Attention, Visual Search and Object Recognition

Antonio J. Rodriguez-Sanchez

Technical Report CSE-2004-11

June 1, 2004

Department of Computer Science and Engineering
4700 Keele Street Toronto, Ontario M3J 1P3 Canada

Attention, Visual Search and Object Recognition

Antonio José Rodríguez Sánchez

PhD candidate under the supervision of John K. Tsotsos

Department of Computer Science
3001 Computer Science and Engineering
York University
4700 Keele Street
Toronto, Ontario M3J1P3
Canada

email: ajrs@cs.yorku.ca
Tel: 416-736-2100 x33257

Attention, visual search and object recognition

Antonio José Rodríguez Sánchez

PhD candidate under the supervision of John K. Tsotsos

Department of Computer Science

York University

1 Introduction

One of the main difficulties that arises when designing automatic systems is developing a mechanism that can recognize, or simply find, an object with the ease with which the human visual system does. Humans can recognize objects effortlessly under variations in location, lighting and viewpoint.

By studying the processes that occur in the human brain when analyzing a visual scene we may be able to construct a model to simulate this behaviour. Some of these processes have been studied using neurophysiological methods. Chapter 2 summarizes key findings about how the human visual system locates objects in a scene.

An important issue for this is attention. The Encyclopedia Britannica defines attention as "in psychology, the concentration of awareness on some phenomenon to the exclusion of other stimuli". Due to the capacity limitations of the brain not all the visual information that impinges our retinas can be processed (Tsotsos, 1990). Attention must be applied to this information in order for the visual system to focus processing on salient information, while filtering or inhibiting other parts of the visual scene. Chapter 3 summarizes several studies that demonstrate that attention is involved in different areas of the brain and describes the attempts to model attention dating back to the early 1970's. Chapter 3 also overviews what *features* are important for analyzing a scene and the techniques the brain may be using to find objects in this scene. A huge amount of psychophysical data has been collected in which the task is to find an element defined by different characteristics among distracters. Several theories have emerged to explain the results obtained in these experiments.

In object recognition, the goal is to locate an object in a scene. Object recognition methods generally apply algorithms based on geometrical methods for finding and recognizing an object in a scene. Chapter 4 briefly summarizes some of the methods used. Chapter 5 concludes the presentation by making several observations based on the reviewed literature.

2 Human vision

Since the foundation of the modern neuroanatomy by Ramón y Cajal (1904), who gave a detailed description of nerve cell organization in the central and peripheral nervous system, great progress has been achieved in understanding the human brain.

This chapter aims to provide a basic overview of the human and primate visual system focused on object recognition. Section 2.1 provides a review of the human retina. Section 2.2 will deal with the Lateral Geniculate Nucleus (LGN). Finally, Section 2.3 introduces the cerebral cortex and its role in object recognition.

2.1 The eye

The eyes (Figure 2.1) are the human organs for producing an image of the real world for the brain. The eye is composed of photoreceptors and retinal cells that convert visual information to neural signals.

The retina is located in the inner surface of the eye (Figure 2.1). It contains photoreceptors that perceive only part of the visible light spectrum. For humans, light with wavelengths between 380 (violet) and 760 (red) nanometres is visible.

There are two kinds of photoreceptors: rods and cones (Figure 2.1). The rods are distributed uniformly across the retina except at the fovea (high density of cones) and the blind spot (where there are no photoreceptors), and have high sensitivity to low levels of brightness (useful for dark situations). Cones have a very high density at the fovea, and a lower concentration than rods in the rest of the retina and require high levels of brightness. We can classify the cones as a function of their wavelength absorbance as red, blue and green cones. These three cone types allow for the perception of colour (Bowmaker and Dartnall, 1980).

The photoreceptors are connected to horizontal and bipolar cells. A bipolar cell can receive connections directly from the photoreceptors or indirectly through the horizontal cells (Figure 2.2a). The direct connections can be either excitatory or inhibitory, the indirect pathway is always of opposite sign. The integration of these two inputs generates a centre/surround (Mexican hat receptive field) response in the bipolar cell.

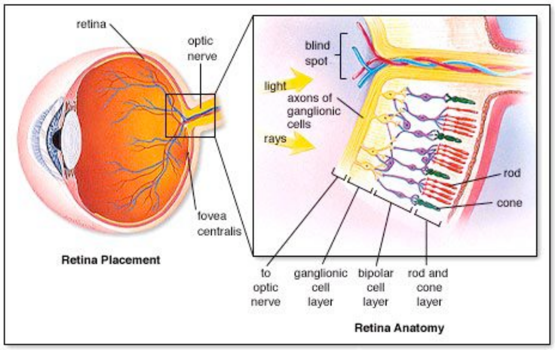

Figure 2.1. The eye and organization of the retina

The output of the bipolar cells are integrated in the ganglion cells (Figure 2.1 and 2.2b) that also have a centre/surround design with an inner circular centre and a surround ring (Figure 2.2b). An on-centre/off-surround ganglion cell would increase its firing rate for light presented to its centre and would decrease its firing rate for light presented in the surround. An off-centre/on-surround ganglion cell would have the opposite response.

2.2 The Lateral Geniculate Nucleus (LGN)

The optic nerve connects the eye to the lateral geniculate nucleus (LGN) and the Superior Colliculus (SC), which is involved in eye movements. The LGN is structured in six layers, the four dorsal layers receive information from P ganglion neurons and are known as the parvocellular layers. The two ventral layers receive information from M ganglion neurons and are known as the magnocellular layers. The layers of each eye are interleaved (Minkowski, 1920). The topographic locations of the retina are maintained in LGN, so that proximal regions on the retina are projected to proximal locations in LGN.

Most LGN neurons have a circular centre-surround structure, but differ in sensitivity in four main characteristics: colour, contrast sensitivity, spatial resolution and latency. parvocellular neurons have a high sensitivity to colour whereas LGN magnocellular neurons are not sensitive to colour (Livingstone and Hubel, 1988). Similarly to the light on-off centre/surround behaviour of ganglion neurons, Parvocellular neurons have a colour opponent centre/surround design with to red-green and blue-yellow opponency.

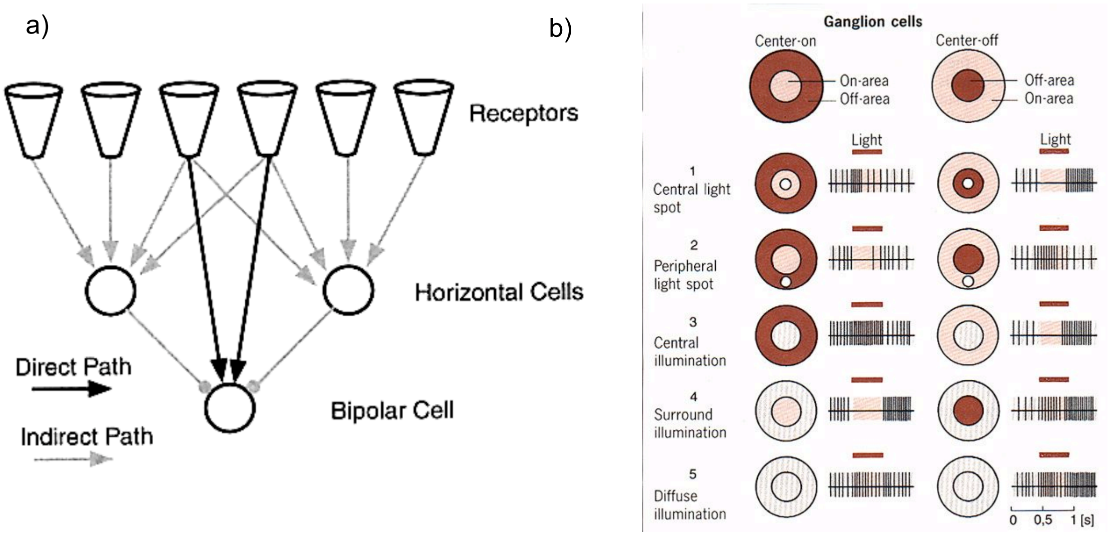

Figure 2.2 a) Centre-surround behavior of bipolar cells **b)** Centre-surround ring of ganglion cells

With regards to contrast, the magnocellular neurons are more sensitive to contrast than the parvocellular neurons in centre-surround difference of brightness (Shapley and Perry, 1986). The Magnocellular neurons have receptive fields two to three times the size of the parvocellular neurons receptive fields and respond to a stimulus before the parvocellular neurons (Livingstone and Hubel, 1988). This property makes the magnocellular neurons suitable for the temporal processing of vision (motion).

These Parvo-Magno differences are maintained in the visual cortex. In a first approximation, the P path is believed to be responsible for shape and object perception while the M path can account for the perception of motion and sudden changes.

2.3 The Primate Cerebral Cortex

Figure 2.3 shows the cerebral cortex. The cerebral cortex is composed of neuronal cell bodies that form the outer layer of the cerebrum. It controls the most complex mental activities such as memory, learning, problem solving, planning, vision, audition and action. The cerebral cortex is divided (sagitally) into two hemispheres, which are each further subdivided into four separate regions (Figure 2.3):

- Temporal lobe: Contains neurons that register sound qualities in the Primary Auditory Cortex. It also has neurons that are involved in the comprehension of the language, memory and learning.

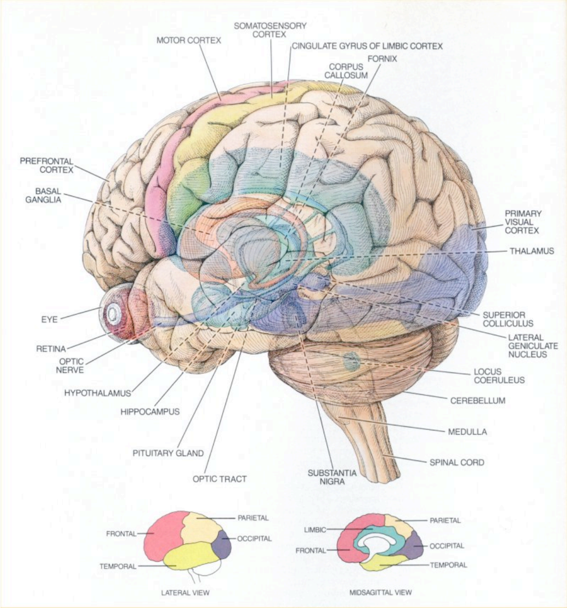

Figure 2.3. Structure of the brain

- Frontal lobe: Holds mainly the Primary Motor Cortex, in which there are neurons that control the muscles of the body. It is organized in terms of the parts of the body that it controls.
- Parietal lobe: Contains the Primary Somatosensory Cortex Located composed of neurons that register the sense of touch, and it is also organized by body parts.
- Occipital lobe: The Primary Visual Cortex is localized in its posterior part and processes visual information. It is the subject of study in this chapter.

Neurophysiological studies usually use non-human primates (e.g. monkeys) because the techniques used for studying the neurons in the brain are usually invasive.

