

Similarity has a long history in mathematical psychology, with two major branches: set-theoretic and geometric. The emphasis in the present work is on *geometric similarity*. In this formulation, objects are identified with points in a space, with categories corresponding to volumes in the space. The *dimensions* of the space are identified with primitive features in the input space. Similarity is conceived of as proximity, and a space which supports a *distance function* or metric is a metric space.

Three properties serve to define a distance function as a metric:

1. The *identity* property asserts that an object should be most similar to itself.
2. The *symmetry* property asserts that the order of presentation should not affect the measure.
3. Finally, *triangle inequality* should be satisfied; two dissimilar shapes should not both be similar to a third.

While these properties must be satisfied to conform to the mathematical definition of a metric, it is less clear that they are relevant to human perceptual processes. Symmetry, in particular, does not hold; but examples where it fails are most readily drawn from the realm of semantic or conceptual constructs (Tversky 1977). To better model semantic information and the symmetry violations, Tversky proposed a set-theoretic framework which counts shared features and independent features to produce a quantitative similarity measure, but dispenses with the notion that the set properties need to be ”dimensionalized” .

Proponents of *prototype based categorization* for shapes have argued that this asymmetry is due to the fact that prominent features of a shape in memory establish it as a prototype, and the absence of this feature can quickly be detected. Edelman (Edelman 1999) argues further that the critique of metric similarity at verbal and conceptual levels is of limited relevance in the assessment of geometric objects, where the objects can be decomposed into objective primitives. The same assessment was made by Tversky and Hutchinson (Tversky and Hutchinson 1986), stressing that this is particularly true when physical stimuli involve a small number of dimensions. When a larger number of feature dimensions is involved, metric models become problematic and the set-theoretic alternative performs better. Uttal, in a review of similarity and categorization, stresses that the boundary between perceptual and semantic representations may hinge on the number of dimensions (Uttal 1988).

The metric space concept was enhanced by the addition of a local density factor by Krumansl (Krumhansl 1978), in a way that compensates for some of the weakness noted by Tversky. In this formulation, the distance between objects is modified by the number of nearby neighbors in the space so as to increase distances in densely populated areas.

Measuring Similarity Experimentally

How can perceptual similarity be measured in a psychophysical setting? A variety of methods have been developed.

When subjects are asked to perform tasks, three major methods of experimentally assessing the perceived similarity or similarity of representation are seen in the literature. *Reaction time* (also called judgement time) are measured, but in some cases short reaction times are interpreted as a sign of similar internal representation (Bower and Clapper 1989), while others interpret long reaction time as confusability (Mumford 1989). *Error rates*, particularly false positives or confusions, are often interpreted as indicators of stimuli sharing similar representations. A final method is to measure *interference*, which is essentially reaction times or error rate effects of attending to, ignoring, or holding in working working memory multiple stimuli.

Levels Of Categorization and Object Recognition

Similarity, in the broad context of category formation and object recognition, is normally framed in a discussion of the properties or features constituting a category. Seen broadly, several problems must be solved for effective and flexible object recognition[7] (Tarr 2000). One is the multilevel nature of categories. Very broad distinctions suitable for concise representation, naming and rapid recognition define the *basic* level of categories (Rosch 1975). *Subordinate* level categories are more fine grained and require more time for naming. The categories 'birds' and 'humans' are entry level, while 'blackbirds' and 'song sparrows' are subordinate categories of bird. A further level of categorization is the individual or exemplar level, where an individual sparrow or human could be identified.

The term *entry* level refers to the level that is accessed first and typically named when a subject encounters a familiar object. This is normally the basic level, but for some categories (notably faces) the subordinate or individual level is the entry level; for anomalous objects in a basic category, such as penguin among birds, the subordinate level may be named.

Another related issue in recognition is the variability in viewing conditions for objects. Objects may be obscured by other objects (occlusion), or parts of an object may

---

[7] The categorization section follows Tarr's review closely; the review provides a concise statement of issues and recent work including some relevant imaging and single neuron studies, but focuses on the structural vs. view based controversy to the exclusion of other important issues such as visual search, and overlooks additional single and multi-neuron studies which will be addressed in a subsequent section here.

shield other parts from view (self-occlusion). For non-rigid objects, recognition must account for variation in the configuration of parts. Viewpoint changes affect the retinal image of an object; changes in size, position in the plane, and rotation in depth.

Categorization level interacts with object recognition, implying the likely prospect of multiple subsystems. There is evidence that the ability to compensate for viewpoint changes depends on the categorization level (Edelman 1995). Subordinate level discriminations – those between very similar objects – increase the costs of recognizing unfamiliar views.

