A Shift in Vision: Recognition, Context and Depiction

by Ron Gallagher

Abstract: Recent work on the human visual system and on the development of computer based visual recognition systems indicates that shape-matching is an unworkable basis for human or machine recognition of objects, scenes and pictures. Studies of 'gist views' have shown that scene recognition is prior to object recognition. Therefore context, broadly construed, has a primary role in object recognition. This paper argues that we recognise objects in pictures using the same mechanisms that we use to recognise objects in the real world. It therefore follows that context, as opposed to shape matching, is the key to understanding how we see content in depictions.

Vision researchers are finding it difficult to shake off a theoretical model of the human visual system which characterizes it as a shape matching system. The predominance of the shape-matching model has focussed on the problems which the human visual system must overcome in interpreting the two-dimensional geometry of light stimulus on the retina as three-dimensional objects. For example, V.S. Ramachandran begins his 1988 paper 'Perceiving Shape from Shading' thus:

> Our visual experience of the world is based on two-dimensional images: flat patterns of varying light intensity and color falling on a single plane of cells in the retina. Yet we come to perceive solidity and depth. We can do this because a number of cues about depth are available in the retinal image: shading, perspective, occlusion of one object by another and stereoscopic disparity. In some mysterious way the brain is able to exploit these cues to recover the three-dimensional shapes of objects (Ramachandran, 1988, pp. 76-83).

This unfortunate emphasis on the two-dimensionality of the retinal image has led vision researchers and philosophers to speculate that the mechanisms whereby the human visual system recognises objects in pictures echoes the reconstruction of three-dimensional objects from the two-dimensional image on the retina.

The importance of the two-dimensionality of the image cast on the retina is also emphasized in David Marr's computational approach to vision and in some of the empiricist work on vision. The assumption which Marr makes is that the human visual system and the computer visual system are dealing with the same high-level computational problem – how to determine 'what' and 'where' from two-dimensional data. Variations on this view dominate vision research literature. One of the consequences of this emphasis on the problem of recovering three-dimensional data from the surface of

the retina has been the assumption that recognising depicted content in pictures presents roughly the same problem for the visual system. That is, in order to recognise content in a picture, the visual brain uses the two-dimensional shapes as the basis for reconstructing a three-dimensional object. The obsession, both in vision science and the philosophy of art, with developing an account of how we can see a three-dimensional object in a two-dimensional plane has blinded theorists to the fact that the two-dimensional outline which an object presents to the eye is largely irrelevant to recognition. Vision research indicates that human beings can categorise a scene as outdoor or indoor in less than 50 milliseconds and categorise an object in less than 150 milliseconds (Oliva, 2004). That is, we recognise scenes and objects at a speed which precludes the possibility that the initial recognition of a scene involves building up objects and scenes from their constituent shapes. Initial scene recognition is almost a reflex and does not, and could not involve higher brain functions such as memory. Some aspects of object recognition are also based on triggered reflexes.

My purpose, in this paper, is to show that recognition is not *primarily* based on shape but is driven by context. The human visual system does not evaluate the precise shape that an object presents to the eye in the initial act of recognition. Evaluating shapes is something that comes after recognition. I will demonstrate this in two ways. First, I will demonstrate the importance of context in that initial recognition moment. Second, I will show that shape matching is unworkable as a means of recognition for any visual system. I will then go on to evaluate the significance of the recent shift of emphasis in vision research towards contextual evaluation and scene recognition for theories of depiction.

Objects in Context: the human visual system primarily recognises objects using context

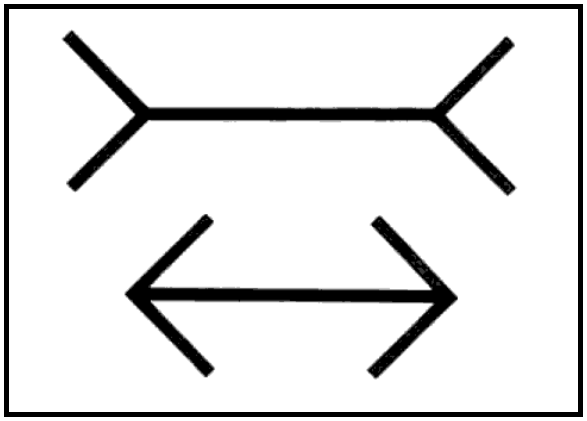

Figure 1 Muller-Lyer

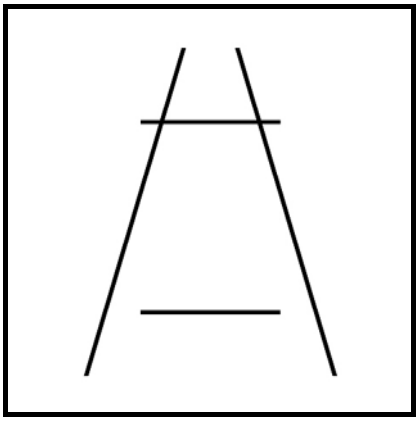

Figure 2 Ponzo

The Muller-Lyer and Ponzo illusions (figs. 1 & 2) demonstrate that what we perceive is radically affected by context. In each drawing the horizontal lines are of equal length, but look as if they are different lengths The apparent length of the horizontals in these illusions is altered by the lines and shapes of their context.