The macaque monkey visual cortex occupies 55% of the neocortex (compared with the 11%, 8% and 3% of somatosensory, motor and auditory areas), there are at least 32 neocortical areas involved in vision (Felleman and van Essen, 1991), and inputs can come from auditory, somatosensory, or visuomotor activity. Felleman and van Essen identified 305 pathways, of which 242 are have bidirectional connections, although they can vary in strength (e.g. connection V1-V4 is robust, but V4-V1 is weak). Area connections are organized hierarchically with upwards, downwards and lateral connections.

2.4 Areas of the Visual Cortex

As shown in Fig 2.4, the visual cortex is organized into different areas. V1 and V2 are the largest, each having an area of approximately 1100 – 1200 mm^2 (11-12% of the macaque neocortex) (Felleman and van Essen, 1991).

Physiological studies show two different pathways with some connections between them: The occipitotemporal pathway (V1, V2, V4, AIT and PIT) is related with object recognition features (color, shape, etc.), while the occipitoparietal pathway (V1, V2, V3, MT and MST) is associated with spatiotemporal characteristics of the scene (direction of motion, etc.) (Webster and Ungerleider, 1998)

Below is a brief overview of the most important areas related to object recognition (ventral pathway) are reviewed.

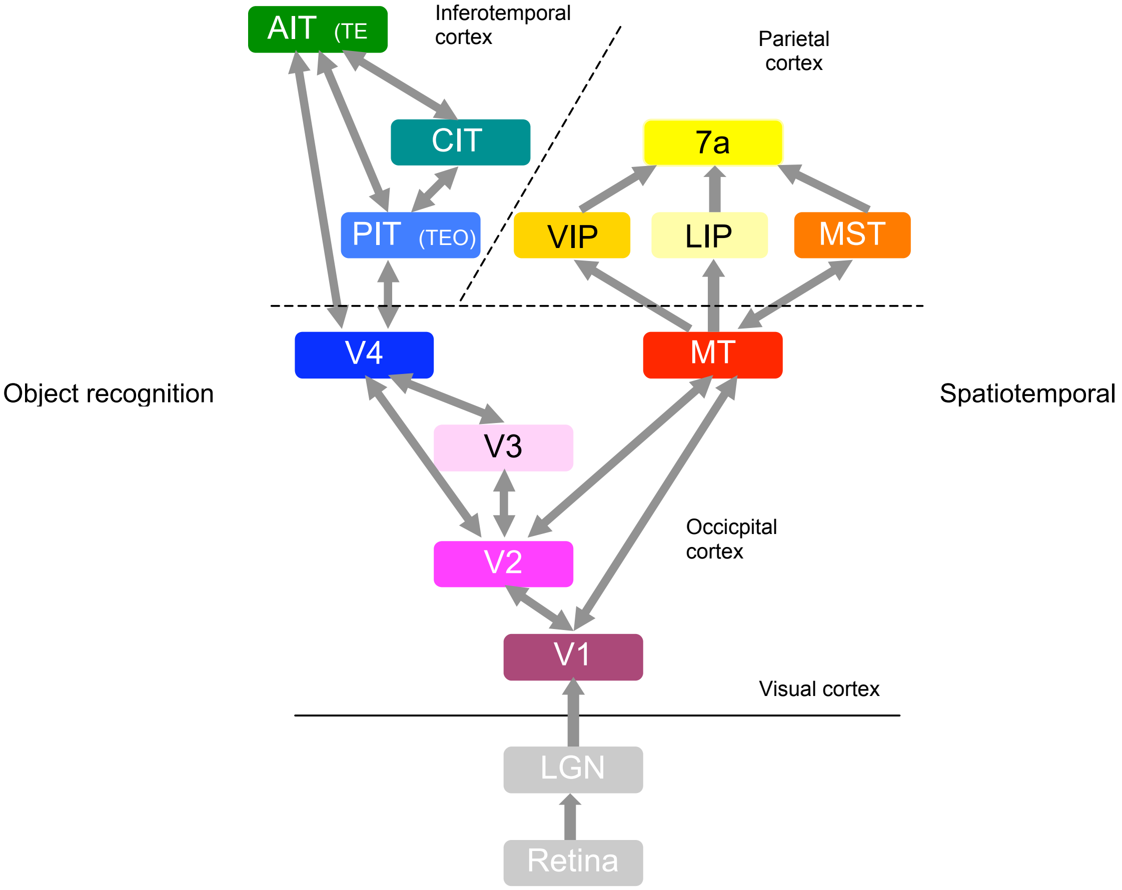

Figure 2.4 Simplification of connection areas from Felleman and van Essen (1991)

2.5 Visual Area 1 (V1)

Hubel and Wiesel (1959, 1962) discovered that cat V1 neurons respond when bars and edges are presented in their receptive field. This same characteristic was later found in monkeys (Hubel and Wiesel, 1968).

Following Rolls and Deco (2002) and Hubel and Wiesel (1959, 1962, 1968), V1 neurons can be classified into three types: simple cells, complex cells, and endstopped cells.

Simple neurons, have a small receptive field (0.25°-1°), are close to the fovea, and their response is based on small areas relative to the background. They respond to bars and edges with different orientations and to spatial frequency (De Valois and De Valois, 1988) in a way that may be modelled by a Gabor filter (Marcelja, 1980; Rolls and Deco, 2002) or a Difference of Gaussians (Hawken and Parker, 1987) (Figure 2.5a):

$$RF(x) = k_c e^{\left(-\left(\frac{x}{x_c}\right)^2\right)} - k_s e^{\left(-\left(\frac{x}{x_s}\right)^2\right)}$$

where x_c and k_c are respectively the space constant and amplitude of the centre Gaussian. Likewise, x_s and k_s are the space constant and amplitude corresponding to the surround Gaussian.

Complex neurons are also sensitive to bars and orientations, but their receptive fields are larger than those of simple neurons and they are less sensitive and are usually directed by simulated motion. Like simple cells, complex cells are selective for bars presented at a preferred orientation in the receptive field and they are tuned for spatial frequency. In contrast to simple cells, they will respond irrespective of the particular position at which a bar is flashed in the receptive field and are largely insensitive to the polarity of the stimulus.

Endstopped neurons require the termination of an edge or bar located in their receptive field in order to respond and can be simple or complex.

There are three paths in V1: stereopsis and motion (magnocellular neurons), color (parvocellular neurons) and form (parvocellular neurons) pathways (Livingstone and Hubel, 1988). In area V1, the retinal fovea has a much larger area of representation than the periphery. There is also *ocular dominance* (LeVay et al., 1975), neurons dominated either by the left or the right eye.

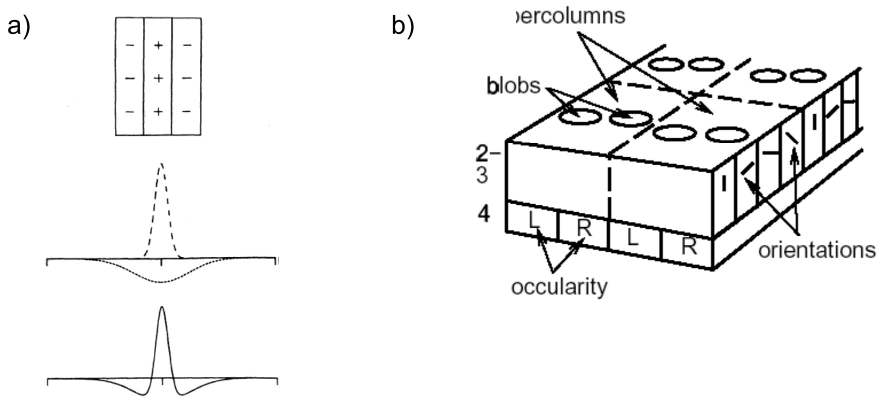

Figure 2.5. a) Fitting of V1 RFs with a difference of Gaussians *Source: Hawken & Parker, 1987)*
b) Hypercolumns in V1 *(Source: O'Reilly & Munakata 2001)*

V1 is arranged into hypercolumns, that contain columns tuned to different orientations in one dimension and alternating ocular dominance in a second dimension (Figure 2.5b). One third of V1 neurons can be activated by the M or P pathways alone. V1 projects mainly to V2, but also to area MT and V3 (Lennie, 1998).

2.6 Visual area 2 (V2)

V2 receives its input mainly from V1. Colour, form and stereopsis/motion V1 pathways continue in V2 (Figure 2.6). Colour path V2 neurons are not orientation selective. Half of them are colour sensitive with centre-surround antagonism and their receptive field centres are larger than their corresponding V1 colour pathway neurons (Livingstone and Hubel, 1988). Neurons in the Form V2 path have selectivity to orientation but not direction and half of them respond to terminations of edges or bars (big increase compared to V1). They are not selective to colour. V2 stereopsis path neurons have orientation selectivity and are occasionally responsive to bars or edge terminations. They responded poorly when only one eye is stimulated, but strongly when there is information coming from both eyes. Poggio and Fischer (1977) found also that these neurons are sensitive to large disparity. For these reason, these neurons seem to be mainly selective for stereoscopic depth and motion.

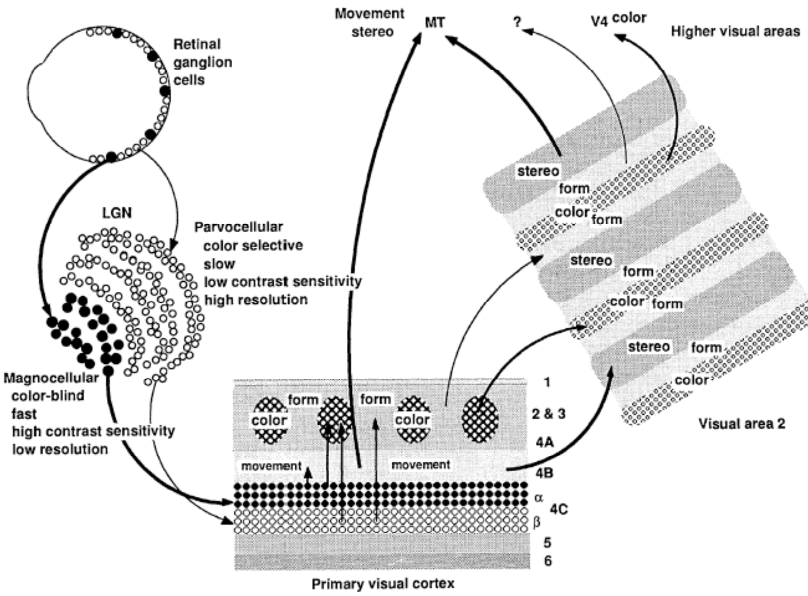

Figure 2.6. Connections LGN-V1-V2. There are three main paths that projects into Visual Area 2: (1) Parvocellular neurons → V1 (4Cβ) → V1 (1,2,3,4A; form) → V2 (form), Parvocellular neurons → V1 (4Cβ) → V1 (1,2,3,4A; color) → V2 (color) and (3) Magnocellular neurons → V1 (4Cα) → V1 (4B) → V2 (stereo). *(Source: Livingstone and Hubel, 1988)*

The main qualities of V4 neurons may be that they show selectivity to luminance (Schein and Desimone, 1990; Heywood et al., 1992, Motter et al., 1994) and colour constancy (Zeki, 1983). They also manifest sensitivity to length, width, orientation, direction of motion and spatial frequency (Desimone et al., 1985).