Another form of interaction is that certain stimulus classes, notably faces, seem to be interpreted chiefly at the subordinate level. There is a long history of claims of specific face detector neurons (Rolls, Baylis et al. 1989). Sub-regions of inferotemporal (IT) cortex, the putative cortical high-level pattern recognition area, were shown to be more active during face recognition tasks (Sergent, Ohta et al. 1992). However, by synthesizing a novel class of stimuli and intensely training subjects to make fine discriminations between members of this class, Gauthier and Tarr (Gauthier and Tarr 1997) make a strong case for an alternative interpretation: that it is chiefly *stimulus expertise* which results in a specific syndrome of configuration sensitivity and automatic assumption of the subordinate level in recognition, rather than the specific stimulus category. This may still be associated with specialized regions of IT cortex; the imaging research of Gauthier and colleagues suggests that such localization does occur (Gauthier, Anderson et al. 1997).

A final area of interaction between categorization levels and recognition involves the type of representation or features used in recognition. Structural description theories assume that *view-independent* or i*nvariant* features underlie the representation of objects, and there is evidence that features such as the major axis of a 3-D shape are used to make entry level categorizations. View based theories build representations from various local features extracted from separate learned views, and have historically been associated with the subordinate level. Advocates of *view-based representations* have recently claimed that basic level categorization can emerge in a natural fashion from the clustering involved in making subordinate level distinctions (Duvdevani-Bar and Edelman 1999). These two strategies are considered in some detail in the next section.

View-Based and Structural Description Theories: Strategies for View Independent Recognition

A growing body of experimental evidence now suggests that performance on recognition tasks is proportional to the distance from the nearest familiar view. This includes both error rate measures (Bulthoff and Edelman 1992) and recognition time (Tarr and Pinker 1989); (Tarr, Bulthoff et al. 1997).

The structural approach derives essentially from the early proposal of Marr and Nishihara (Marr and Nishihara 1978) that the ultimate task of object recognition is the recovery of 3-D structural relationships from the 2-D retinal projection. If such a structure could be derived, then recognition of the object would be largely viewpoint independent. Faced with evidence that recognition is not invariant but varied linearly

with distance from learned views for "paperclip" objects (Bulthoff and Edelman 1992), the recognition by components (RBC) approach (Biederman and Gerhardstein 1993) was refined to suggest that the maximal viewpoint invariance occurs when three conditions hold:

1. Objects consist of geon-like parts. Geons are a set of simple volumetric components which can be configured to produce a wide range of everyday objects.
2. These parts form qualitatively distinct configurations in different objects.
3. These parts are visible over the range of viewpoints for which invariant performance should occur.

These predictions were tested by substituting a unique geon for one of the cylinders in the chain, which resulted in nearly invariant performance. A subsequent study by Tarr and colleagues (Tarr, Bulthoff et al. 1997) revealed that the single geon case is exceptional, and generally performance falls off with distance from a learned view with three or five geons. This 'paperclip with added geons' image set from the Tarr group (henceforth denoted here as the paperclips+ set) was selected for study in the simulations described later in this thesis.

In contrast to the experiments just described, in my simulations the original gray scale images are reduced to silhouettes, reducing the 3-D information available by shading and occlusion in the raw synthetic images. This is done chiefly to simulate putative edge extraction mechanisms in early visual layers in an attempt to rely on form alone. Another recent recognition study by Hayward (Hayward 1998) found no significant difference in viewpoint-dependent performance between silhouettes and shaded, part-boundary-visible versions of objects, indicating that features in the boundary contour are largely responsible for recognition and changes in performance.

Network Implementations of View Interpolation

The view based approach has been developed extensively with feed-forward neural networks, stemming from a general strategy first described by Poggio and Edelman (Poggio and Edelman 1990). This strategy is essentially view interpolation by regularization, or by normalization[8] in Tarr's terminology (Hayward and Tarr 1997). Normalization refers to the concept that different, perhaps novel views are mapped by some computational process to a representation derived from the trained views. This mapping occurs by transformation of several views of an object in a high dimensional measurement space to a lower dimensional shape representation space. A learning process operating over the presented views (e.g. the adjustment of network parameters) ensures that the transformation approaches the same point in the representation space for all trained views.

The dimensions of the measurement space correspond to an assembly of tuned filters. This transformation occurs by approximating the statistics of activation and their changes with basis function units (Poggio and Girosi 1990). The statistics are captured

[8] I will use the term normalization; regularization implies a certain underlying mathematical approach is used (Poggio and Girosi 1990); a major result in the present work is to demonstrate an alternative mathematical approach and network realization to accomplish the normalization.

by the selection of centers and widths of Gaussian radial basis function (RBF) units and weight values from each unit in the basis function unit to an output summing layer.

A more elaborate version of this approach is found in the Chorus system (Duvdevani-Bar and Edelman 1999); (Edelman 1999). In Chorus, a set of prototype objects is chosen as representative of a larger world of objects. The *n* prototypes are drawn from a smaller number of categories. RBF classifiers for each of these prototype objects are designed (trained). With advance knowledge of all the prototype objects, optimal views for each can be chosen which achieve the best normalization (constant activation of the classifiers for all views) as well as maximizing inter-cluster distances in the space of all prototypes. Any known or novel objects applied to the measurement units (200 tuned filters) map to a point in the *representation space* whose dimensions correspond to each prototype. This point is signified by activation values on each prototype unit.

