Shape Recognition, the magnitude of the challenge a machine learning approach

Julian Yarkony*
Department of Computer Science
University of California, Irvine
Irvine, CA 92617
jyarkony@uci.edu

Abstract

Shape recognition is a holy grail of computer vision and a great wonder in the field of cognitive science. However surprisingly little is known about how this is done. In this survey the shear magnitude of the challenge will be explored in great detail with emphasis on how large search spaces in space time, the need for generalization make the task difficult.

1 Introduction and overview:

Shape provides the user of a visual interface with more knowledge about a given object that any other feature (color,lighting, time, texture etc). The human brain uses shape to great effect and is incredibly adept at determining equivalence and similarity between shapes which it has observed in the past and current shapes in its perceptual view. The human brain is also able to infer shape from partial and incomplete observations. Furthermore it has the ability to extract features of shapes and use them to infer similarity to new and unobserved shapes, and is able to determine which aspects of shape are important and which are not. Not only do humans posses these abilities but a huge range of animals from advanced mammals such as dogs dolphins, and humans to very simple animals such as insects (which often have very adept visual systems). Palmer divides the ability to process shapes into two separate parts, shape equivalence and shape similarity. Shape equivalence involves the process of determining weather two shapes presented in two different contexts are the same, regardless of transformations, occlusions, and context alterations, Shape similarity is the process of determining weather shapes are similar (purposfully vague). Shape equivalence can be thought of as a discriminative model asking weather two things are the same while shape similarity can be thought of as a generative model which attempts to describe the commonalities of a family of shapes. Similarity is used to infer relationships between shapes and hence the properties of the objects that they represent. After this introductory discussion here this survey will take a machine learning approach and discuss the aspects of shape recognition which make it such a computationally complex problem. Next a discussion of the use of features and feature extraction in determining shape equivalence and shape similarity in the context of discriminative models and generative models. Afterwards a discussion of comparisons of templates vs descriptions in terms of shape recognition and similarity and their various insights. Then context will be discussed and how this greatly

*PhD student in Dr. Donald Hoffmans lab

Figure 1: A motivating example of a computer recognizing the parts of a person. This is from work done by Deva Ramanan.

complicates the problem of shape recognition. Finally concluding remarks will be given which discuss how spectacular the brains abilities are relative to current computer vision technology.

2 Invariance and variance of shapes

Shapes are rarely presented in an upright form at the center of the object view and scaled to the same size as they were when and if it was presented previously. More often the perceptual system must be able to search out for shapes in massive regions where the shapes have undergone a transformation which makes it difficult to find. Such transformations include but are not limited to translation, rotation, dilation, and reflection. The presence of translation means that the shape can be anywhere in the image, hence all possible locations must be checked. The positions or windows where the shape can occur are overlapping and hence a check for the shape must be done hundreds of times for even a small image unless elaborate mechanisms are used. Rotation means that even if the image is found it will not be easily recognizable. This is because the shape will look different from with different rotations hence adding 360 degrees of uncertainty. More insidious is dilation, this means that one can be searching for the image shape and be right on top of it and yet be unable to discern it from its surroundings as it is too small to recognize. If the image is too big on the other hand one could be giving a false negative in every position where the image is. Hence different window sizes or scales of analysis must be chosen multiplying the amount of computation needed in image search. Reflection adds extra difficulty as any recognition problem must be phrased as to avoid having to deal with numerous local axis that an object can have. Reflection in real images, objects and shapes is not restricted to rotation about the x,y,z but can include all sorts of elaborate types of reflections of an object (see chirality picture below). In chemistry this is called chirality as individual parts of molecules can rotate and the same chemical equation is preserved see figure 1. All of these difficulties exist in a context of a world without noise. When these transformations combine they will produce different objects if done in different orders as matrixes are not associative. This single feature adds a great deal to the problem of shape recognition as it means that procedural analysis is diminished as an analytical tool. Image analysis is