As V4 is not the only place where the processing of colour occurs, V4 is responsible for more than colour constancy. There is evidence that V4 is important for the perception of form and pattern/shape discrimination as was shown in several studies where area V4 was ablated in monkeys (Heywood and Cowey, 1987; Merigan et al., 1998).

Wilson et al. (1997) and Wilson and Wilkinson (1998) constructed a model that fitted their psychophysical results on glass patterns. Glass patterns are random dot patterns in which the orientation of each dot pair is tangent to the contours of a global pattern. They found that subjects were more sensitive concentric glass patterns than for radial, hyperbolic and parallel glass patterns. Their results were supported by a later functional magnetic resonance imaging (fMRI) experiment (Wilkinson et al., 2000).

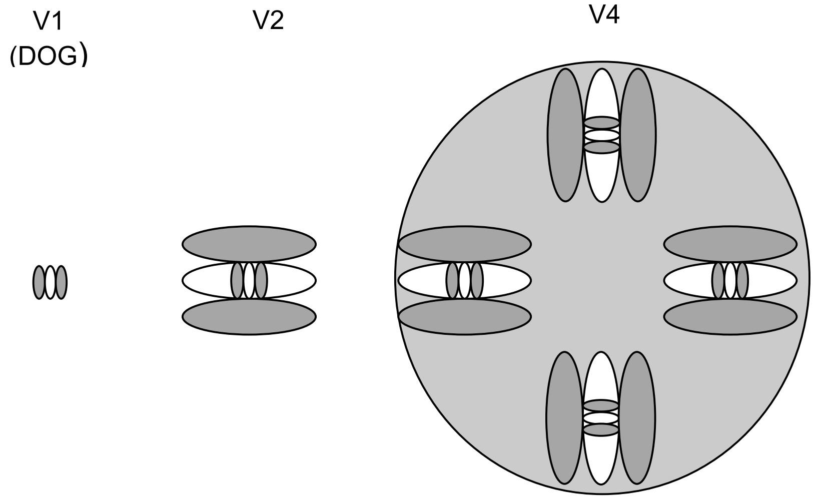

Figure 2.7. V1-V2-V4 processing proposed by Wilson et al. (1998)

2.7 Visual area 4 (V4)

V4 neurons have receptive fields that range from 2° to 4° (Felleman and Van Essen, 1991). Some studies suggested that V4 is the centre of colour processing due to the large number of neurons with high sensitivity to colour (Van Essen & Zeki, 1978). Later studies suggested that this number may not be so high (Schein et al., 1982; Heywood et al., 1988).

Wilson et al. (1997) and Wilson and Wilkinson (1998) proposed a model for the integration of shape in V4 neurons. They based their model in psychological results from glass patterns, which are random dot patterns in which the orientation of each dot pair is tangent to the contours of a global pattern. They found that subjects were more sensitive to concentric glass patterns than to radial, hyperbolic and parallel glass patterns. Their results were supported later by an fMRI (functional magnetic resonance imaging) experiment (Wilkinson et al., 2000). In their model V4 receptive fields perform a summation from V2 neurons in a circular shape (Figure 2.7). Their model gives an array of concentrically organized units sensitive to contour curvatures and forms.

2.8 Inferotemporal cortex (IT)

The macaque monkey's inferotemporal cortex (IT) receives mainly inputs V2 and V4. IT is believed to be an area involved in object recognition and discrimination. The first studies supporting this claim were based on research on monkeys in which IT was ablated (Gross 1972, Dean 1976). Ablation resulted in deficits on tasks that involved visual discrimination or object recognition. IT neurons are view-independent, translation, space and size invariant and respond mainly to objects and faces (Rolls and Deco, 2001).

There is more recent evidence that IT is involved in object recognition. Tanaka et al. (1991) divided IT in two main parts: Posterior IT (PIT or TEO) and Anterior IT (AIT or TE). In TEO (close to V4), most of the neurons were activated maximally by a simple combination of features such as bars or disks varying in size, orientation or colour. Tanaka et al. (1991) called these neurons *primary cells.* TE (comprising two-thirds to three-quarters of the IT area) neurons required more complex features for maximal activation. Tanaka et al. (1991) called these neurons *elaborate cells*, supporting the Felleman and van Essen (1991) architecture where TE receives its input from TEO. TEO may be responsible for medium complexity features whereas TE is responsible for high complexity features (Tanaka, 1996).

Tanaka et al. (1991), in a series of complex experiments recording IT neuron activation from anesthesized monkeys, found that primary neurons were selective for orientation, bars, disks and texture; while colour was generally not a relevant feature. Elaborate cells only responded to different shapes (e.g. stars, combination of a disk and a bar, T shape, rounded tongue, etc.) or to a combination of different features (shape and texture, shape and colour, texture and colour, texture and colour and shape). Tanaka et al. (1991) also found neurons that responded only to faces or hands. These elaborate neurons were also sensitive to the orientation and size of the stimulus. Tanaka et al. (1991) also reported that neurons in anterior IT (TE) had larger receptive fields (from 12.38° ± 8.89° for primary cells to 13.62° ± 7.32° for elaborate cells) than posterior IT (TEO) neurons (from 3.66° ± 4.34° for primary cells to 10.12° ± 7.96° for elaborate cells). Receptive fields of TE neurons usually include the fovea. An important observation about elaborate cells was that objects were coded by combinations of active neurons, each representing to a particular feature. A single neuron was enough for a face or a hand. This neuron did not respond to other objects and had different tunings for different faces, this supports Rolls (1987) in that face neurons encode different faces in a distributed way.

IT is not retinotopically organized. As a result, the connections between TEO and TE have to be in feature space. TE can perform position invariance because it receives inputs from V4 and TEO receptive fields neurons at different retinal positions (Tanaka, 1996).

Fujita et al. (1992) showed that anterior IT not only responded to fairly complex objects, but also that it had a columnar organization in a similar way as for orientation of stimulus contours in V1 (Hubel and Wiesel, 1962) and for area MT (Albright et al., 1984). Marr and Nishihara's (1978) assertion that "a useful representation system for object recognition should satisfy two conflicting conditions: ability to reflect the degree of similarity between two shapes in their descriptions and sensitivity to small differences between the two" may be solved in a columnar way. A column may encode a feature common to similar shapes and the activity of individual neurons may account for small differences between those objects. Related features overlap creating a continuum of features (Tanaka, 1996; Figure 2.8). This columnar organization in IT was demonstrated in TE using optical imaging (Wang et al., 1998) and extracellular recordings (Tsunoda et al., 2001) in awake monkeys. Wang et al. (1998) found several properties of this columnar organization: First, single features activated different spots. Secondly, different parts had different selectivities, some regions were activated by only one stimulus (of 16), while other regions were activated by more than three stimuli. Lastly, faces in different orientations activated partially overlapping regions, these regions were arranged in a continuous map of the view angle along the cortical surface. Face recognition was composed of one component common to all the orientations of the face and a second component that depended on the view of the face. Due to the columnar organization of TE, many TE neurons in a column may represent a complex feature and slight changes in features may be the result of differences of neurons with similar but subtle changes in selectivity in those columns (Tanaka, 1996).

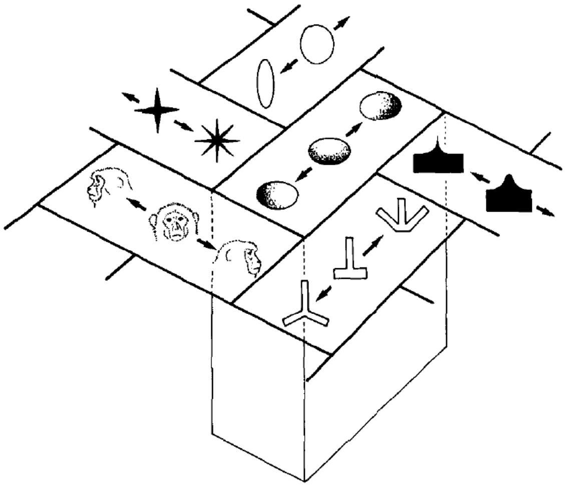

Figure 2.8. Columnar organization in IT *(Source: Tanaka, 1996)*

The feature-based model is consistent with several studies in monkeys (Desimone et al., 1984; Tanaka et al., 1991; Kobatake and Tanaka, 1994) and fMRI studies in humans. Tanaka (1996) found that neurons in IT responded under complex shapes, and other neurons responded to the combination of such shapes with colour or texture. Recently, Tsunoda et al. (2001) proposed that an object corresponded to a combination of different active and inactive feature columns, supporting and extending the feature-based representation. They found in their monkey neuron recordings the existence of inhibitory mechanisms inducing feature columns to be active or inactive.

3 Attention

The visual system can only analyze a limited amount of information that impinges in the retina (Tsotsos, 1988; Tsotsos, 1990; Tsotsos, 1992). The filtering of irrelevant information is performed by attentional mechanisms.

Attention is normally directed where we foveate. To bring attended objects or locations to the fovea, attention usually interacts with eye movements, although attention can also be activated without involving eye movements (Bushnell et al., 1981; Hoffman, 1998).

Some authors argued that attention is a mainly bottom-up process (Koch and Ullman, 1985; Itti and Koch, 1998), but others proposed that attention may operate in two ways (Tsotsos et al., 1995; Desimone and Duncan, 1995): In a first step, a top-down bias primes the neurons that encode the object to attend, and a second step comprises a competition in which the primed neurons inhibit the effect of the rest of neurons.

In these theories, there is cooperation among different neurons that respond to the same stimuli, while the ones that respond to different stimuli compete. The final response will depend in the strength of the interactions of neurons than respond to the same stimuli, and the strength of the neurons competing against them (Duncan et al., 1997).

There is no specific location for attention in the brain. Attention can be found in every area of the visual cortex and other areas of the frontal and parietal cortex (Bushnell et al., 1981; Kastner and Ungerleider, 2000), as Duncan et al. (1997) states "Selection of objects for the control of action arises through cooperative and competitive activity across multiple brain systems".

Section 3.1. will present evidence for attention from neurophysiological and fMRI studies. Section 3.2. deals with another very important evidence for attention: visual search. Section 3.3 presents most of the models of attention.

3.1 Neurobiological evidence for attention

Evidence for attention in physiology was first found in the Superior Colliculus (Goldberg and Wurtz, 1972; Wurtz and Mohler, 1976), and later in other parts such as the posterior parietal cortex (Bushnell et al., 1981), V2 (Wurtz and Mohler, 1976; Motter, 1993; Luck et al., 1997), V4 (Fischer and Boch, 1981, Moran and Desimone, 1985; Motter, 1993; Luck et al., 1997; Reynolds et al., 2000) and IT (Moran and Desimone, 1985, Chelazzi et al., 1993; Chelazzi et al., 1998). While several studies failed to find attention effects in V1 (Moran and Desimone, 1985; Luck et al, 1997), others reported some modulation in this visual area (Motter, 1993; Press et al., 1994). The huge amount of feedback connections existing in the cerebral cortex would also support the idea of visual attention from a neurophysiological perspective (Moran and Desimone, 1985; Felleman and van Essen, 1991).