Given this representation space, categorization can be performed by various strategies. Nearest neighbor match chooses the category of the object with minimum distance; another, k-nearest neighbors, examines the category of the k nearest neighbors and selects the category based on majority vote.

Critique of the feed-forward view interpolation theory

Supporters of the view based stragegy generally attribute recognition time effects to a normalization (i.e. orientation correction) process, similar to that assumed for mental rotation processes. However, Chorus, a well knnown computational view-based model with claims for biological relevance, does not actually predict any differences in reaction time for the normalization process. The one-shot feed-forward flow through the network is the same for any view presented to the network, whether novel or previously learned. A previously proposed network, with a spreading activation architecture, had a more natural interpretation for reaction time (Edelman and Weinshall 1991) . In general, some form of iterative computation and competitive interactions progressing toward a decision state have been invoked to explain reaction times in connectionist models, while more abstract theories such as the diffusion model (Ratcliff, Van Zandt et al. 1999) claim to explain response time distributions and differences in response distributions for error and correct responses.

Nonlinearities in response time vs. distance from familiar views have been noted by several investigators under certain testing conditions. Hayward and Tarr used a set of qualitatively distinct single part geons previously used by Biederman and Gerhardstein (Biederman and Gerhardstein 1993), but changed the experimental conditions to eliminate possible opportunities to learn multiple views and exploit local diagnostic features (Biederman and Gerhardstein 1993). They designed objects and training viewpoints such that for ± 45° rotations from the trained view, one direction resulted in no qualitative changes, while the other produced qualitative changes, such as the disappearance of areas of curvature. For these conditions, they found that response time varied between quantitative and qualitative conditions (610 vs. 650 ms) and error rates

also varied similarly (13.5 vs. 4.5 %). Hayward and Tarr conclude, then, that these results contradict a normalization process based simply on magnitude of rotation.

The source of the well-established reaction time effect is arguably an artifact of some matching process of stored codes and codes formed in the early visual pathways, possibly involving synchronization processes. Based on these nonlinearities, Edelman has argued that the mental rotation hypothesis has weak support based on the evidence mentioned above, and that the disappearance or reduction of delay with practice represents a faster path to recognition, not an increase in the rotation rate.

The subject of visual search and attention has to date been given relatively little attention in the literature on view-based object recognition, but problems with the feed-forward recognition model are also apparent in this context. It is easy for humans to search a visual scene for a familiar object and to know that it is not there, but feed-forward models do not readily address this case. Extensive experiments by Miyashita indicate a repeatable, stimulus-specific response in anterior ventral IT cortex during a 16 second delay interval in a delayed match to sample task (Miyashita and Chang 1988). The stimulus is not present during this interval, and the response is statistically distinguished from the period when the stimulus *is* present. This is interpreted by Miyashita as a neural correlate of short-term memory for the particular shape. A feed-forward model considers the network weights to be the essence of memory, and predicts no stimulus specific response during a delay period.

Single unit studies addressing attention and search aspects of object recognition in IT cortex have led to considerably different interpretations of the functioning of IT than those cited by Edelman and Tarr. These are described in more detail in a later section, but for now I note the findings of Eskandar et al. (Eskandar, Optican et al. 1992) that the best prediction of the stimulus from spike trains results from interpreting the trains as a multiplication of a target code and the incoming stimulus code during a search process. This could be interpreted as an intermediate computation (a weighting process) in an RBF-style computation leading to activation in a certain area. Alternatively, it might be interpreted as a cooperative synchronization process, also ultimately resulting in activation in a few areas which are structurally and dynamically suited to synchronize with the stored memory representation. The latter type of computation is the focus of the theory and experiments here.

A final issue I raise regarding the neural correlates of psychological phenomena was pointed out by Tsuda (Tsuda 1992), that of the difficulty of breaking the life of an organism into clean epochs of learning and recognition. It seems likely that normal exploratory behavior involves both of these activities proceeding in parallel, or at least that a system is poised to be able to rapidly switch from one to the other as the dominant mode. Dynamical models involving continual *bifurcation* (dynamical parameter changes), but more explicitly recast in terms of synchronization dynamics, may be a more appropriate architecture for combining learning and recognition in a natural way (Skarda and Freeman 1987). The approach in this thesis, while consistent with oscillatory representations and synchronization-based computational strategies, does not yet step up to the challenge of dynamic shifting between learning and recognition modes.

In summary, while view-based recognition has deservedly emerged as a leading computational theory of object recognition and representation, several issues have not been addressed: Primed search (search for an object held in short term memory in a visual scene), variance in reaction times, and the longer time scale contextual shifts between learning and recognition. While these issues are not addressed or resolved theoretically or experimentally by the present work, we will return to these subjects in discussing the relative merits of three recognition approaches with claims to biological relevance.