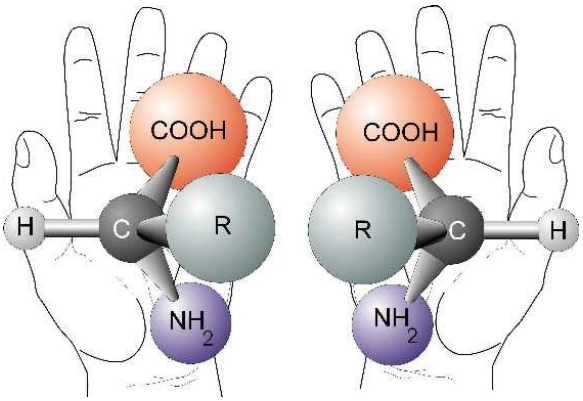

Figure 2: No matter how one of the shapes is rotated they will not become aligned. Chemical formula and shape are preserved under rotation and translation but rotation will not put the shapes into exact alignment. Imagine a fare more complex object with many parts which can be on opposite sides or face different directions but still as a group have the same fundamental shape. Recognizing similarity or equivalence in such situations can be quite difficult and adds great richness to the problem of shape recognition. Unlike in a reflection across the z axis in which all parts of the object switch sides real life objects need not have all their members switch sides making analysis under reflection extremely difficult

heavily complicated by these difficulties however the human brain is able to effortlessly analyze millions of large images from the eye each day. This is a great power of the human visual system which computers are far from being able to compete with.

3 Worse than variance the problem of generalization

A more challenging problem (though one which has been studied very intensively in machine learning literature,is one of generalization of shapes and other types of data). Generalization is the ability to determine what characteristics of members of the group are salient, which are unimportant and be able to determine weather an unseen instance is a member of that group. Sadly it is not true that there are a few traits each of which must be satisfied but instead there are many traits some of which must be satisfied, some never satisfied, some satisfied only when another is satisfied, or 3/5 (not any other number) of a set of traits must be satisfied. Learning more complex models requires more data or stronger and more informative priors. Amazingly the brain can do this with great ease and accuracy for example the brain will recognize chairs of all different variety, appearance, complexity and size as chairs because it is able to determine the importance, unimportance and relations between the traits of all members of the set of chairs while on the other hand being able to identify objects that may look more like the basic chair than some elaborate chairs as members of their own correct classes (tables, couches, etc).
Using machine learning one can put the problem of shape generalization in the context of generative models. Generative models attempt to understand the relations of the variables (Chairs: how many legs, what angles are they relative to each-other, is there a long flat surface etc) within instances of a type of objects. Object generalization can be thought of as the following problem: Given a set of input vectors or objects generate and recognize members of that set (datum have high that have high likelihood's of being generated to put it in the context of generative models such as gaussian mixture models and restricted boltzman machines). This can be done in the context of features, both binary and continuous and need not require the inputting of maps of pixels describing an image. Recognition is

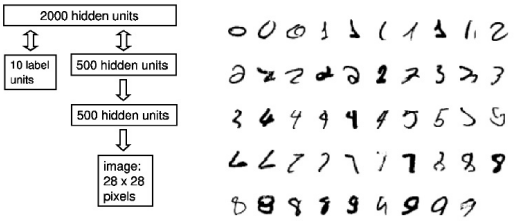

Figure 3: On the left is a restricted boltzman machine, a generative model capable of recognizing and modeling digits. One the right are different handwritten digits from the MNIST data set of which is modeled by the restricted boltzman machine. The restricted boltzman machine is capable of recognizing an generating these characters despite no direct knowledge or preprocessing using linear transformations to make the input data more uniform. This picture comes from the restricted boltzman machine draft from Geoffrey Hinton (one of my scientific heros)

done in the context of determining how likely the chair gaussian or the chair hidden layer is to generate this a given datum. Interestingly can be done quickly and does not require the generation and analysis of hundreds of millions of datum nor does the training of the generative model require enormous quantities of data though more data can be helpful. Generalization adds to the problems created by object variance as by combining the two the overall problem for the perceptual system is to find objects in an image given that you do not know what is there, what size it is, where it is, or even the important parts which make it up. However the perceptual system in both humans and animals is able to tackle these problems with elegance, and ease computers on the other hand are far behind.