Bushnell et al. (1981) found higher neuron responses to attended stimuli versus unattended stimuli in the posterior parietal. They also found shifts of attention without involving eye movements.

Moran and Desimone (1985) recorded neurons from monkeys while performing several tasks involving attention. They found that in V4 when attention was directed to the ineffective stimulus, the neuron's response was greatly attenuated compared to when attention was devoted to the effective stimulus, although both, effective and ineffective stimulus were in the neuron's receptive field. They failed to find differences in V1 neurons' response when attending to the effective stimulus vs the ineffective stimulus.

Motter (1993) showed by means of a series of neurophysiological experiments that attention was present at early stages such as V1 and V2. He found that 1/3 of the neurons in area V1, V2 and V4 had an increase in the response in the presence of focal attention. From these neurons, ~70% of the neurons only raised their response when there was more than one stimuli, supporting some competition between stimuli.

In contrast Luck et al. (1997) found no effects of attention in V1, and in V2 and V4 only when the competing stimuli and the target were inside the receptive field of the neuron. An important finding of this study was that attending to a location when expecting a stimulus in the absence of visual stimuli increased the neural baseline activity.

Chelazzi et al. (1993, 1998) found in single unit recording experiments with macaque monkeys that the inferior temporal cortex (IT) is implicated in filtering the objects to which we attend and foveate in visual search tasks. They demonstrated that this filtering results from competition between neurons in the IT cortex. Their experiments consisted of a sequence of stimuli in which they first showed an object that activated or inhibited the recording neuron at the fixation point. After a delay of 1500 ms, two objects appeared, one of which could be the one shown previously. In this case the monkey had to make a saccade to this object. An interesting finding from these experiments is that in the delay period, neurons were at some level of activation after having beein shown the stimulus to which they were selective. This may support a top-down bias that favors the relevant stimulus and thus several of the leading models, Tsotsos' Selective Tuning Model (Tsotsos et al., 1995), the Biased Competition Model (Duncan and Desimone, 1995) and Ferrera and Lisberger's (1995) model. In this way, a neuron is inhibited through competition when an ineffective stimulus is attended, even in the presence of an unattended effective stimulus in the neuron's receptive field. This competition is directed by a top-down bias from working memory, where the information about objects is stored.

From these studies, there seem to be some discrepancies about how attention works in single unit recordings. While some studies showed that the effects of attention appear only when there was simultaneously an attended and ignored stimulus inside the neuron's receptive field (Moran and Desimone, 1985; Luck et al., 1997), some other studies suggested that there is also attention even if only one stimulus was inside the receptive field of the neuron (Motter, 1993). The three studies used the same visual areas (V1, V2 and V4). The answer to this difference may be that there is modulation when one stimulus is inside the receptive field and the other is outside the RF, but this modulation appears to be larger when both stimuli are inside the RF (Chelazzi and Desimone, 1994 in V4; Treue and Maunsell, 1996 in MST; Luck et al., 1997 in V2 and V4). Also, the effects of attention in the case of inside/outside stimuli seemed to be stronger when the stimuli are difficult to discrimate (Luck et al., 1997). The effects of attention were directly related to the neuron's receptive field size (Kastner et al., 2001). That is, lower effects were found in V1 neurons than in V4 neurons.

With respect to the discussion about if there is attention in area V1, more recent studies with fMRI (functional magnetic resonance imaging) in humans found the effects of attention in the human visual areas V1, V2, V4, IT (see Kastner and Ungerleider, 2000 and Pessoa et al., 2003 for a review) and even as early as LGN (O'Connor et al., 2002). Also, with fMRI the modulation of attention was found to appear in V1 with a delay of 150-250 ms (Martinez et al., 1999), this may explain why other studies using single unit recording failed to find such modulation (Moran and Desimone, 1985; Luck et al., 1997) in this area.

Objects' attentional selection may be accomplished through the inhibition of the neurons' responses to irrelevant stimuli (Moran and Desimone, 1985, Tipper, 1985; Chelazzi et al., 1993; Chelazzi et al., 1998; Kastner and Ungerleider, 2001; O'Connor et al., 2002).

3.2 Visual search

Another way to study attention is by means of psychophysics. Most of the experiments in psychophysics to analyze the mechanisms of attention involve visual search tasks.

We are constantly performing visual search tasks in our day-to-day activities, e.g. when we are looking for a blue shirt among other shirts in our wardrobe, the blue shirt is the "target" shirt, we call the other shirts "distractors".

Pashler (Pashler, 1998) defined visual search as "the task of attempting to find a specified target or targets in a visual display". A typical visual search task consists of an observer looking for a target in a visual display, the observer responds "yes" if the target is present and "no" if the target is absent.

In a laboratory environment, an example of a visual search task is to find an item of a specific colour among items of another colour, or finding an "X" or a "T" among "L"s (Figure 3.1a). Two measures are usually used in visual search: reaction time (RT) and accuracy. When RT is the measure of interest, the visual stimulus is present in the display until the subject decides whether or not the target is present. RT is usually plotted as a function of set size (the total number of items in the display). Harder search tasks require longer RTs (e.g. as in Figure 3.1a, it is easier to find "X" among "L"s than to find the "T"). When accuracy is being measured, the stimulus is only presented briefly, followed by a mask. In this case, accuracy is plotted as a function of the stimulus onset asynchrony between the stimulus and the mask. Here, difficult search tasks would require longer stimulus onset asynchronies.

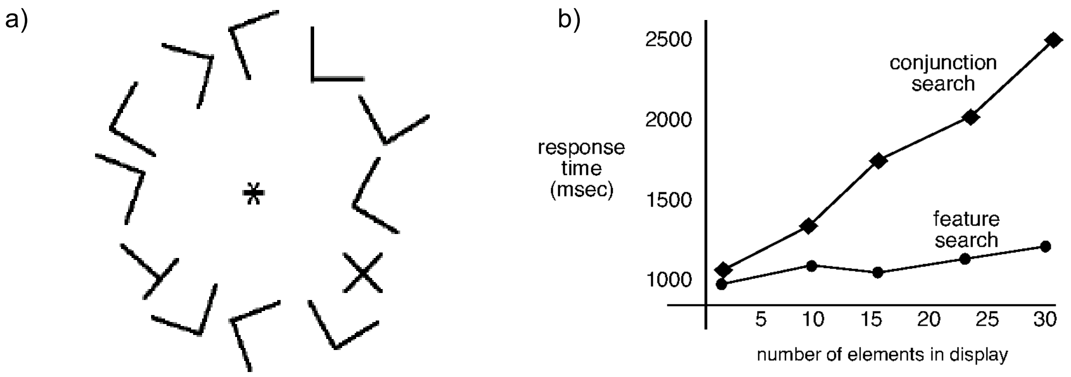

Figure 3.1. a) Visual search task *(Source: Wolfe, 1998)*
b) RTs for conjunction and feature searches

When we perform visual search tasks and we measure RT vs set size we usually observe two extreme cases (Figure 3.1b): *conjunction search* and *feature search*. In feature searches, (Figure 3.2 left) the target is defined by a single feature, and RT vs set size slopes are near zero msec/item. In conjunction searches (Figure 3.2 right) the target is defined by the conjunction of two different features, and slopes are greater than zero msec/item (e.g. 20-30 msec/item in the case of searching an S among mirror-Ss or L among Ts).

3.2.1 Studies in Visual Search

3.2.1.1 The Feature Integration Theory (FIT) (1980, 1990)

The Feature Integration Theory (FIT), proposed by Treisman & Gelade (1980) held that in feature search the array was searched in parallel while in conjunction search the array was searched serially. This is because the targets in feature searches would be identified without locating them, they did not need attention and could be identified in parallel. Conjunction searches needed attention to be focused in order to combine the objects' features to form the complete object, therefore attention needed to focus on every object serially to find the target.

Treisman and Sato (1990) modified the original theory to account for new findings in visual search that contradicted its predictions (Nakayama and Silverman, 1986; Pashler, 1987; Wolfe et al., 1989; Duncan and Humphreys, 1989). They conceded that "conjunctions for highly discriminable features could be rapid or even parallel".

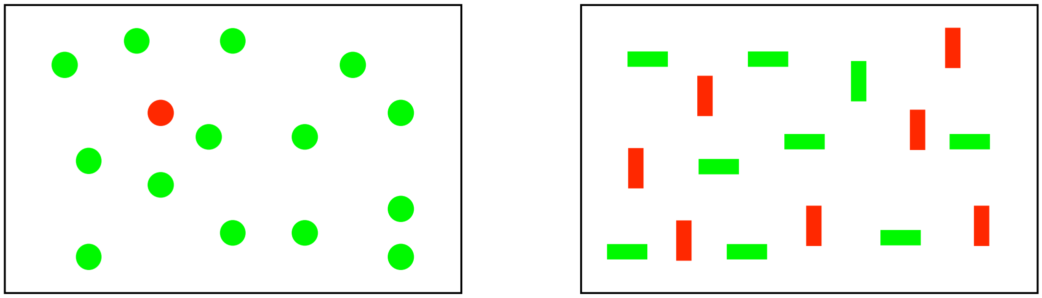

Figure 3.2. (left) In feature searches, a target is defined by a single feature: here a red circle is set against green circles. The target seems to pop out. **(right)** In conjunction searches, the target object is defined by a conjunction of two features: here, the two features are orientation and colour; The target is a vertical green bar and the distractors are both green and horizontal and red and vertical bars

They tested three possible strategies for visual search: (1) Spatial selection according to FIT, (2) the temporal coincidence hypothesis, that states that objects features are put together when their neural onsets coincide in time; and, (3) that conjunctions are processed by specialized detectors. They explained their results by postulating that serial search was not performed item by item, but rather in groups. That is to say, subjects consider groups of objects serially, supporting partially their FIT. Distractor locations were inhibited by feature inhibition.

3.2.1.2 Criticism of FIT

Treisman and Gelade (1980) performed a series of visual search experiments that seemed to confirm the predictions of their model. But, since its inception, the FIT has been controversial. It has been modified several times by their authors (Treisman & Sato, 1990; Treisman, 1993) and dismissed by other models (Duncan & Humphreys, 1989; Wolfe et al., 1989; Wolfe, 1994; Grossberg et al., 1994).

Atkinson et al. (1969) were the first to criticize the notion of a dichotomy between parallel and serial search before the FIT was proposed. They argued that the results obtained from a serial search task may also be explained by models based on limited capacity parallel search (Kinchla, 1974; Ratcliff, 1978). Nakayama and Silverman (1986) reported that when visual search involves a conjunction of motion and stereoscopic depth the slopes RT vs set size were near zero. In a series of experiments, Wolfe et al. (1989) found that searches in triple conjunctions were easier than for simple conjunctions. The FIT on the contrary, predicts the same or steeper slopes for triple conjunctions. Recent experiments have shown (McElree & Carrasco, 1998) that conjunction searches can also be performed in parallel.