PERSPECTIVES FROM COMPUTER VISION

Most contemporary work in the psychology of vision and perception uses computational and signal processing concepts. The converse is not true, in that many algorithms proposed for recognition have no ready interpretation in neural network terms. In reviewing developments in computer vision most relevant to the present work, I will first focus on a few classical dilemmas related to object recognition. I then review some recent work claiming to be biologically motivated, and finally mention some recent algorithmic approaches which share aspects of the computational style. In spite of the emphasis here on dynamics and neuroscience, the synchronization opponent lattice network also has something in common with recent trends in computer vision including nonlinear diffusion, deformation, and feature histogram methods; thus it may be improved by drawing on continued progress in those areas.

One algorithm (geometric hashing) is presented which may seem a bit out of context with the rest of the discussion. I include it because it handles two problems – invariance for discontinuous point sets and embedding of objects in a scene – which I do not believe can be handled by any methods discussed here, including my own.

Classical Pattern Recognition in the Image Domain

Some attempt must be made to situate the present work in relation to the long and diverse history of image recognition methods. To concisely present the history and recent trends of such a vast field is challenging; I will emphasize the areas of transformation and multiple scales that characterize recent geometric methods, and will stress the way in which transformative methods can blur traditional distinctions between structural and syntactic approaches and scale issues.

Several surveys on image processing methods identify the major classical methods as either statistical or structural (Freeman 1985); (Leedham 1991); (Del Bimbo 1999). Del Bimbo describes more recent approaches as “shape through transformation”. Thus classical recognition methods – both statistical and structural (or *syntactic*) – are relatively passive, in that they do not modify the base image. They merely subject it to some interpretive framework, such as a particular feature set. In contrast, my method (and others I will survey) *modifies* the image in some way prior to measurement on a modified image, or possibly measurements over a sequence of modifications.

Classical Methods and Dilemmas

Statistical Methods

In statistical approaches, pattern data is represented by a feature vector which is used as input to some classifier or decision process. Features may characterize global form (area, elongatedness, major axis orientation) or local elements (corners, characteristic points). Shapes are viewed as points in shape feature space. For effective recognition, the requirement is to choose features such that patterns of the same class are tightly clustered in N dimensional space corresponding to N features, and patterns of different classes are in other tightly clustered regions well separated from each other (Duda and Hart 1973).

A key problem in statistical methods is the reduction of the dimensionality of the feature vector. This may be accomplished by a feature selection process, in which low significance features are deleted, or by a feature space transformation method, or both. Classically, a particular class was represented by a template with matching against templates; this matching was considered to be intractable for large numbers of objects due to the need to compare with inputs which have been rotated, scaled, partly occluded, non-rigidly transformed, or presented under varying lighting conditions. Recent schemes employing normalization (the RBF networks underlying Chorus) and interactions among multiple well chosen prototypes, or the sophisticated weighting of a large feature set (Mel 1997) have overcome this to some extent.

Another approach to the use of features is to create a transformed representation space on the basis of correlations among the dimensions to enhance cluster tightness and inter-class separation. Feed-forward supervised networks, or competitive networks such as self organizing maps can use feature vectors as input, and via training transform the features into activation levels in a set of network elements corresponding to classes.

Decision methods may generally be classed as non-parametric or parametric (Leedham 1991). Non-parametric methods include linear discriminant functions, minimum distance classifiers, and nearest neighbor classifiers.

The most widely used parametric decision rule is the Bayes classifier. The main distinction from non-parametric methods is that the decision rule involves class conditional densities and *a priori* probabilities of occurrence of classes. Bayesian classifiers are particularly important with large object databases, where setting classifier decision boundaries properly and defining the optimal feature set are crucial for good recognition performances.

The description of statistical pattern recognition methods presented here thus far has been in general terms, applicable to any data set. Recognition of object shapes in a statistical framework poses additional problems unique to this class of data. Non-rigid objects are composed of parts which can assume different poses – human and animal figures are good examples.

The changing projections of three dimensional objects seen from different viewpoints constitute the *stimulus identity problem.* Different features and feature conjunctions will be present in each view. This problem has been addressed by

geometric methods seeking invariants (treated in the next section), or by neural networks exploiting regularities in the changing distributions of raw features (the Chorus RBF ensemble approach). Recently, however, progress on stimulus equivalence within a "raw feature" paradigm has been demonstrated, by careful design.

Mel, describing the design goals for a recent high performance feature based system (Mel 1997), notes the following expectations on feature sets to overcome these problems:

1. Features should be large in number; sparsely occupied high dimensional representations are most robust to noise.
2. Features should be useful; they may be sensitive to object quality (occlusion, poor lighting) but should be robust in the face of pose or configuration changes.
3. They should be dominated by spatially local features; this is particularly important for non-rigid objects , which preserve local but not global structure in any particular view.
4. They should be driven by multiple visual cues to maximize discrimination, represent diverse objects, and buffer representation against degradation which affects different cues (feature channels) more or less severely.