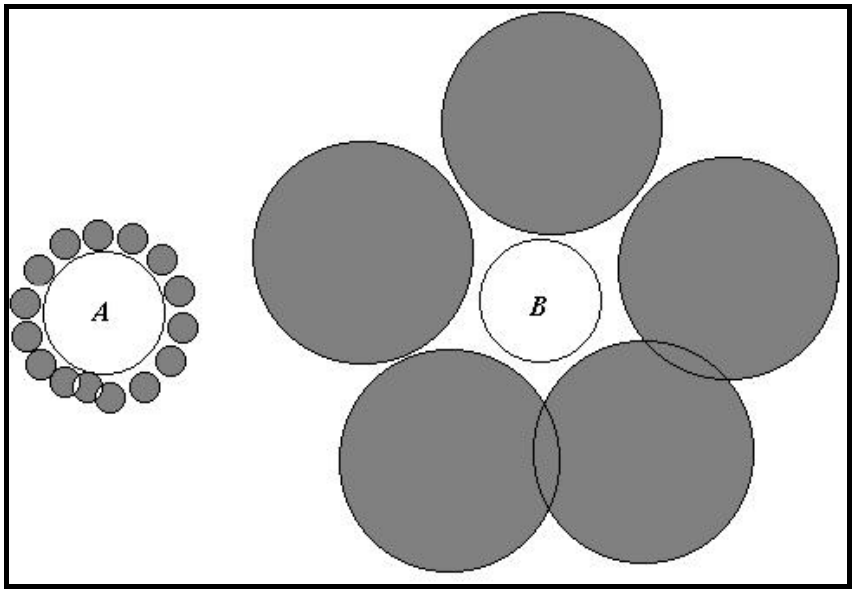

Figure 3 Ebbinghaus – circles A and B are the same size

In the Ebbinghaus illusion (fig. 3) and its variants (Howe and Purves, 2004) we can see that our perception of the size of the central circle in each group is altered by the surrounding shapes.

Figure 4 The Simultaneous Contrast Illusion – the round dots are all exactly the same shade of grey (Kaiser, 2007)

The simultaneous contrast illusion (fig. 4) shows how the brightness of a shape is affected by its surrounds.

These illusions have been devised by vision researchers to highlight various aspects of how the human visual system works. This family of illusions have been used to demonstrate that many primates use context as a primary indicator of the identity, real-world shape and size of objects both in pictures and in reality (Sigala and Logothetis, 2002).

However, it is misleading to call these visual effects illusions. While it is true that the horizontals on the page are the same length in both the Muller-Lyer and Ponzo drawings,

it is not the case that perceiving them as different lengths is a failing of our visual system. In fact, our interpretation of these drawings indicates that our visual system is trying to do exactly what it is supposed to do – perceive objects in the world based on their contextual setting. The reason simple line drawings can powerfully depict objects is because our visual system always tries to interpret the stimuli in the visual field as three dimensional objects in the world. When we look at the drawings above, our visual system is trying to determine what kind of real-world objects could have given rise to the light stimulus, not measure the light stimulus. Your visual brain is not reading the lines as marks on a two-dimensional canvas. Our visual system reads the lines as a co-existing complex of real-world edges and hypothesizes about what in the world could have been the source of the stimulus. Context, in the broadest sense of the word, is paramount in generating these visual-object hypotheses. In the Ebbinghaus drawing, although the central circles are identical, the surrounding circles make them look different.

Figure 5 Ducks

However, it is not surprising that we see the central circle on the right as smaller than the one on the left. Like the duck above (fig. 5), the central circle in the Ebbinghaus is surrounded by smaller versions of itself. The mother duck would look small if it was surrounded by geese. It is natural that our visual system should ascertain the size of objects *relative* to their surroundings because the apparent size of an object varies according to how far away it is. Thus we rely on context to ascertain size; our visual system does not calculate the size of an object based on how big or small it is in our visual field. In the Ponzo drawing (fig. 2) our visual system does not see a configuration of marks but sees a real-world scene involving far and close objects. The top horizontal seems further away than the bottom horizontal because of the converging tracks. Our visual system is forcing us to see the shapes as if they exist in the world and we cannot help but assume that the top line is further away than the bottom line and therefore represents a longer object. The visual system calculates the width, height and size of objects using a comparative system. It does not use the absolute size of an image on the retina as a guide to the size of the object. The reason the human visual system uses a number of different strategies to identify an object is that the light stimulus on the retina

is very ambiguous. It is almost always true that the stimulus on the retina could have been caused by a number of different source phenomena. Consequently the visual system uses context and heuristics to narrow down the possibilities. Of course, it doesn't seem as if you are guessing every time you look at an object – it seems that you rarely are mistaken about what you see. This conviction gives rise to the widespread assumption that when vision researchers talk about the light stimulus being ambiguous they are referring to peripheral cases. When you look at a live dog, for example, you don't assume the light source is ambiguous. After all, the light is reflecting off the dog and you see the dog very clearly. Where is the ambiguity? Why would our visual system need to evaluate the dog in terms of the dog's context when it can evaluate the furry shape in front of us? It is hard to believe that our visual system doesn't simply see the shape of the dog and match it up with similar dog shaped things it has encountered in the past. That certainly seems to be a more logical way of evaluating what is in front of us. It is the way that most artificial intelligence visual recognition systems have been designed – to a large extent they are shape matching systems and that is why they don't work.[1] The reason that artificial recognition systems are so unsuccessful is because shape matching is a very poor basis for visual recognition. There are three main reasons why this is the case:-

1. Objects look different from different angles and in different light;
2. Many objects change shape eg animals, clothes, plants;
3. Objects rarely present themselves whole and unobscured.