Wolfe (1998) explained four reasons why this serial/parallel dichotomy should be rejected, these are summarized as follows:

- The inference that the results using RT vs set size (Figure 3.1) are due to parallel and serial mechanisms is incorrect. As already mentioned, it can be due to other mechanisms such as a limited capacity parallel search.
- Serial search is based on a number of unfounded assumptions. The first assumption is that in trials involving a single target, on average half of the items

are searched, but in target absent trials, all the items are searched. This implies no items being checked twice (Horowitz & Wolfe, 1997) and that there are no misses. The second assumption is that serial search presumes that only one item at a time is analyzed, evidence exists to show the opposite (Gilmore, 1985, Grossberg et al., 1994).

- The time that a focus of attention is spent at a location (dwell time) in serial search is considered fixed and is in a range of 25-50 ms/item. Duncan et al (1994) found dwell times of several hundred ms, supporting parallel models of visual search.
- Over the past 15 years, experiments have shown a continuum of slopes (Figure 3.3) for visual search tasks. Duncan and Humphreys (1989) stated that visual search is harder when target and distractors are more similar, but is easier when this similarity decreases.

Although the FIT has been criticized, Treisman and colleages were not completely wrong. Recent studies have shown that in very demanding visual search tasks, search is strictly serial (Horowitz and Wolfe, 1998; Horowitz and Wolfe, 2001; Woodman and Luck, 2003).

Some alternative theories have been proposed to the FIT in order to explain the results from visual search that contradicted the FIT.

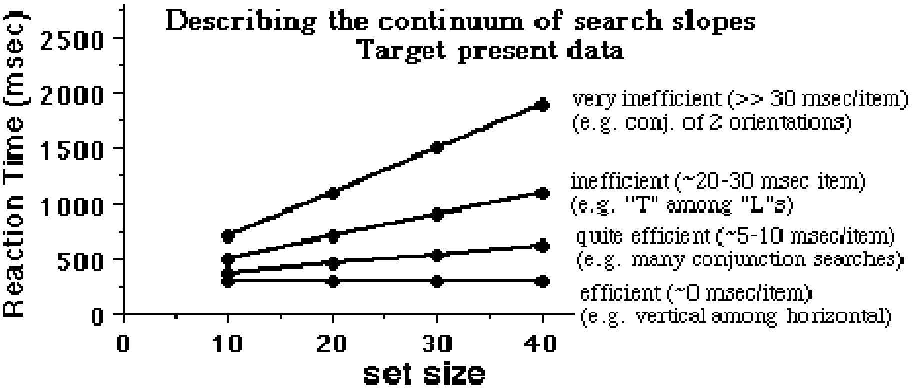

Figure 3.3. Visual search continuum. From efficient to very inefficient. (*Source: Wolfe, 1998)*

3.2.1.3 Pashler (1987)

Pashler (1987) proposed a parallel self-terminating process for visual search. He showed that parallel searches occurred in groups of up to eight items. When there were more that eight items, subjects analyzed the visual field in sequences of groups containing up to eight items. He also pointed out that the strategy used in conjunction searches was for items having common features to the target.

3.2.1.4 Attentional Engagement Theory (1987, 1989, 1992)

Duncan and Humphreys' (1989) theory was composed of three components: (1) A perceptual input description (colour, size, etc) that was accomplished in parallel, (2) selection, performed by comparing the input descriptions with an internal template, and (3) a virtual short-term memory that saved the selected information. The visual input for Duncan and Humphreys (1989) had a hierarchical organization in the same way as Barrow and Tenenbaum (1978). In this hierarchy, the *structural units* segmented from the visual field (similar to the object files of Kahneman and Tresiman, 1984 and the 3-D model of Marr and Nishihara, 1978) had a set of qualities, such as relative location, motion, colour, texture, size and shape.

Quinlan and Humphreys (1987) carried out a series of experiments with feature search, conjunction search and *both search* (two features separated in the visual field) with two and three features. From their results, they argued that visual search is determined by the *fidelity* of the information needed for discrimination, that is, how good (in terms of difference) is the information in the display for discrimination.

A visual display was decomposed at different levels. The top level was decomposed into regions and these regions were decomposed into smaller regions at each successive layer until there were only individual items at the bottom. For conjunction searches it is necessary to search among the individual items, while for feature searches, the target is identified at higher levels. This approach also accounted for by the fact that the difficulty of feature discriminations increased at higher similarities, and evidenced a clustering when performing visual search (Quinlan and Humphreys, 1987; Humphreys et al., 1989). Attention selected among these different levels of visual description the highest possible level for which the task is performed fastest.

They proposed that the serial/parallel visual search dichotomy does not occur, but, that there is a continuum in search efficiency. They also proposed that target-distractor and distractor-distractor differences play an important role (Duncan and Humphreys, 1989). Search is mainly parallel in groups of elements (Humphreys et al., 1989). They supported this statement with Pashler's (1987) experiments. Humphreys et al (1989) showed that items in a visual search task are usually clustered in a familiar shape and they stated that "visual search cannot be understood independently of the processing mediating grouping and segmentation".

According to Humphreys et al. (1989), combined-form information accounted for discrimination and the grouping of this information accounted for performance (Humphreys et al., 1989). When the homogeneity of distractors is high, the group of distractors forms a single *object description* and search is fast. Adding distractors having similar features to the target affected the grouping of distractors. These distractors were considered in the group of possible targets, and for such, search was slower and set size was important.

Duncan and Humphreys (1989) found that when the target was very dissimilar to the distractors, search was always fast (independently of the similarity or dissimilarity among distractors). When target and distractors are similar, the heterogeneity in distractors became important for visual search. The worst case (slowest search) was where the target was similar to the distractors and these distractors were heterogeneous, in this case grouping is difficult (the distractors were not similar among them, and they were similar to the target), a search that was close to a serial search applied.

As a summary, visual search is fast under the following conditions:

- Similarity is high among distractors (Duncan and Humphtreys, 1989; Humphreys et al., 1989).
- High dissimilarity between target and distractors (Duncan and Humphreys, 1989).
- Small ratio of eccentricity/stimulus size (Humphreys et al., 1989).

According to Humphreys et al. (1989), this was an indication of grouping of objects in the visual field.

3.2.1.5 Guided Search (1989)

Wolfe et al. (1989) presented an alternative to the FIT. Their model was based in Hoffman (1979) in which a serial stage preceded parallel process.

In Guided Search, a top-down process influenced parallel processes for filtering irrelevant information in a visual search task. The parallel stage directed attention to the possible targets, dividing the visual field into distractors and candidate targets. They argued that conjunctive searches do not have a slope as do feature searches due to noise in the connection between the parallel and serial stages.

3.2.1.6 Grossberg et al. (1994)

The authors proposed an alternative to Treisman's FIT (1980) and Wolfe's Guided Search (1989) for visual search.

In their model, an attentive visual object recognition system interacted with visual search. The object recognition system was a combination of form, colour and depth information in an ART network. In ART networks, neurons can be activated by bottom-up processes and a top-down process primes neurons.

As with Quinlan and Humphreys (1987) they used evidence that visual search tasks are facilitated by clustering (Quinlan and Humphreys, 1987; Humphreys et al., 1989). Items were evaluated simultaneously in clusters by a mechanism involving attention in those groups of items (Chelazzi et al. 1993). The algorithm proposed involved several steps that were repeated until a match is found:

1. Perform a preattentive analysis of the scene and recording of features
2. Group items in selected regions, based on clusters of features. A top-down process can prime some regions over the others.
3. Select a candidate region. Here the bottom-up and top-down processes influences the selection.
4. Compare with the target.
    a. If there is a complete mismatch of the features then return to step 3 to select another region.
    b. If there is a partial mismatch, return to step 2 for selecting a subregion inside the selected region

The authors performed several visual search tasks (form-colour conjunctive search, colour-colour conjunctive search, multi-target and single-target conjunctive search and triple conjunctive search). They compared the results of applying their algorithm with the psychological results from Treisman and Gelade (1980), Wolfe et al. (1989), Morkoff et al. (1990) and Treisman and Sato (1990) regarding response times (RT) and slopes.

Responses using their approach were very similar to the results obtained by previous psychophysical experiments (Treisman and Gelade, 1980; Wolfe et al., 1989, Morkoff et al., 1990; Treisman and Sato, 1990). For the calculation of response times, they used two linear equations for target-present search and target-absent search, both of them had four parameters (durations of step 1, steps 2 and 3, step 4 and number of candidate regions). The target-absent equation search was similar to the target-present equation with the difference that the number of candidate regions was multiplied by two. It is not surprising that their results had a relation in RT slopes 2:1 for the target-present and target-absent conditions (as has been shown across the visual search literature). Also, for the different tasks they needed to adjust parameters in order to obtain good results.

3.2.2 Basic features in Visual Search

3.2.2.1 Colour

Colour has always been used as one of the basic features in visual sarch experiments in theories (Treisman and Gelade, 1980; Quinlan and Humphreys, 1987; Humphreys et al., 1989; Wolfe et al., 1989). Efficient search can be performed with heterogeneous colours (up to nine distractors) as soon as they are widely separated in colour space (Wolfe, 1998). Treisman and Gornicam (1988) found asymmetries in search: to search for a target magenta among red distractors was easier than finding a red target among magenta distractors. They argued that it is easier to find a variation of red than a red target among variations of red. Since magenta contains blue, an explanation is that the target contains blue among distractors that do not contain this colour. While, finding a red item among magenta distractors implies to finding the reddest item.

V4 seems to be the main area where colour is processed (Van Essen and Zeki, 1978). But, there are other areas where the processing of colour occurs, namely V1, V2 and IT. In V1 and V2, wavelength information is processed, although some studies suggested that

V1 and V2 do more than that, such as correcting changes in luminance due to variations in chromaticity (Yoshioka et al., 1999). V4 may be a second stage, where colour constancy is computed, the final step is IT where there may be an association of colour with form (Zeki et al., 1999).

3.2.2.2 Orientation

Orientation is another broadly accepted basic feature. Usually, V1 is considered to process orientation due to the fact that it processes bars and edges (Hubel and Wiesel, 1959, 1962, 1968).

Subjects are capable of discriminating lines as little as 1° or 2° difference in orientation. Although, for efficient visual search, a minimum of 15° is required. When distractors are heterogeneous, search becomes inefficient (Wolfe, 1998). Search asymmetries were also found in orientation: it is easier to find a vertical target among distractors that are tilted 20° than to find a 20° tilted target among vertical distractors (Wolfe et al., 1992). In the same study, Wolfe et al. (1992) found that search is efficient even with heterogeneous distractors when the target is the only vertical, horizontal, left- or right-tilted element.

3.2.2.3 Spatial frequency

Spatial frequency seems to be related to size and scale (although there are no experiments that confirm this association). Spatial frequency is related to cycles per degree gratings. Size visual search finds the element that has a unique size. If the size difference is enough, then a target specified by one size can be found quite efficiently among distractors of another size (Treisman and Gelade, 1980; Quinlan and Humphreys, 1987; Duncan and Humphreys, 1992). Treisman and Gornicam (1988) found another asymmetry in size: it was harder to find a small target among big distractors than a big target among small distractors.