4 Invariant features hypothesis

The horrors of generalization and variance can be tackled by a powerful (yet woefully insufficient) technique and theory called the invariant features hypothesis. This states that objects are recognized using features which are present in the object under rotation, dilation, translation, and reflection. Such features can include, number of angles of a certain degree, percent of object which is a certain color (the parameters of the color histogram), number of lines with certain length ratios in close contact. Once large numbers of features are extracted from a particular object they can be turned into binary (if desired, though this is not needed) and then used by a classifier to determine weather or not an object is in a particular class. The theory of invariant features was heavily pushed by McCloch and Pitts two of the founders and AI and neural networks and until recently was long dominant in cognitive science circles. Classifiers or discriminative models are a large set of tools in machine learning which ”learn” to distinguish between members of different classes given a set of training data. Two important questions come up in this theory and heavily undermine it. The first is the question of the brains ability to distinguish between shapes which are identical in all but rotation such as mach’s square diamond (see below) and the question of the degree of information held in data which is so reduced through the process of ignoring so much information as is the result of only looking at invariant features.
The invariant features hypothesis would imply that the human perceptual system would recognize mach’s square/diamond as the same object as they differ only by a rotation and the invariant features are unaffected by that rotation however it recognizes 2 the mach’s square and diamond as separate objects depending on rotation. This implies that the perceptual system is not using invariant features but is instead drawn to different features depending on presentation. This does not mean that invariant features can not be used in computer vision

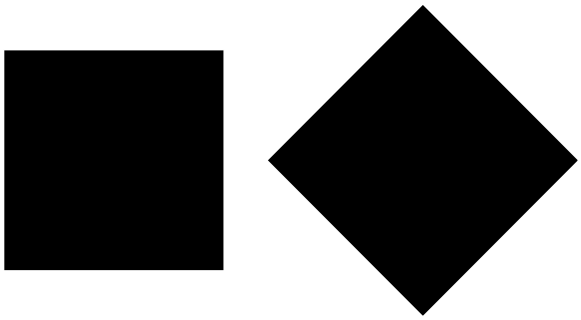

Figure 4: These two shapes are identicle in all ways except for a 45 degree rotation. The invarient features hypothisis would state that these two shapes would be recognized identically by the human perceptua system however they are not which significantly damages the credability of this theory.

techniques but if the most powerful perceptual system (the brain) is not using this method then it implies that there is a better method that needs to be discovered for our computer vision systems to advance beyond a certain point. Another problem with invariant features is that it requires that all the variant features be thrown out before analysis. An enormous amount of information may be contained in these features and that may vastly weaken the ability of a computer perceptual system to learn models and do classification. Despite its weaknesses the invariant features hypothesis and the techniques it provide CAN be used in computer vision.

5 Templates, their importance and their weeknesses

Templates are a naive solution based on a convolution operation designed to detect particular objects. They are mocked in perception and vision literature as biologically infeasible (though palmer points out that they must be used in some point in the vision process). Templates involve placing a template atop of the pixels on an image. Next a weighted sum is taken. If the output is above a certain threshold the image is considered a member of the set described by the template. However templates can not be rotated as they would not align with pixels under rotation, they are not able to handel insertions in the input and can become out of phase quite easily, they can not deal with dilations as srinking the template would require multiple convolutions and there is not guarentee that the alignment would not result in a requirment of hal pixels. Templates from a machine learning point of view represent the need to have multiple layer of analysis before classification as a direct approach looking at the data set is insufficient to properly represent of the complexities of the data. It implies that features should be extracted and multiple layers of representation holding descriptions or the presence or absence of global and local features must be used in complex visual analysis

6 Comparing images vs Comparing Descriptions;Feature lists

Rather than comparing images as is done in templates in order to determine weather two objects are the same or similar it has been suggested that shapes are compared by comparing descriptions . A shape can be broken down into a series of labels, numbers, class memberships etc. One example of this is a feature list which are a means of describing an object in terms of a binary or numeric representation. Features can be simple invariant features, or complex difficult to derive features. A figure can be inputted into a feature