These factors alone are enough to defeat any visual system which relies on shape matching.

[1] In fact, machine-vision systems use a variety of non-human strategies to overcome the limitations of shape-matching eg pixel counting and measurement. Such systems are set relatively simple and well-defined tasks using prepared images and cannot recognize objects or shapes in a naturally occurring scene.

Figure 6 Isabelle Bulthoff's drawing of an office scene illustrating some difficult cases for visual processing.

Isabelle Bulthoff's drawing of an office scene (fig. 6) demonstrates a number of common problems for machine based visual recognition systems. Liter and Bulthoff comment that sophisticated artificial recognition systems have rudimentary recognition abilities and need to be primed with artificial subjects and artificial conditions in order to complete their tasks. For example, no artificial visual system can track objects which are obscured by other objects such as the chair behind the desk in Bulthoff's drawing. They comment:

> notice that the bounding contour of the chair in the lower right corner is identical to the bounding contour of the shadow on the back wall. Clearly no one would attempt to sit in the chair projected on the wall. Likewise, no one would attempt to sit in the chair atop the desk, though its image size is identical to that of the chair seen through the door on the back wall. Another difficulty that is apparent upon viewing this scene is that objects must be segregated from the background before they can be recognized (Liter and Bulthoff, 1998).

Indeed, segregating objects from their background is one of the primary functions of vision and, as Liter and Bulthoff comment, is no trivial matter. Unentangling objects from each other and their background is impossible for any system that relies on shape and property matching. Taken together with the other two difficulties mentioned, changing

angle and changing shape, the prospects look bleak for any theory of visual recognition which appeals to resemblance or shape property matching. As Liter and Bulthoff show, such a mechanism is simply unworkable.

Liter and Blanz provide an interesting illustration of the number of styles of chair one might encounter. Clearly one does not recognise a novel chair as a chair based only on its shape. When, in the first few hundred milliseconds of looking at an office scene we recognise chairs it is because we expect to see chairs. The exact shape of the chairs is not an issue – where they are situated in the *scene* is far more relevant.

Figure 7 Computer simulated chairs used by Blanz et al (1999)

It is only in the last 5 years, due primarily to the impact of change-blindness and inattention blindness research (Levin, 2002 and Noë, 2002), that vision research has shifted its focus away from shape-matching and object recognition to the problems of scene recognition. This shift has been accompanied by a focus on context as the primary factor in the initial recognition of scenes in the world.

Shape-matching is unworkable as a primary recognition process

The main obstacle to the context recognition argument is that it's hard to shake the intuition that we recognise things on the basis of their shape. Indeed, with a line drawing what else could we be using to recognise an object other than shape? It is clear when you look around you can recognise things because they have edges which demarcate where one surface meets another, or which indicate a shape boundary of some kind. Logic and intuition tell us that it must surely be the case that we use these object boundaries to identify what we are seeing. However, attempts to develop an account of how these edges

and boundaries can be used by the human visual system to recognise objects using shape matching have been unsuccessful because there are a number of insuperable problems which make shape-matching unworkable as the basis of a theory of how people recognize objects.

- **There are too many possible shapes to match**. The light stimulus stops at the retina. There are no shapes in the brain, just electrochemical signals. If recognition is based on matching the shape of the object we are currently looking at with shapes of objects we have seen, we need a mechanism for the light geometry on the retina to be matched with stored representations of past retinal geometries. This is an absurd scenario because the retina processes millions of views of thousand of scenes every day. It is far more likely that somehow the visual brain picks out objects and stores some kind of abstract notation of significant objects and looks for notation triggers for these objects, as opposed to their shapes, when trying to recognise what is in a scene. That is, we need a theory of object and scene recognition that is not based on shape matching.
- **We don't necessarily notice similar shapes**. Psychophysical tests have shown that even when a shape, such as a silhouette, is the perfect match for the outline of an object, subjects often fail to identify the object. That is, subjects do not notice the shape similarity between silhouette and object (see fig. 12 and the following analysis for demonstration of how similar shapes play different roles).
- **Similarity isn't enough to trigger recognition.** We encounter thousands of similar shapes in a day but don't mistake them for each other.
- **Speed of recognition precludes shape-matching**. If recognition is based on matching the shape in view with a stored shape how do we do it so quickly? There are infinite permutations of the ways the surfaces of the objects in your office, for example, can present to you. Does the human brain store millions of views of thousands of objects? Does it store 'visual models' of thousands of objects? If so, how can human beings categorise a scene in less time than it takes for a signal to go from the retina to the memory centres?
- **Even objects as familiar as chairs can have bizarre and unpredictable shapes.** Most days we will encounter a chair shape we have never encountered before and yet we identify it as a chair despite never having encountered the shape before. (see fig.7)

- **The light stimulus is ambiguous**. All theories of vision agree that the light stimulus that arrives at the eye is too ambiguous for human beings to recognise objects in the world based on retinal geometry. J. J. Gibson argues that because light obeys simple laws, and that the environment we encounter is regular, the light stimulus, together with information gained by using our sensorimotor system, eg moving our eyes, our heads etc, is adequate for recognition. Thus Gibson argues that the light stimulus is augmented by information about how our body is interacting with the environment. In a sense, he is maintaining that our embodiment provides a context for the stimulus at our retina. For example, if we know that we are upright and 20 metres from an object which is on a level plane with us we can judge its size and height because of what we know about how things look under those conditions. We can learn more about the object's shape by moving nearer or moving our head. The obvious problem with Gibson's account for theory of depiction is that none of these ruses work for pictures. We can't discover more about pictorial objects by moving our heads. Hence context is even more important in picture recognition than it is in real-life recognition.