3.2.2.4 Motion

Another widely accepted basic feature is motion. It is easy to find a moving target among static distractors. Spatiotemporal features have been demonstrated to be the first features present in humans for recognizing objects (more than colour and orientation) (Xu and Carey, 1996).

3.2.2.5 Other features

Although the preceding features are the ones found in most studies of visual search, letters appear also to be a basic feature in visual search (Treisman and Gelade, 1980; Quinlan and Humphtreys, 1987; Humphreys et al., 1989). The idea that letters are basic features was supported by several studies Malikowski and Hubner (2001). Junctions (as in letters) can be combined hierarchically (Humphreys et al., 1989).

Other studies showed that basic features are not necessarily low-level features (e.g. letters). Levin and Takarae (2001) performed a series of experiments where subjects quickly found targets in heterogeneous and complex categories. One of these features may be shape. Shape has been controversial in the literature. Some studies show that shape can be a basic feature with no possibility of reducing it to orientation and curvarture (Wolfe, 1998).

Other features not considered here are curvature and depth (Wolfe, 1998). Letters, depth and shape may be related with object-based attention, or they may be considered high-level features. For this reason we should not consider them as basic.

3.3 Models of Attention

Here I present the most influential models of attention following the historical evolution of those models. From the evolution of the models of attention, we can extract that they seem to converge in having a hierarchical neural network, competition among units (usually winner-take-all) where the stronger units inhibit the response of the weaker ones, and a gating process for selecting the focus of attention.

Broadbent (1958) was the first to propose a filtering of the information that flows to the brain. He suggested that the brain's limited capacity influences the processing of sensory information. There is a selection of the incoming data on a basis of physical characteristics (*early selection*). In Broadbent's theory, after short term memory, there was a filtering process in which the information not selected by the filter was simply excluded (Broadbent, 1958) or attenuated (Broadbent, 1971, 1982).

There is evidence that selection is performed not only on the basis of physical characteristics. Information may also be selected on the basis of its semantic content. Deutsch & Deutsch (1963) considered that the limitation is not in the sensory system, but in the response system (*late selection*). See Figure 3.4a for a comparison between Early selection and Late selection models.

Later, Treisman (1964) modified the first Filter Theory presented by Broadbent (1958) where messages were not entirely filtered, but attenuated. Semantic selection criteria could be applied to all of the messages.

For Milner (1974), attention was accomplished by means of feedback pathways and synchronized firing was used by attention to select relevant elements.

Treisman and Gelade (1980) and Treisman and Sato (1990) FIT (Feature Integration Theory) was a model that worked in two stages (Figure 3.4b): In a first preattentive stage features were registered in parallel and objects were recognized in a second and serial stage that required focused attention to integrate objects' separate features (colour, orientation, spatial frequency, movement, brightness). In their theory, the role of attention was to combine features into objects, without which features couldn't be associated with each other. They considered that parallel locations were essential in detection. Perceptual grouping controlled a later stage that involved attention. In the Feature Integration Theory, attention to spatial locations associated the corresponding features with an object (binding problem).

Following Milner (1974), von der Malsburg (1984) proposed that timing correlations bound features to the same object in order to recognize an object by its conjunction of features.

Koch and Ulman (1985) proposed a system with different base features (color, orientation, direction of movement, etc). Each feature was represented in a topographical map, and the set of these maps formed the *early representation.* From the early representation, the *central representation* selected a location in the visual scene. Koch and Ulman used a Winner Take-All (WTA) to select these locations. When the selected location was inhibited, a shift towards the next most salient location was performed.

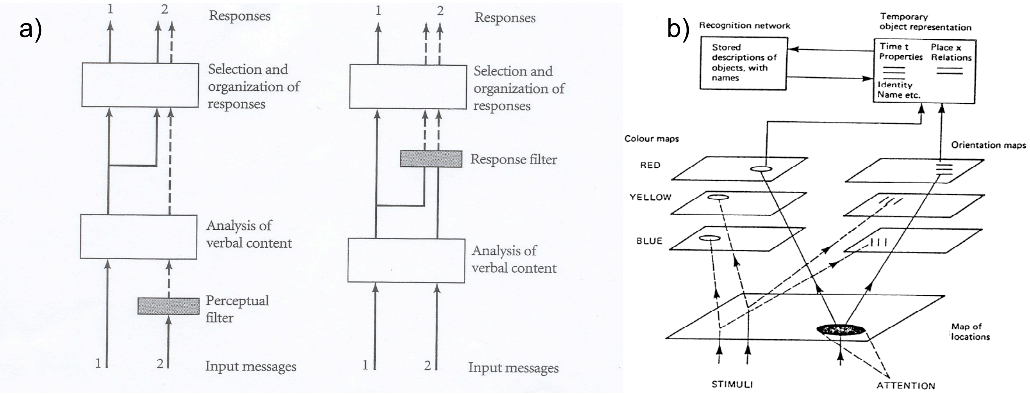

Figure 3.4. a) Early selection vs Late selection
b) Feature Integration Theory (FIT) *(Source: Treisman and Gelade, 1980)*

Koch and Ullman (1985), as previous models, did not test their model, but it was later implemented and tested by means of a computational system (Itti et al., 1998) and was a basis for others (Olshausen et al., 1993).

Guided Search (Wolfe, 1989) was presented as an alternative to the Feature Integration Theory (Treisman and Gelade, 1980). A top-down process influenced parallel processes for filtering irrelevant information in a visual search task. The parallel stage directed attention to the possible targets dividing the visual field into distractors and candidate targets.

Sandon (1990) proposed a hierarchical multi-scaled network with a module of attention which used lines or edges for a low-level attention mechanism and corners for a high-level attention mechanism. Attention guided features from a region to higher layers of the network, where object recognition was performed. The module of attention performed a winner-take-all among the values of the features.

SLAM (Phaf et al., 1990) followed McClelland and Rumelhart (1981) in the construction a model of attention that recognized words. Their model was a hierarchy with different levels from bottom to top: feature level, letter level, word level and higher levels that delivered top-down selection to the word level. Features, letters and words were represented by nodes that accounted for the different possible combinations among features, letters and words (e.g. form-position, colour-position and form-colour for features). Both bottom-up and top-down connections were included. Inter-level connections were inhibitory, while between level connections were excitatory or inhibitory.

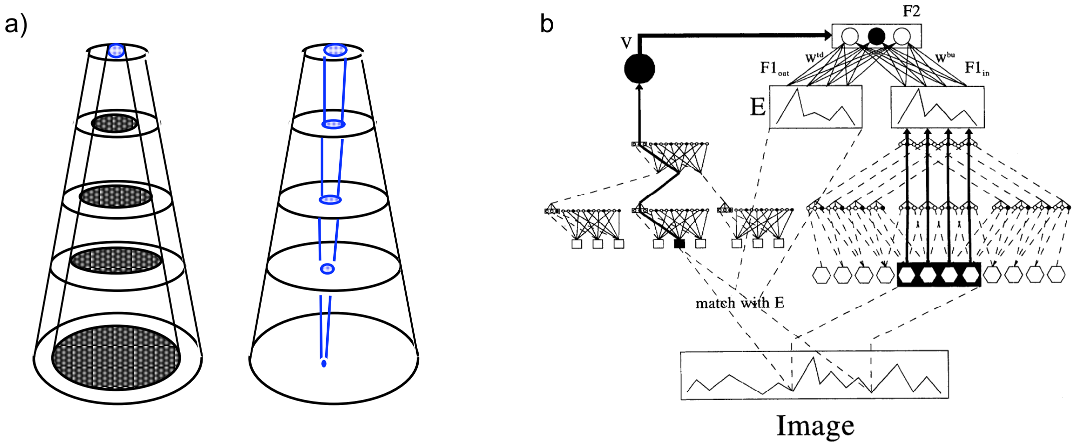

Figure 3.5. a) The selective tuning model, 1st bottom-up information flow, 2nd top-down
b) SCAN with the gating and classifier network *(Source: Postma, 1997)*

The Selective Tuning Model (Tsotsos, 1990; Tsotsos, 1993; Tsotsos et al., 1995) was a hierarchical system with bottom-up and top-down attention (Figure 3.5a). Top-down attention was performed by inhibiting the response of neurons at lower levels through gates at every level. The model worked in two main steps: a feedforward process and a top-down process that selected the location of interest by means of a hierarchy of winner-take-all at every level. The winner-take-all proposed by Tsotsos is a variation from Koch and Ullman's (1985) winner-take-all in order to assure convergence. Selection was accomplished by gating networks at each level of the hierarchy.

MORSEL (Mozer, 1991; Mozer and Sitton, 1998) contained a hierarchical feedforward network that transformed features operating for simple patterns and location dependent into features which responded to complex patterns and were location-invariant. Layer neurons had the same receptive field size. Some overlapped, but others didn't and went from small number of features at low-level to higher number of features at higher levels. The model incorporated an attentional network that gave rise to an attentional map with attentional units which permitted features corresponding to a location to be connected to the network. Unattended locations (the ones not selected by the attentional units) were inhibited. The attentional network had two sources of input: exogenous (from sensory data) and endogenous (from learning, priming or cueing).

Attentional units competed one against the other in a winner-take-all fashion. Mozer tested his model in letter recognition and visual search.

VISIT (Ahmand, 1991) was composed of three networks and working memory (Figure 3.6 left). A *gating network* reduced the activity of other regions except the attended one, this attended region was selected by a *priority network* which performed bottom-up or top-down attention. These two networks communicated by means of a *control network*. The working memory saved temporal relevant information. Ahmad applied several modifications to VISIT in order to account for visual search (Figure 3.6 right): A group of feature maps were used for basic features (orientation, colour, etc.), each one communicating to the gating network and the priority network. The gated feature maps only passed the features that was present in their gate units. Then an OR operator was computed between them to check for feature combinations. The priority network performed a classification in order of relevance.

Niebur et al. (1993), Niebur and Koch (1994) and Usher et al. (1996) proposed that a neural model of attention was based in first instance on oscillations, and later on the temporal correlation for a few ms among sets of neurons. Their network was composed of three layers (with a possibility of adding new layers): V1, V2 and V4. V2 projected in V4 and excited pyramidal neurons and/or constrained the action of interneurons that inhibited pyramidal neurons. A time-dependent Poisson process was used for simulating V2 neuron pulses with a mean rate λ that was the addition of λ_{spont} (spontaneous firing rate) and λ_0 (the stimulus-dependent rate) modulated by the Poisson distribution over time:

$$\lambda(t) = \lambda_0 P(t) + \lambda_{spont}$$

$$\lambda_0 = \lambda_{\max} \times overlap(stimulus, receptivefield) \qquad \lambda_{\max} = 200Hz$$

In V4, a competition within microcolumns of neurons was performed, and neuron responses were simulated over time.