The use of these principles led to the creation of his SEEMORE system, which achieves recognition rates above 90% in a 100 object world, even for scrambled images. The high performance achieved with these first order[9] feature channels is interpreted by Mel to support the idea that a simple feature space is all that is needed and attempts to extract structural information or otherwise "bind" collections of features may be unnecessary for biological systems. It is easy, however, to construct images with identical first order statistics which will fool such a system but are readily distinguished by humans. It seems likely that some of SEEMORE's recognition success depends on diversity in first order statistics of the object world, along with limited use of second order statistics for some feature channels.

Structural or Syntactic Methods

The other major family of classic pattern recognition approaches, chiefly developed for image or shape processing applications are structural or syntactic methods (Pavlidis 1977). Here, the input image must first be segmented into primitives; the primitives must be recognized, and spatial or topological relationships between these primitives extracted. Finally, with this information, a syntactic analysis and classification on that basis can proceed. None of these problems are trivial.

Within computer vision, structural methods based on raw image data have been largely superseded by related methods which capture structural information implicitly by *multi-scale* representations or by deformations. In the psychological examination of human vision, structural approaches still command a good deal of support. In part, this is

[9] First order features implies that no information on the spatial proximity of other features is present. Second order features would capture adjacencies of feature pairs at one or more scales, with increasing high order features preserving this trend.

due to the fact that task specific or language mediated descriptions of objects offer evidence that *compositional* representations are used. Statistical approaches, and feed-forward neural networks have been problematic in regard to this issue.

Compositionality is essentially the separability of the components of a composite representation, i.e. the ability to use or talk about them independently after the formation of that representation (Van Gelder 1990). Recurrent networks have been demonstrated to exhibit a so called functional compositionality, in which tree structures can be represented and their constituent parts derived (Pollack 1990).

Some Geometric Methods for Shape Description

Finding a representation of shapes which is invariant to viewpoint has been approached from a variety of methods which are difficult to justify biologically. However there is evidence that classes of stimuli, such as point sets in a regular geometrical arrangement, are recognized even in a noisy background (Uttal 1988). It is unlikely that other methods discussed here involving local receptive-field computations (e.g. SEEMORE) would handle this situation well. The first geometric method to be examined deals explicitly with point sets and is designed to work in scene analysis.

Geometric Hashing

Geometric hashing (Wolfson and Yehezkel 1992) was proposed as a means of performing model based recognition in scenes, with robustness to partial occlusion and to transformations in the plane. This is accomplished by considering an object as a point set, and by remapping coordinates of every point in terms of all possible triplets of non-collinear points. For four points $A,B,C,D \in \boldsymbol{R}^2$, affine invariant coordinates of D are coordinates with respect to axes defined by $\overline{AB}$ and $\overline{AC}$. Affine transformations (translation, scaling or rotation in the plane) will produce a new set of points A', B', C', D'; the coordinates of D' in the A', B', C' coordinate system are unchanged. First, signatures are generated from interest points on one or several views of an object. Interest points are endpoints or intersections of segments extracted by some edge extraction procedure. During scene analysis, interest points are selected and processed by a similar computation. The signature generation procedure is outlined here:

procedure signature_generation
  for each model object
    extract m interest points for the object
    for each ordered non-collinear triplet (affine basis) do
      a) compute coordinates of all m-3 model points in the affine coordinate frame for the current basis;
      b) use the coordinate as an address to a hash table;
      c) record in the table entry a pair {model, basis} for which the coordinate was obtained
    end for
  end for

end procedure

The complexity of signature generation is of order m^4 per model. The creation of the hash table is viewed as a learning process, in which a memory is formed relative to different foci of attention.

The corresponding matching procedure is then:

procedure match_model_in_scene

a) extract n interest points from scene

b) choose arbitrary ordered triplet of non-collinear points, compute scene points referenced to this triplet as affine basis.

c) for each such coordinate

check the appropriate entry in hash table;

for every {model, basis} pair , tally a vote for the model and affine basis.

endfor

d) If a certain {model, basis} pair scores many votes, decide this is the object.

e) Consider all {triplet, image point} pairs which voted for winning {model, basis} pair

f) find the affine transformation giving the best least-squares match between corresponding point pair views.

g) transform the whole low level representation of model according to affine transform and verify it vs. the scene.

end procedure

Multiresolution Methods

One major problem with feature vector classifiers is that the relevant features of an object tend to vary with scale in a way which is unknown *a priori*. Overcoming this defect is a major goal of *scale space* approaches, such as geometry-driven diffusion. While such methods can adaptively tune feature representations for shapes with detail at many scales, the mapping of the resultant curve family to a representative feature vector can be computationally expensive, and some of the features advocated are difficult to discover (i.e. the detection of singularities in evolved curves). Wavelet decompositions also perform well in terms of capturing details at multiple spatial scales, but early formulations had problems with translational and rotational invariance; newer methods, such as *steerable pyramids* (Simoncelli, Freeman et al. 1992) claim to overcome these limitations.