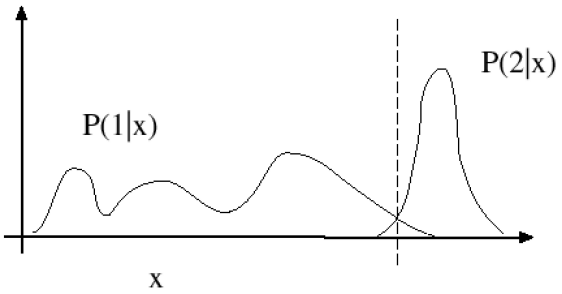

Figure 5: This demonsrates at its funemental level what a classifier is (a template is a type of classifier). The vertical line in the middle represents a proposed value for the optimal boundary between the data classes of the given input space. The line (or lines) of optimal (depends on receiver operating characteristic) classification is often difficult to determine and the space of inputs is often highly fragmented and noncountius in regards to the class (almost like a map of early modern central europe with tiny states and states with non contiguous or swiss cheese territories). This graphic comes from Forsythe's slides on computer vision

Figure 6: This map demonstrates the problem of classification given highly non-linear state spaces. Imagine the problem of class segmentation in a state space such as this. Note that this segment of map would look like a spec on a map of modern india. This comes from google image search for maps of early modern metre (central) europe during the days of the holy roman empire (which was neither holy, roman or an empire)

extraction system and relations, local color histograms, the presence or absence of local groups of angles can be observed. Features can be highly elaborate and can require deep analysis or be very simple. Once this is done lists can be compared using a dot product (or a more complex method) to evaluate similarity. More elaborate forms of classification are desirable as a dot product is unable to take account the relative importance, unimportance or correlation's between different members of the feature list. To clarify this point imagine multiple instances of the same feature (this is quite possible as many features may be so closely correlated as to be essentially the same feature); this will heavily bias the result of the dot product and hence cause inferior results on similarity detection tasks. Comparing objects before feature extraction is comparable to direct training of discriminative models without first generalization and representing them; a procedure which Geoffrey Hinton asserts is heavily flawed as each layer of representation attempts to do to much. Feature lists in this way are a representative model which can but put data in a generative context allowing for more advanced and accurate classification. Two problems are inherit in feature lists. First the features have to be determined by some means and this implies that they must be hard coded. This means that a person must determine features relevant for each object a highly time intensive procedure or worse a general set which would be so large as to make its use difficult. The second is given the feature list, is the determination of weather a feature is present. This often means having to deal with all the problems of translation, rotation, dilation, and reflection. This is in some ways turning the problem of object and shape recognition into dozens if not hundreds of instances of the same problem and is highly inefficient.

7 Getting a good view of an object reference frames and their associated heuristics

The complexity of analysis can be vastly reduced given that one can put the object being analyzed inside of a reference frame or a specific context. This is difficult as to put a reference frame one has to know what it is first which is the whole point . This process can be simplified by knowing about its orientation (deliberately vague) or its spatial/physical structure. For example trees are tall and are oriented along their vertical, and missiles are long and thin and oriented in their direction of travel. Some structural heuristics which helps inform insights into orientation are: gravitational orientation (which orientation gravity would force upon the shape), axis of symmetry, axis of elongation, contour orientation (see picture below), textural orientation (does the texture imply something about the orientation of the object), contextual orientation (what is the orientation of the objects around it)(see picture below), and motion orientation (what structure does motion place on orientation). This paragraph may sound vague, thoughtless, or poorly described, however these ideas have not fully been formalized into a unified theory. They do however provide insights into future approaches to the problem of imposing structure on unknown objects. Perhaps using EM to itteravtivly gain insight into the parameters of an object using these heuristics may be useful (perhaps a majumder, julian paper).