Grossberg’s (Grossberg et al., 1994; Grossberg et al, 1998; Grossberg, 1998) ART (Adaptive Resonance Theory) model had bottom-up activation and top-down priming that modulated a neuron. To be active, a neuron had to receive a minimum amount of bottom-up activation and top-down priming. Top-down priming had a stronger effect over the neuron and could even inhibit bottom-up activation. Grossberg (1998) constructed his model with LGN, V1 and V2 layers containing inhibitory and excitatory neurons

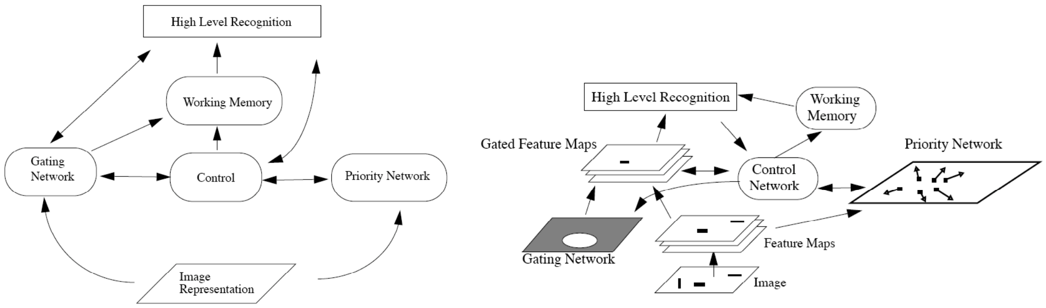

Figure 3.6. Architecture of VISIT (left) and VISIT for visual search (right) *(Source: Ahmad, 1991)*

The Biased Competition model (Desimone and Duncan, 1995) was similar to the Selective Tuning Model (Tsotsos, 1995). Attention was performed by inhibition through a bias that favors some stimulus over others incorporating a top-down winner-take-all strategy.

VAM (Schneider, 1995) follows von der Marlsburg (1981) and constructed two different path for attention: a what path and a where path. The where path was oriented to select locations that shared features with the target, and the what path was concerned with differences among elements.

SCAN (Signal Channeling Attentional Network, Figure 3.5 b) (Postma, 1997) was a model composed of two main components: a hierarchical gating network and a classifier network. The input of the gating network was the image, a subimage (*input pattern)* was selected in an attentional way to an *output pattern*. The gating network was composed of a data part connected to the input image and a control part that determined how strong attention was to the subimages.

The classifier network consisted of two layers: an input/output layer that received the output pattern from the gating network and a category layer. The category layer was based on a WTA strategy and generated an *expectation pattern*. A threshold element determined if there was a match between the expectation pattern and the input. In case of mismatch, a new generation pattern was generated until a match was found (Figure 3.2b).

Olshausen et al. (1998) presented a model based on Anderson and van Essen (1987) that was position and scale invariant and tried to perform a transformation from the retinal reference frame to an object-centred frame (Figure 3.7a). To accomplish this, they used *shifting circuits* and *control neurons*. The control neurons dynamically conducted information from lower levels of a hierarchical network to higher levels of the network. By means of the shifting circuits and the control neurons, the window of attention changed in size (scale invariant) and position. This model was extended in the similar SIAM (Heinke and Humphreys, 2003).

Itti et al. (1998) followed Koch and Ullman (1985) to construct a purely bottom-up model of attention (Figure 3.4b) in which different feature maps for colour, intensity and orientation in a Gaussian pyramid give rise to a saliency map. To account for the lack of top-down attention, they used a normalization operator to provide saliency maps for each feature (colour, intensity and orientation). These three saliency maps were combined in a general saliency map. A winner take all selected the most relevant region based in the values from this saliency map.

Feature Gate (Cave, 1999) was a hierarchical neural network in which feature gates at each location accounted for attention controlled by competition among locations favored by a bottom-up and top-down system.

Deco and Zhil (2001) and Deco and Rolls (2004) showed a model where features were represented by a population of excitatory neurons (pyramidal neurons) at every location. Inhibition was performed by a population of inhibiting neurons (V4). A high-level map integrated the different features at each location (IT), and top-down was provided by a memory module. They tested their model in visual search tasks with bars and colours.

Lee et al. (2004) constructed a model composed of three modules: a bottom-up module, a top-down module and an integration module. The bottom-up module extracted features as colour, aspect ration, symmetry and shape. The top-down module directed the finding of targets by means of the target features. The integration module integrated inputs from both modules through what they call an Interactive Spike Neural Network (ISSN) that is reminiscent of the spike synchronization from physiology. They tested their system with success to find people with different characteristics. For that, the bottom-up module had the features corresponding to faces (skin color, ellipse shape, …) and the top-down module had features corresponding to a specific feature (e.g. the color of the shirt). The model used spike synchronization between the bottom-up and top-down modules to find the target, this worked quite efficiently in their test cases (people), but more and more studies in physiology are discarding this way of performing attention, and their results could be accomplished also with a winner-take-all with inhibition of return.

All of the models of attention have a hierarchical structure in different levels. In some of them attention was deployed in a bottom-up fashion, in others was top-down. Koch and Ullman (1985) bottom-up model has been very influential and followed by other models such as Itti et al., 1998. Bottom-up models attend to locations based on a *saliency map.* This saliency map usually incorporates responses from maps of features (colors, orientation, luminance, etc.). Attention is then deployed to the strongest feature values. This has been shown successful for attending objects that pop-out. But, bottom-up attention models fail when attention is task directed, that is, look for an object with specific features as in visual search. A second group of models would be the oscillation models. there are a number of models whose base is oscillations following Milner (1974) and von der Marlsburg (1984) (Niebur et al., 1993; Niebur and Koch, 1994; Schneider, 1995; Usher et al., 1996; Lee et al., 2004), although, a neuron's synchronization for attention is usually discarded by physiologists,. Lately, since the Tuning Selective Model (Tsotsos et al., 1995) top-down models of attention are gaining supporters. Evidence from neurophysiological studies (Fischer and Boch, 1981; Moran and Desimone, 1985; Motter, 1993; Chelazzi et al., 1993; Luck et al., 1997) recent and fMRI studies (Kastner & Ungerleider, 2000; Kastner et al., 2001; O'Connor et al., 2002) confirm this modulation. Most of the bottom-up and top-down models coincide in a winner-take-all strategy for attention. A large number of models have an abstract base, being more a theory than a model (Broadbent, 1958, 1971, 1982; Deutsch & Deutsch, 1963; Treisman & Gelade, 1980; von der Malsburb, 1984; Koch and Ullman, 1985; Wolfe, 1989; Phaf et al., 1990 Ahmad, 1991; Niebur et al., 1993; Grossberg, 1994; Desimone and Duncan, 1995; Schneider, 1995). Only some of them have a strong computational base, and even less have been tested computationally (Sandon, 1990; Olshausen e al., 1993; Tsotsos et al., 1995; Postma, 1997; Itti and Koch, 1998; Mozer and Sitton, 1998; Deco and Zhil, 2001; Heinke and Humphreys, 2003; Lee et al., 2004). Only a small amount of models have been tested in the recognition of basic objects or visual search tasks (Grossberg et al., 1994 and Deco and Zhil, 2001). During the last decade models of attention are getting more and more similar. A number of laboratories have constructed their model in a computer and have tested them for simple visual tasks. Results are promising and efficient. They do not need to compute complex structures of indices as the current object recognition systems in computer science (see chapter 6). They only extract the basic features thought to be used by humans such as edges, colours, intersections, etc.

3.4 Object-based attention

The models addressed previously usually consider the focus of attention in spatial terms. It also has been proposed that attention selects discrete objects and that limits of attention are imposed by the number of objects that can be attended simultaneously.

Posner et al. (1980) considered that attention was centred on a location and was decreased away from its centre. In a series of experiments he found that responses were faster when showing a target in a valid cue location than when the target appeared elsewhere. He proposed the *spotlight* model in which attention was directed to a location instead of to objects. But, Neisser (Neisser, 1967, 1979; Neisser and Beklem, 1975) had showed that superimposing images (Figure 3.8sa) at the same location, subjects attended to one or another. As the images were superimposed, selection was not spatial for this case.

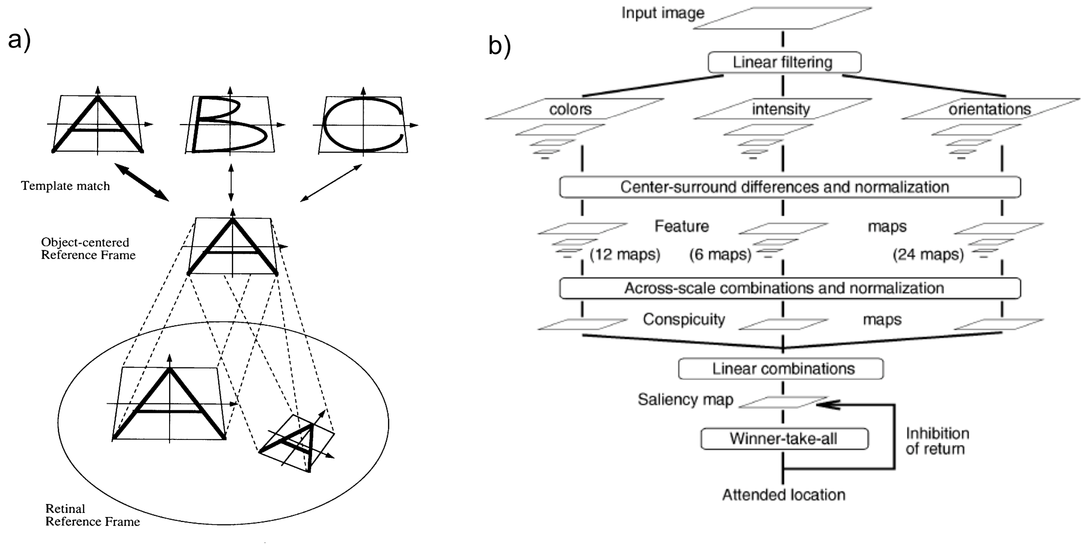

Figure 3.7. a) Olshausen et al. (1998) scale and translation invariant model *(Source:Olshausen et al., 1998)*
b) Itti et al., 1998 model *(Source: Itti et al., 1998)*

Other experiments reported that subjects were fast when reporting two properties of the same object while they had difficulty in describing two properties of different objects, even if they were in the same location. This was called the *same object advantage* (Duncan, 1980). This same object advantage was shown in the automatic spread of attention. Subjects were faster detecting changes in location that belong to a same object than the same change in an equidistant location that belonged to another object (Egly et al., 1994; He and Nakayama, 1995). Other support for object-based attention came from multiple object tracking, neglect patients and simultagnosia. In multiple object tracking subjects tracked several moving objects in the display with moving distractors (Pylyshyn and Storm, 1989). Subjects tracked up to five targets at the same time, which suggested that attention was split between the target objects. Neglect patients could not perceive stimuli in the contralateral visual field to where they had the lesion. Some patients showed an object neglect, neglecting half of the object (Behrmann and Tipper, 1994; Driver and Baylis, 1998). Simultagnosia patients could not perceive more than one object at a time. They could not specify whether two parallel lines were of the same length or not, but if the two lines were joined by means of lines, they could tell if it the new object was a trapezoid or a rectangle (Holmes and Horax, 1919) (Figure 3.5b).