In order to develop a workable theory of object recognition it is necessary to discover how the human visual system makes a virtue of the ambiguity of the light stimulus. How it picks out objects despite their shape. How it can quickly categorise objects it has never seen before. How it can categorise scenes so quickly.

The Primacy of Context: gist views and multiple visual systems

Context is the key to all these questions. Ambiguity is resolved by context – to some extent we know what objects to expect to see in an office scene, or a landscape or a portrait. Shape problems can be resolved contextually – we know that chairs will be near tables. Context will enable us to categorise novel objects in terms of what we expect in a scene.

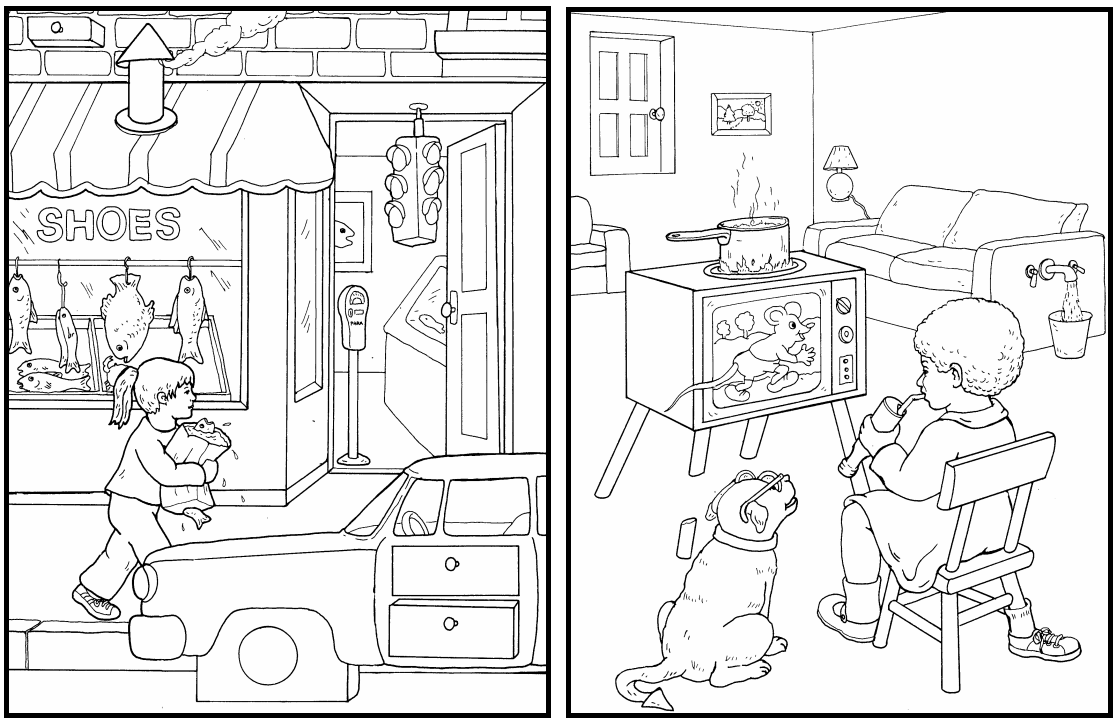

Figure 8 Drawings from *What's Wrong With This Picture*

The power of the recognition context can be experienced when you look at these drawings (fig. 8) from Anna Pomaska's book *What's Wrong With This Picture* (Pomaska, 1983). When objects are presented out of context they are the last things we notice rather than the first things. In figure 8 we read the scene as a street scene (40 milliseconds), we identify a child (150 milliseconds), a shop (200 milliseconds), and a car (300 milliseconds). Once our eyes begin to saccade around the pictures, we notice the anomalous parking meter, the fish in a bag and finally we notice that the car has a square wheel. The curious thing about this dawning of recognition is that we try and make the object fit the context despite the fact that it is anomalous. We don't notice the anomalous object until after we have categorised the scene.

Evidence of this incredibly swift scene categorisation has been brought to light in the study of 'gist views'. This research indicates that in the first 100 milliseconds we categorise the scene we are looking at as a landscape, a street scene, an interior etc. That is, the first thing our visual system does is set up the context for the objects in the scene. This initial categorisation is called the 'gist' view.

Figure 9 This is the kind of resolution your visual system provides in the first few hundred milliseconds of looking at a scene.

I have blurred the picture in figure 9 to emulate the kind of resolution your visual system is getting when it takes in the gist of the scene. It is black and white because 95% of our visual field is monochrome. It is blurred because the resolution of 90% of our visual field is very coarse. In order for us to see colour and detail and actual objects, our eyes must scan the scene in a series of saccades. This takes time. Our eyes saccade around a scene at the rate of about 10 times a second, each time fixating on a point.