Transformational or Deformation Methods

Scale spaces are a generic term for families of derived images or shape which attempt to capture aspects of shapes at various spatial frequency bands. In *smoothing* approaches to scale space shape characterization, a gray-scale luminance image is subject to a smoothing evolution by a family of Gaussian kernels with increasing neighborhood size. Alternatively, in a more geometrical, abstract formulation, a curve may be evolved by displacement at each point by moving in the direction of the normal vector by an amount proportional to the curvature at that point. Each evolved shape in this iterative process can be characterized by some feature. Zero crossings of derivatives, inflection points, curvature extrema, and symmetry axes have been used as features (Kimia and Siddiqi 1994). The set of features extracted after the evolution process captures the shape characteristics at a variety of scales. Extrema that survive larger smoothing extents may be considered more significant, and might be weighted more heavily during feature based distance computations.

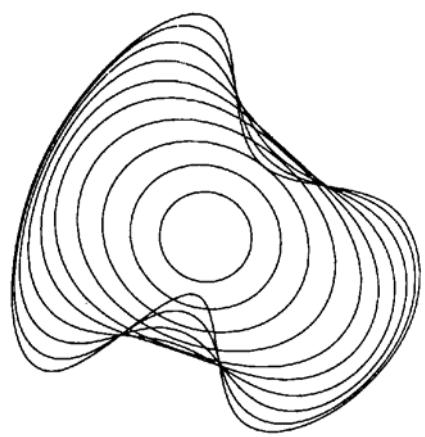

Fig. 2. Example of curve evolution by geometry -driven diffusion. Smoothing occurs by displacing each point from the original curve proportional to the local curvature. The series of curves generated serve as the basis for characterization of the original shape. From Kimia, B. B. and K. Siddiqi (1994). Geometric heat equation and nonlinear diffusion of shapes and images. Computer Vision and Pattern Recognition, Seattle, IEEE Computer Society. Used with permission, IEEE.

The literature on curve evolution is primarily concerned with theoretical problems and extensions and short on applied comparisons to other approaches. Curve evolution scale space methods are argued to give good qualitative descriptions of shape, but are rather expensive to compute and do not allow easy reconstruction in contrast to decomposition methods (e.g. wavelet transforms) which also capture information at various spatial scales. The listing of inflection points at each scale or iteration involves

difficulties in the data representation and efficient computation of a similarity function, since the number of inflection points for each shape is not constant. In addition, if the shapes to be characterized are not simple solids bounded by a closed curve, missing interior components would need to have separate evolutions and feature vectors, possibly leading to ambiguities, singularity problems or problems handling an extended set of vectors for each component when applying a similarity function.

Many of these defects have been reduced by a closely related, more recent approach utilizing *shocks* (Tek and Kimia 1999). Shocks are the sites where diffusion wavefronts from an object outline collide internally (on the medial axis of a shape), and externally as well. Resulting shock graphs and their grammars have been used to describe shapes, with similarity functions on the resulting graphs defined. The representation has been proven sufficient to reconstruct the local shape from the medial axis, tangents, velocity and acceleration of shocks (Giblin and Kimia 1999) . Medial axis representations are normally sensitive to deformations in the outline, but methods have been developed to distinguish stable from unstable shocks to ameliorate this problem (Giblin and Kimia 1999).

While the neural mechanisms which might implement such transformations are rather opaque, there is evidence that axis representations or skeletons are computed in some fashion and influence the response of cells in primary visual cortex (Kovacs and Julesz 1994); (Lee, D. et al. 1998).

Morphological Scale Space

A related approach relying on nonlinear transformation of the image at multiple scales is designated *morphological scale space* (Korn, Sidiropouls et al. 1996). In this approach, morphological operators of increasing scale are applied to the original image or curve, again resulting in a family of transformed images. A *pattern spectrum* has been proposed to characterize such an image family (Maragos 1988). The spectrum consists of the accumulation of successive differences in area between a pattern and its successor as opening and closing morphological operators of increasing scale are applied. Similarity functions can be applied to the resulting histogram.

Comparison of Computational Methods and Psychological Responses

The application of geometric algorithms for shape similarity to the problem of image retrieval in multimedia databases has motivated studies of how well a particular algorithm corresponds to human judgements on the same task. One such study with a large and diverse set of images (Scasseleti, Alexopoulos et al. 1994) found that each of several algorithms performed very well for certain target images, but poorly on others. Turning angle, the most robust of the algorithms across the images, was the best match to human preferences on only 8 of the 20 target images used. Turning angle methods require a search or fitting procedure to insure the best alignment between the feataure vectors prior to computing the distance function; also, such raw curvature descriptions are sensitive to scale. Descriptions based on sets of inflection points, like the sign of curvature approach, reduce scale sensitivity in comparison with raw curvature.