8 Context is key, the horror of spatial and temporal dependence

One deeply disturbing idea which is highly highly consistent with analysis of human perception is the idea that accurate recognition of the shape of an object is highly dependent on the shape of the objects around it, the motion of the object, the context of the object etc. This is not true simply for visual illusions but is all to often present in everyday visual analysis. This means that an object can not simply be analyzed by looking at its own features but instead requires a degree of computational complexity and analysis which is extremely great. One example which demonstrates the phenomena of time dependence on the per-

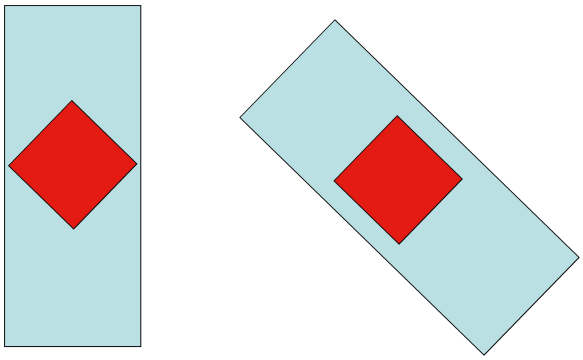

Figure 7: This demonstrates how contextual dependence influences perception. The tilted square inside the tiled box is generally perceived as a square while the other is perceived as a diamond.

ception of shape is the famous walking group of dots, if any particular frame is shown then the group of dots appears as nothing, however when they are shown with time they appear to be the outline of a person. The true horror comes from the requirement that spatial and temporal dependency imposes which is that in order to understand an objects shape, orientation and behavior one must often look at the objects around it yielding a highly iterative procedures for maximizing the degree of consistency between object relationships in the image. Approaches to to similar problems include dealing with local windows of activity and these are used to great affect in machine learning (Pierre Baldi applied similar techniques to second structure prediction in proteins). The dependency of objects upon each-other turns inference of shape and other properties from a linear time operation based on the number of shapes needed to be classified into a quadratic time or even worse. However objects and methods do exist for handling such dependencies, hidden markov models, markov random fields and LR-NN (a new method being developed by Julian Yarkony) are designed to handle such problems. None of these are even comparable in robustness and accuracy with even simple animal brains.

9 Conclusion, The brain is a true marvel

In the above sections the challenges which a cognitive system faces when processing shape are detailed,and the ability of the brain to handle the these challenges relative to current machine learning technology is been discussed. The brain is able to identify, classify, and infer properties of large groups of objects in infinitesimal periods of time. It is able to build accurate generalizations of large groups of visual data and infer properties of previously unseen objects. Modern machine learning techniques may be limited according to Dr Hoffman by the classical computing approach of the computer because he and many others believe the brain is a quantum computer capable of computations which are not simply faster but completely alien to modern classical computers. Quantum computers take advantage of entanglement and superposition of bits (known as cubits) in order to do exponential time algorithms in linear time and find global (not a typo) optima of complex functions in linear or even sub-linear time. If the brain is a quantum computer that would explain much about our inability to match its algorithms as human made computers are limited to a less efficient set of algorithms. The brains ability to handle variation, and generalization in real time sets a high bar for any machine learning techniques in the future weather the brain is a quantum or classical computer or the computers of the future are classical or quantum.

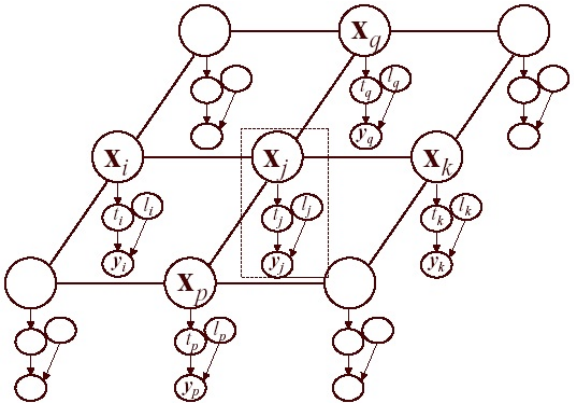

Figure 8: This is a markov random field a machine learning model taking into account long range dependencies via propagation in chains. Inference and parameter estimation in such models can be challenging do to the advanced and complex structures that they attempt to describe. However this is a functional real life example of a mathematical model which when considering a local classification takes in long range effects. This comes from the university of toronto Probabilistic and Statistical Inference group website

10 Bibliography

1. Vision Science, the shape recognition chapter (chapter 8)
2. Jim Davis's slides on computer vision
3. Bioinformatics by Pierre Baldi and Soren Brunak
4. All work on the interface theory of perception being developed by the Hoffman group of which I am a new member.
5. Computer vision a modern approach by Forsyth