From an evolutionary perspective being able to see highly detailed coloured shapes in the ‘recognition moment’ is not a priority. Being able to make extremely fast judgements about what is facing us is a high priority. Thus our visual system has developed a way to roughly categorise a whole scene based on very general, and very coarse, features in less than 100 milliseconds. In her paper ‘Gist of a Scene’, Aude Oliva writes:

> Behavioral studies have shown that observers can recognize the basic-level category of the scene (e.g., a street; Potter, 1976), its spatial layout (e.g., a street with tall vertical blocks on both sides (Schyns and Oliva, 1994), as well as other global structural information (e.g., a large volume in perspective) in less than 100 msec. Observers may also remember a few objects (e.g., a red car and green car),

the context in which they appear (e.g., parked on the side) and other low-level characteristics of regions that are particularly salient (Oliva, 2004, p. 251).

Other studies confirm that the gist of a scene is available for purposes of categorisation of a scene on or before the first saccade – sometimes less than 50 milliseconds (Castelhano and Henderson, 2005). Oliva makes a distinction between an object-centred categorisation and a scene-centred categorisation. The gist of a scene is developed before objects are identified; consequently the 'spatial envelope' needs to be defined before the object can be recognised. The categories which comprise a scene-centred description are more abstract than an object-centred description. Oliva and Torralba (2002) provide examples (fig. 10) of these two types of categorisation of the same scene.

Object-centered:
Sky, buildings, people, cars, trees, road

Scene-centered:
Large space, Man-made scene, Semiclosed space

Object-centered:
Lamp, sofa table, window

Scene-centered:
Small space, Man-made scene, Enclosed

Figure 10 Oliva and Torralba's object-centred and scene-centred descriptions

In their view, we initially use such categories as natural/man-made, large space/small space, open/enclosed. After tests with observers, who were asked to develop a vocabulary for scene-description, they further refined these categories into measures of the volume of the space and the scene properties such as, depth range, openness, expansion, ruggedness, verticalness, naturalness, busyness and roughness. The theory is that we categorise the 'spatial envelope' using something like these properties before we go on to spot a definite object. This basic level categorisation enables us to recognise a street-scene, a forest, a highway, a panorama, an office or a living room in our first glance.

It is clear from this recent work on gist that before we recognise objects we identify the spatial envelope. The spatial envelope provides the context for object identification. It is only after this initial categorisation of a scene that we populate it with objects.

The neurological mechanisms for gist recognition are the subject of some debate. There has been very little work done in this area; vision research has tended to concentrate on object recognition.[2] However, there is general agreement that we do not build up the gist of the scene from the parts. We take in the whole scene at once.

[2] Some earlier work on scene perception can be found in Kaplan and Kaplan (1982) and Potter (1976).

Consequently, many of the hierarchical models of object perception, and indeed their neural models, do not work for gist recognition. Oliva observes:

> speed and accuracy in scene recognition are not affected by the quantity of objects in a scene, and recognition can be achieved equally well even when object information is degraded so much that objects cannot be locally recovered (Oliva, 2004).

One of the impediments to developing a neurological account of gist recognition has been the assumption that there is either a local-to-global or global-to-local feedback mechanism which drives the process. However, any cortical feedback mechanisms are likely to be too slow for gist recognition (Rasche and Koch, 2002). For example, in the famous Fuchs saccade study of how our eyes read a face (see fig. 11) we can see that the gaze moves around and fixates on key areas in the matter of a second or two. The problem, as Fuchs points out, is that there isn't enough time for given fixation to get feedback from the higher areas to tell the eye where to go next.

> Saccades are so rapid that they are over before visual feedback can help guide them to the target. Nevertheless, they are very accurate. Therefore, the neural command that drives the eye muscles must be programmed very precisely in advance (Fuchs, 2006).

Fuchs project is to discover 'how a sensory stimulus elicits an appropriate eye movement response' (Fuchs, 2006). One of the puzzles is how the eye can move so accurately to its target when the gist view is so rough. One suggestion is that our eyes are pre-programmed to read a face. However, there must be some kind of early feedback to the visual system that tells it that a face, or even landscape, is in view.[3]

[3] The question is: Which area of the brain is giving feedback to which area? Rasche and Koch (2006) tentatively suggest that there are a number of local feedback mechanisms routed through the lateral geniculate nucleus (LGN). Epstein and Kanwisher (2006) found a region of cortex referred as the parahippocampal place area (PPA) which responds more strongly to pictures of what they call 'intact scenes'. What Epstein and Kanwisher call 'intact scenes' correspond roughly to Oliva and Torralba's (2005) scene-centred classifications. This work on scene recognition and gist promises to become immensely important in the study of how we see paintings in terms of their whole composition.

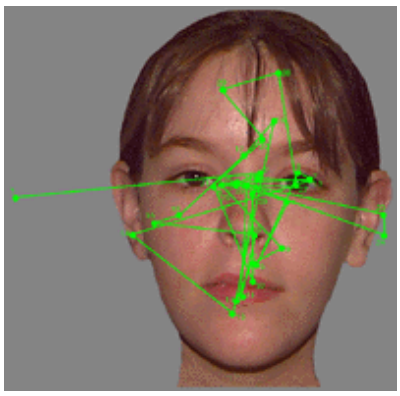

Figure 11 A trace of the saccades made by looking at a face over a few seconds

Remember, the gist view is not just coarse but it is black and white. Our early visual processing registers the whole picture coarsely in black and white. We are able to take in a whole scene with our whole retinal array in around 100 milliseconds – in this first glance we are not attending to anything in particular and somehow on the basis of that first glance we are able to direct our fovea towards salient objects in the scene. It is clear that the human visual system has evolved a number of different methods for evaluating the light stimulus on the retina. Each time we look at a scene, a number of competing and parallel processes spring into action and evaluate it. We do not rely on the light stimulus because we know it is ambiguous. The human visual system has evolved to overcome this ambiguity and the best way to eliminate ambiguity is to obtain and analyse multiple versions of the scene in different ways. It has been known for some time that the visual system processes colour, motion and edges using specialized modules. It also seems to be the case that there is a specialized system for face and hand recognition. Goodale and Milner in their book *Sight Unseen* maintain that there are at least two visual systems (Goodale and Milner, 2004).