This finding of an apparent psychological primacy of local contour based measures corroborates the finding of Hayward already mentioned (Hayward 1998) that contours of binary images (silhouettes) are recognized with comparable performance to gray scale images, and that contour (or some transformation of contour) is the major information source for human object recognition and similarity.

Summary: Situating the Soca Approach in Computer Vision

The framework and implementation I develop here draws from many classical computer vision concepts, as well as the image transformation, and representation space concept introduced earlier in the context of the Chorus system. Statistics are used to match histogram templates; currently a nearest neighbor decision function is used. These statistics are over an abstract representation space which is designed to achieve the goals of within-class tightness and inter-class separation, where each class is a depth rotation-invariant description of a three dimensional object. The rotation invariance is formed by a transformation method involving diffusion and blurring as part of its mechanism, like the heat equation deformation methods. These scale space methods typically avoid *creating new spatial structure*; in contrast, the procedure described here is completely dependent on creating fine structure, and on cooperative interactions derived from those structures.

A fundamental aspect of the Soca network implementation is that there is local, receptive field like processing, with a diffusive "spreading of activity" character. However, this activation is not to be understood as a monotonic variable associated with detection of some feature. In image processing terms, the process can be considered as a nonlinear filter with feedback, or iterative nonlinear convolution. This combination of diffusion and highly nonlinear (non-monotonic) transfer function forms a representation determined by both local features (e.g. curvature and corner elements) and medium-scale structural relationships. The scale of interactions is determined by a window proportional to the number of network iterations used to generate a representation meeting some criteria. A particular juxtaposition of local curvature changes may, with appropriate network parameters, result in a unique distribution or histogram in the representation space. This type of process is, to my knowledge, a unique approach to combining local feature and structural information; thus it represents one of the main contributions of the thesis.

Forming such a representation - one that captures the co-occurrence of local features - is a hotly debated subject in neuroscience, referred to as the *binding problem*. While the problem is typically presented in terms of separate channels (such as color and shape), the situation of decomposing an image or outline into a set of orientation frequency detectors presents the same difficulty. Opinion on the subject ranges from claims on the neural correlates of binding to assertions that there *is* no problem. This will be discussed in some detail in a subsequent section on neuroscience.

PERSPECTIVES FROM THEORETICAL COMPUTER SCIENCE

Computer science has certain perspectives and emphases that result in a characteristic way of framing the problems of similarity and recognition. In this section, I focus on two such perspectives. First, any spatial or temporal pattern can ultimately be represented as a string in some alphabet. One traditional theoretical approach is to consider families of such strings as a formal language, and to frame recognition problems in terms of recognizing a language. This way of formulating recognition problems is also relevant to the present thesis because it allows the tools and language of symbolic dynamics to be applied. Symbolic dynamics will be taken up in more detail in the section on dynamics and representation, and in theoretical discussions on proving the representation-forming capability of Soca style transformations.

Dynamical Recognizers and Computational Mechanics

The problem of learning to accept positive exemplars of a language while rejecting negative exemplars is known as language induction. Classical machine learning approaches to this problem construct a finite state automaton to affect recognition. Formally, a finite state recognizer is a quadruple $\{Q, \Sigma, \delta, F\}$, where Q is a set of states (with q_0 denoting the initial state), Σ is some finite alphabet, δ is a transition function mapping $Q \times \Sigma \Rightarrow Q$, and $F \subset Q$ is a set of final or accepting states. A string of tokens from alphabet Σ is *accepted* by the recognizers if, starting from initial state q_0 the sequence of state transitions indicated by the tokens in the string ends up in one of the final states in subset F.

A pioneering attempt to formulate the language induction problem as a dynamical system, in the form of a recurrent neural network, was the study of Pollack (Pollack 1991). The dynamical recognizer is a quadruple $\{Z,\Sigma,\Omega,G\}$, where $Z \subset R^k$ is a state space; $z_k(0)$ is the initial condition. Σ is the input "alphabet", where a particular closed interval in Z corresponds to each element in this alphabet. (This correspondence between *intervals of state-space* and *symbols* is a cornerstone of *symbolic dynamics*, which will be mentioned again later). Ω is the dynamic, a sequence of transformations $\omega_i{:}Z \rightarrow Z$ (one for each token) with an associated set of dynamical parameters; these parameters are fixed for a particular recognizer during the induction (training) process. $G(Z) \rightarrow \{0,1\}$ is the decision function which maps one or more states in the sequence produced by the dynamic to an accept/ reject decision. In Pollack's work, only the final state and token are used in the decision function. Within this general framework, the dynamics and decision function are normally much weaker in computational power than a Turning machine. Pollack notes that G may be generalized to a graded function indicating "fuzzy" acceptance, or could return a more complex categorization or representation.