- One is for navigation and establishing where we are in relation to things. The proceeds of this system do not present to consciousness.
- The other is perceptual and is responsible for working out what we are looking at. The proceeds of this system are our visual phenomenology.

We also know that there are specialized reflex systems which enable us to react quickly to things, such as things flying directly at us. The ‘gist view’ system enables us to categorise a scene as outdoors, indoors, large space, small space etc within the first 30 or 40 milliseconds of a view. This response is so fast that our visual system cannot possibly be evaluating the scene in terms of its constituent objects. In tests, the category response times were so fast that it is clear that the higher brain and memory are not involved in this

initial categorisation. Thus the ‘spatial envelope’ of a scene is determined by a reflex action before our eyes can make their first saccade to an object in the scene. In the Pomaska pictures above (fig. 8) we categorised the scenes as street-scene and living room based on a gist view long before we identified an actual object. In a sense, our visual system works outside-in rather than inside-out; from context to object, not from object to scene. The thrust of object recognition research until a few years ago has been how we build an object up from its parts (eg geon theory). This idea that we recognise things by identifying their components has also influenced theories of depiction. Most theoretical work on depiction assumes that we when we look at a sketch, for example, we build up an object or scene from the individual marks. This is the wrong way of thinking about it.

We can see in the drawings in figure 12 that there is a diversity of ways that a line can refer. It is context which determines the kind of edge which a line demarcates.

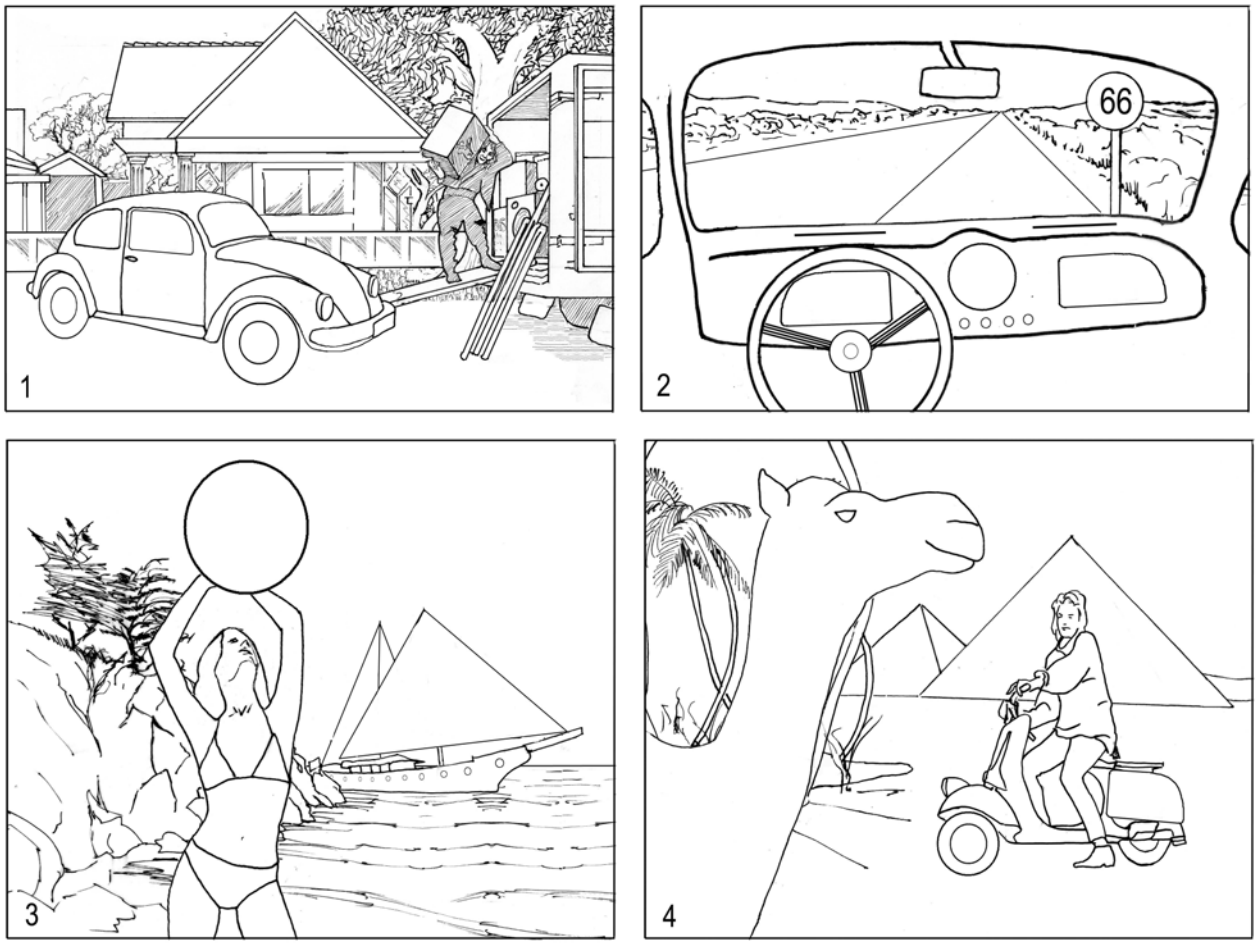

Figure 12 Lines refer according to their context

In picture 12.1 there are lines on the Volkswagen which indicate cracks around the door, the bound of the curved roof, and the edge of the running board. These are all just lines but they all demarcate different kinds of boundaries. They represent different

structures in the world. But lines can refer in even more ways. The line up the middle of the road in 12.2 represents a pigment line and the lines radiating out from the centre of the driving wheel are wires. In 12.3 there is a line which represents the horizon, and a line which indicates an indentation at the base of the girl's neck. According to John M. Kennedy there are at least eight ways that a line can refer, and the multiple functions of lines reflect the ambiguity of how edges may be perceived in the real world (Kennedy, 1974). His argument is that the exact configuration of an edge or boundary in the real world is inherently ambiguous. We need to assess any edge according to its context. Just as the individual lines which I highlighted in the above drawings only refer because of how they relate to other lines, the shapes which we see only refer because of their relation to other shapes. Each drawing above features a triangular shape which we interpret as a different kind of object in each context: the gable of a house; a receding road; a sail; and a pyramid. In each case the shape in the drawing is the same, but when you looked at each picture you instantly recognise the object. It wasn't the shape you recognised but the object.

It seems to be the case that our visual system does not pick out similar shapes. That is, it does not pick out the constituent shapes of an object and compare that shape with shapes in other objects. That is not how recognition works. There is no shape matching going on.

Notice that there is a circle in each of the drawings but in the first drawing it is seen as a car wheel, in the second as a road sign (a flat disk), in the third as a sphere (a beach ball) and in the fourth as a scooter wheel. It is irrelevant to our visual system that these are all just flat circles in these drawings. Flat circles are not objects and, just as in the case of the Ponzo drawing, it is three dimensional objects that our visual system is primed for and recognises.

In order to demonstrate that in the initial recognition moment you do not register all the flat circles, I've embedded some additional circles in the drawings. For example, the port holes on the boat and the tachometer on the car dashboard are circles. I have also repeated a shape in the drawings which signifies a very different object in each case. The window of the Volkswagen, the right glove compartment on the dashboard and the back of the scooter are all exactly the same shape. I have done this to demonstrate how irrelevant the shape is in the initial recognition of the Volkswagen, dashboard and scooter. It is the overall configuration of parts, in a very general sense that triggers recognition. Oliva's and Fuchs' work demonstrates that we recognise these scenes as a suburban street,

a highway from a car, the seaside and the pyramids, long before our eyes focus on an object or discern a distinct shape.

Categorising the 'spatial envelope' from the gist view is just one of dozens of visual evaluations which commence as soon as we look at a scene. It may be that the gist view acts as an initial diagnosis of the scene and that what follows are a series of visual tests which confirm of disconfirm the diagnosis. Our visual system throws a multitude of tests at any scene to resolve ambiguity. It is known from the work of neurologists such as Hubel (1988) and Zeki (1993) that our visual system separately evaluates the light stimulus for motion, edges, and colour. It is clear that there are dedicated neural modules which abstract the retinal stimulus and parallel process it for basic features. Recent work on the neurology of object recognition indicates that even at the object level the visual brain is testing the light stimulus for quite specific triggers. Clearly the primate visual system uses multiple strategies for evaluating the object stimulus. These object evaluation strategies are subsidiary to the initial evaluation of the spatial envelope and the scene context.[4] Consequently an account of how the human visual system evaluates objects cannot explain how we identify objects in pictures or in real life. Visual recognition does not *begin* with object identification it *ends* with object identification. Prior to object recognition, multiple brains systems and brain processes are engaged in assessing the situation in which you find yourself. These processes contextualise the objects in view and are a necessary prerequisite to recognition. When you look at a scene your visual system attempts to ascertain what in the world could be in front of you by making an initial diagnosis and then simultaneously testing the scene for hundreds of quite specific things. The process is analogous to the way blood tests are used to diagnose disease. The doctor makes an initial diagnosis and the nurse takes a syringe full of blood and does a number of tests which are looking for something specific – elevated uric acid levels, low white-cell count etc. These tests confirm or disconfirm the diagnosis. The doctor can't just do one test which shows you have gout, for example. There are a series of tests for symptoms and each test is quite specific. The process is essentially indirect. The visual system also does dozens, possibly hundreds, of simultaneous tests on a scene and calculates what is in view by a process of trial and error. Of course, as in the case of the doctor, it helps if the visual system makes a good initial diagnosis. This initial assessment, which we might call it the 'gist assessment' ensures that the right tests are carried out.

[4] See Antonio Torralba et al (2006) for an account of work evaluating the role of context in identifying elements in a scene.