The Soca network and recognition method I describe later is quite consistent with this extended dynamical recognizer framework. A key difference is that the Soca net operates on an image "string" in parallel (thus the state space has higher dimensionality R^N, where N is the number of pixels or sampled image elements), and the tokens are used only once as the initial state. Such a parallel recognizer framework for *picture languages*

with constrained states and transformations was studied in a series of papers by Rosenfeld (Rosenfeld 1979). In Rosenfeld's formulation, the transition function at each pixel is now a function of several tokens in some spatial neighborhood; this is the cellular automata formalism, which is described in the dynamics section later. The decision function is necessarily modified by this larger state space. Rosenfeld proposed several possibilities:

1. every spatial element reaches an accepting state
2. any element reaches an accepting state
3. one particular spatial element reaches an accepting state.

In the Soca network and recognition strategy, my approach is to form a metric representation space, but that space consists of *statistics measured instantaneously during a high dimensional, parallel dynamics*, rather than a direct map of the input features or measurement space. These statistics naturally support an acceptance function; simply define some threshold distance for each classifier, and accept an object as an instance of language L if it satisfies this distance test. The distances might vary by class, depending on the cluster density of that class in the representation space. Another contribution of the thesis, then, is adding another type of decision function to the repertoire defined by Rosenfeld. While such a distance threshold decision function is common in statistical pattern recognition, it is novel for processes operating with local dynamics.

Other researchers have recently been concerned with decision functions over spatial patterns processed by cellular automata, a form of spatially-extended dynamical systems closely related to those used in the present work (Mitchell, Hraber et al. 1993); (Mitchell, Crutchfield et al. 1996); (Hordijk, Crutchfield et al. 1998). Genetic algorithms were used to generate and test particular one dimensional cellular automata (CA) which decide, for example, whether a random initial condition has majority ones or zeros. The group then examines space-time plots (i.e. plots of successive iterations of a 1-D spatial array) of the resulting successful computations and develops an explanatory framework based on physical metaphors; this framework is designated by this group as *Computational Mechanics*.

Computational Mechanics seeks to reconstruct the computations embedded in space-time behavior in terms of regular domains, particles, and particle interactions. *Regular domains* are regions visible in space-time plots consisting of words (spatial configurations) in the same regular language, i.e. regions that are computationally homogeneous. *Particles* are localized boundaries between such domains; they serve as information carriers. *Collisions* between particles are the loci of information processing. This processing can be conceived in terms of operators such as decay of one particle to many, reactions (state transitions between language domains at collision sites), and annihilations (the disappearance of an interface as one language domain dominates future spatial evolution at a collision site). The computational strategy can then be expressed in the more concise language of particles and their interactions, substituting for a more verbose description in the language of CA rule lookup tables and raw spatial configurations.

While the computational mechanics group does not explicitly state this, the decision function in the majority task can be considered a type of *synchronization* – until

all cells reach the same language domain (which, in this simple case, is all 0 or 1) the system is undecided. Consider a k-block as an adjacent set of k cells, with the entire CA consisting of overlapping sets of such k-blocks. Each k-block of cells in this automaton is defined by the state transition graph of a finite state automaton (FSA) with k states. As the automata evolves, at each time step we can label each state with the fraction of k-blocks which currently hold that value (we say they occupy the state). Synchronization in this context implies that over time, the occupancy statistics of the graph converge to sharp peaks, or an unchanging sequence of sharp peaks; particular sub-graphs of the state-transition graph are active, while others become blocked, as their predecessor states become unreachable within increasing spatial "territories".

Note that particles have a characteristic velocity, and for certain kinds of terminating conditions (such as a particular site or region reaching a value in a set of accepting states *F*) one possibility for variance in the temporal processing is dependence on the emergent particle velocities on initial configurations in a family of inputs, when a "synchronization" decision function is reached.

Summary: Situating the Soca Approach in Computer Science

In summary, the Soca system extends the tradition of dynamical language recognizers over spatial configurations, and attempts to unify this approach with traditional metric representation space of statistical pattern recognition. Similar to the work of Mitchell, Crutchfield, and their colleagues, the approach taken here is to discover successful computations within a particular family of spatially distributed computations, then analyze the result. I have generally proceeded with more constraints on the search process, guided by general principles of pattern recognition.

The decision functions used here also involve synchronization in the sense defined above, but the synchronization is partial and not defined to contiguous regions as in the regular domains. Another key distinction of the Soca work from the research in the computational mechanics group is that the present search strategy focuses on solving the decision problem within a fixed number of iterations, rather than an open ended synchronization process. This led to the hypothesis that dynamical changes (non-stationary parameters or rules) might lead to superior performance relative to constant dynamics, by forcing more rapid synchronization.