Recent neurological work on the visual brain indicates that some of these tests and the brain processes which enable them to be carried out are radically counter-intuitive. Work by Keiji Tanaka at the RIKEN Brain research Institute in Saitama suggests that when we identify objects a specific set of neurons tests the stimuli for specific object types (Tanaka 2003). That is, the visual brain is object testing at a neuron level using a procedure that does not appear to involve higher memory functions. Brain scan work on recognition by Rodrigo Quiroga at the University of Leicester suggests that there are individual neurons which fire when an iconic object or person is seen in a photograph, drawing or even the name of the object is seen (Quiroga et al, 2005). Quiroga found that there were neurons which were triggered by famous scenes, such as a picture of the Sydney Opera House or the Eiffel Tower, and by pictures of famous people such as Halle Berry or Mother Teresa. The idea that individual neurons can be dedicated to specific objects in the world had previously been derided by vision researchers.

Research such as this, together with work on change-blindness and gist views, indicates that theories of recognition based on shape-matching and resemblance are simply naïve. The evidence from this work on the human visual system indicates that the mechanisms of recognition are radically counter-intuitive. It is clear that context, in the broadest sense of the term, creates the visual system's initial diagnosis of a scene and determines what kind of results the visual system is looking for in the tests which it conducts. This work has fundamental implications for theories of depiction. In particular, a shift away from shape-matching and object recognition theories towards and an emphasis on context evaluation and scene recognition will provide fresh approaches to a recognition-based theory of depiction.

Bibliography

Blanz, Volker, Tarr, M. J., and Bulthoff, H. H. (1999), 'What object attributes determine canonical views?' *Perception*, **28** (5), 575-99.

Castelhano, Monica S. and Henderson, John M. (2005), 'The influence of color on perception of scene gist', *Journal of Vision*, **5** (8), 68-68.

Epstein, R. and Kanwisher, N. (1998), 'A Cortical Representation of the Local Visual Environment', *Nature*, **392**, 598-601.

Fuchs, Albert (2006), 'fuchs@u.washington.edu Community of Science page', accessed 15th May.

Gibson, James J. 'Old and New Assumptions for a Theory of Visual Perception from the Purple Perils', http://www.trincoll.edu/depts/ecopsyc/perils/folder2/oldnew.html>.

Goodale, Melvyn A. and Milner, A. David (2004), *Sight unseen: an exploration of conscious and unconscious vision* (Oxford: Oxford University Press).

Howe, Catherine Q. and Purves, Dale (2004), 'Size Contrast and Assimilation Explained by the Statistics of Natural Scene Geometry ', *Journal of Cognitive Neuroscience* **16** (1), 90-102

Hubel, D.H. (1988), *Eye, Brain and Vision* (New York: Scientific American Library).

Kaiser, Peter K. (2007), 'The Joy of Visual Perception: A Web Book ', (Toronto: York University).

Kaplan, Stephen and Kaplan, Rachel (1982), *Cognition and Environment* (New York: Praeger).

Kennedy, John M. (1974), *A Psychology of Picture Perception* (London: Jossey-Bass).

Levin, Daniel T. (2002), 'Change Blindness Blindness As Visual Metacognition', in Alva Noe¨ (ed.), *Is the Visual World a Grand Illusion?* (Thorverton: Imprint Academic).

Liter, J. C. and Bulthoff, H. H. (1998), 'An introduction to object recognition', *Zeitschrift Fur Naturforschung C-a Journal of Biosciences*, 53 (7-8), 610-21.

Noë, Alva (2002), 'Is the Visual World a Grand Illusion?' in Alva Noe¨ (ed.), *Is the Visual World a Grand Illusion?* (Thorverton: Imprint Academic).

Oliva, Aude (2004), 'Gist of a Scene', in Laurent Itti, Geraint Rees, and John Tsotsos (eds.), *Neurobiology of Attention* (New York: Academic Press).

Oliva, Aude and Torralba, Antonio (2005), 'Building the Gist of a Scene: The Role of Global Image Features in Recognition', <http://cvcl.mit.edu/Papers/OlivaTorralbaPBR2006.pdf>, accessed 15th May.

Oliva, Aude and Torralba, Antonio B. (2002), 'Scene-Centered Description from Spatial Envelope Properties ', *Biologically Motivated Computer Vision: Second International Workshop, BMCV 2002, Tübingen, Germany, November 22-24, 2002. Proceedings* (Lecture Notes in Computer Science 2525; Berlin: Springer).

Pomaska, Anna (1983), *What's Wrong with this picture* (New York: Dover Publications, Inc).

Potter, M.C. (1976), 'Short-term conceptual memory for pictures', *Journal of Experimental Psychology: Human Learning and Memory*, **2**, 509-22.

Quiroga, R. Q., et al. (2005), 'Invariant visual representation by single neurons in the human brain', *Nature*, **435** (7045), 1102-07.

Ramachandran, V. S. (1988), 'Perceiving Shape from Shading', *Scientific American*, **259** (2), 76-83.

Rasche, Christoph and Koch, Christof (2006), 'Recognizing the gist of a visual scene: possible perceptual and neural mechanisms', *Neurcomputing* <http://www.elsevier.com/wps/find/journaldescription.cws_home/505628/description#description>, accessed 15 May.

Sigala, N. and Logothetis, N. K. (2002), 'Visual categorization shapes feature selectivity in the primate temporal cortex', *Nature,* **415** (6869), 318-20.
Tanaka, K. (2003), 'Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities', *Cerebral Cortex,* **13** (1), 90-99.
Torralba, Antonio, et al. (2006), 'Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search', *Psychol Rev,* **113** (4), 766-86.
Zeki, Semir (1993), *A Vision of the Brain* (Oxford: Blackwell Scientific Publications).