One interesting recent study showed that the units stored in visual working memory were objects (Luck and Vogel, 1997). They showed that objects defined by a conjunction of four features can be retained in working memory just as well as single-feature objects. They proposed that visual working memory stores integrated objects rather than individual features.

It is for those reasons that some authors consider that attention is not only location-based, it is also object-based. In some cases attention can be object-based, and in others location-based or even both at the same time (Scholl, 2001). Attention to objects and groups of objects or elements may be the result of the same circuitry (James, 1890; Scholl, 2001).

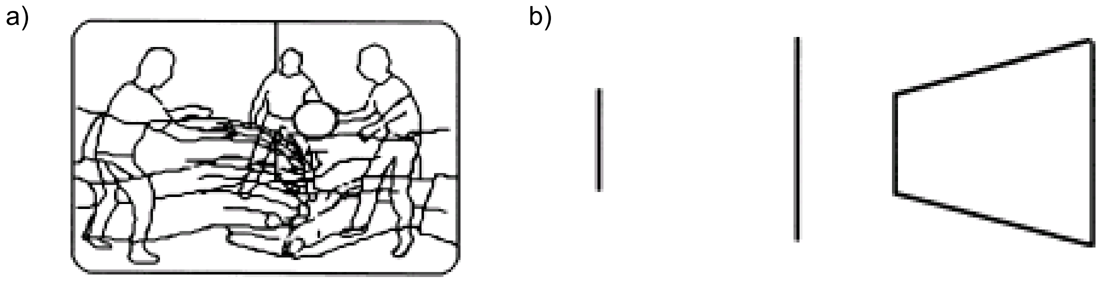

Figure 3.8. **a)** Two superimposed images are seen as two different ones
b) Simultagnosia patients cannot judge if the lines are parallel but can specify if the shape is a rectangle or a trapezoid *(Source: Scholl, 2001)*

Some studies supported that when attending to an object, its features were automatically stored in virtual working memory (Kahneman and Henik, 1981; Duncan 1993). An fMRI study (O'Craven et al., 1999) showed that there were different activations on parts of the brain when attending to one object than when attending to the other, having both of them placed at the same location. They used superimposed faces and houses. When the subject was attending to the face, the face fusiform area became active, but when the house was attended, the parahippocampal place area became active, supporting Neisser's work. Treue and Martínez-Trujillo (1999) in single unit recording and Sáenz et al. (2001) in fMRI found a feature-based mechanism of attention. Attending to features can even precede attention to spatial locations (Hopf et al., 2004).

Among features, spatiotemporal characteristics seem to be the strongest ones (even more than color, orientation or size). This was supported by the work performed with moving objects tracking (Pylyshyn and Storm, 1998) and studies on infants (Xu and Carey, 1996), where they showed that infants up to ten-months old used only spatiotemporal information to recognize an object. That is, when showing an infant two parts that move in the same direction and at the same velocity, they considered them as the same object, while for older infants or children, the other features of these two parts (size, color) suggested that they were different objects.

Regarding these studies about object-based attention, some authors developed several theories. Kahneman and Treisman (1984) and Kahneman et al. (1992) proposed object *files* that saved the objects' properties and whose maintenance and update was performed based on spatiotemporal factors. Three operations are: (1) Correspondence operation determined if the object was a new object or it was an object that has moved from a previous location (2) Reviewing operation extracted an object's previous characteristics and (3) Impletion operation used the reviewed and current information to develop a perceptual phenomenon (e.g. object moving). Pylyshyn's (1989) visual indexing theory augmented this theory and proposed an indexing system for objects independently of their spatial distribution. The theory is very attractive from a computational point of view. But, the object files theory can contradict some studies about the areas of the brain involved on object recognition. This object files would be located in the ventral pathway (involved in object recognition). But if, their maintenance and update is based on spatiotemporal factors, there should be strong connections between the ventral and dorsal pathway (involved in the analysis of motion). Neurophysiological studies have not found strong connections between the ventral and dorsal streams (Felleman and van Essen, 1991).

This evidence for object-based attention does not have to be considered as a contradicting spatial-based attention. Attention to objects (or its features) and attention to spatial locations coexist. Depending on the task, attention will be modulated for an object's features or spatial locations, in visual search tasks attention to features have been shown to precede attention to locations (Hopf et al., 2004). Among features, spatiotemporal features seem to be the most important (Pylyshyn and Storm, 1998), but color, orientation, size and shape are other important features for attention.

4 Object Recognition

Object recognition is the problem of finding and recognizing an object in a scene among other objects. In this chapter a review of object recognition will be presented. Given the breadth of the field, a review of all relevant methods is beyond the scope of this chapter. For more complete reviews see (Besl and Jain, 1985; Bennamoun and Mamic, 2002). First, several definitions of the problem of object recognition will be presented. Secondly the two main types of approaches (bottom-up and top-down) will be explained. Section 4.1 will deal with 2D object recognition methods. Section 4.2 investigates 3D object recognition algorithms, they are divided into object-centred and viewer-centred methods.

There are many definitions for object recognition. Bennamoun and Mamic (2002) considered that a typical vision system consists of five modules: sensing, preprocessing, segmentation, description and recognition, where object recognition comprises the last two. Description is related to the features that differentiate one object from another, while Recognition identifies the object per se.

For Besl and Jain (1985), the problem of object recognition comprised the following steps:

1) Given a set of objects, examine each object and label it.
2) Given an array of pixels from a sensor and a list of objects, those questions arise:
    a) Is the object present in the scene?
    b) If so, how many times does it appear?
    c) For each occurrence find its location in the scene and determine its translation and rotation parameters referred to a known coordinate system.
3) A third optional stage incorporates in the system any unknown objects in the scene (learn from experience).

There exist two basic approaches to solve the problem of object recognition. The bottom-up approach only uses information that comes from the sensors and makes no a priori assumptions. Of particular interest is Marr's theory (1982), that consists of three stages for deriving shape information from the intensity values of the image. The first stage is the *Primal Sketch* that corresponds to the properties of the 2-D image, mainly intensity changes and geometry (blobs, edges, virtual lines, etc.). The second stage is the *2 1/2-D sketch* that accounts for the properties of the image in a viewer-centred frame (distance from the viewer, discontinuities in depth, surface orientation, etc.). Finally, the *3-D representation* is an object-centred representation and its spatial organization by means of a hierarchical representation with volumetric primitives (spatial configurations of sticks or axes) and surface properties.

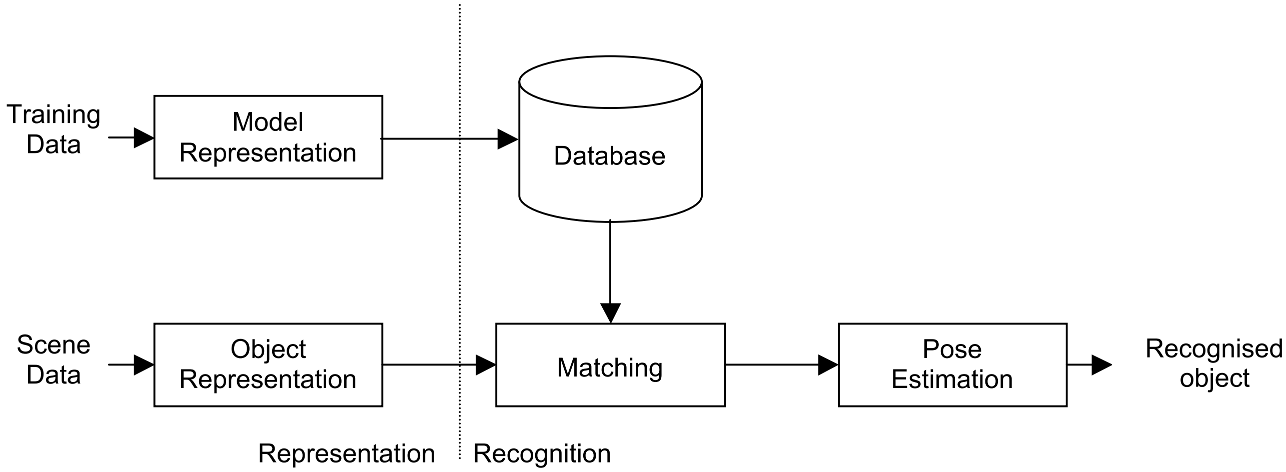

Figure 4.1. Model-based paradigm of object recognition *(Source: Bennamoun and Mamic, 2002)*

The top-down approach presupposes the existence of different objects in the scene and attempts to locate them by using different algorithms. The classic paradigm of this type is the *Model-based object recognition*. A database contains a model representation of the objects to recognize, these models are matched with the representation of the object taken from scene, corresponding the best match to the recognized object (Figure 4.1). Model-based recognition systems are composed of a representational and a recognition step. Recognition is further composed of matching and pose estimation. Matching consists of finding the object from the scene that corresponds to the model. After this, pose estimation is the evaluation of the position and orientation of the object in the scene.

4.1 2D object recognition

Although the majority of the literature presents 3D object recognition methods (see Besl and Jain, 1985 for a review), 2D recognition methods are also very prominent in the literature. 3D object recognition methods require explicit models of the structure of an object, due to this, it requires more information about the object to find or recognize (e.g. several views). 3D modeling processing is tedious and sometimes it is not possible to obtain a 3D model (we have only a “model picture” to find in a scene). In those cases, we need a 2D object recognition method. 2D object recognition is very popular in medical images (MRI, X-ray, etc.). The most basic 2D object recognition approach would be a correlation between the model and the scene, but success of this approach would depend on the luminance, scale, translation and rotation of the object in the image. In this section approaches that apply dynamic programming, probabilities, features and neural networks will be reviewed for the problem of 2D object recognition.

4.1.1 Dynamic programming

Amit (1997) used landmarks in a template image. Dynamic programming was used for the best match between a template graph and the image. Sixteen masks (Figure 4.2 right) were used for the landmarks. Three of these landmarks formed a triangle (Figure 4.2 left). Each triangle had a cost function associated that penalized its shape deviation from the template. The cost functions were translation and scale invariant. By means of dynamic programming matching was accomplished in polynomial time. Amit tested his approach only with MRI images (Figure 4.2 left). He obtained a good level of classification but with too many false negatives. Amit's approach is robust only under small amounts of rotation or scale variations. Dynamic programming algorithms are usually slow and require large amounts of memory.

4.1.2 Probabilities

Leibe and Schiele (2003b) presented a method for object categorization mainly applied for segmentation. Their method created hypotheses without prior segmentation. Hypotheses were constructed matching a rough segmentation of the object with the model database. A later figure-ground segmentation was based on the probability that a given pixel is figure or ground. The probability was calculated as a function of this hypothesis. They also calculated what they called *interesting points*, these points would be the points with more information about the object, but using these points they did not get very high levels of categorization (53.3% only using the first hypothesis). When using all the patches performance increased to 87.6%. They tested their method with cars and cows taken from the ETH-80 database (Liebe and Schiele, 2003a), their conclusion was that their performance is better than using gradient-based methods. This is true since their method used more information than the Gradient, no other comparisons with other methods were performed (e.g. Laplacian) and no other objects were used. Also, due to the recognition process for generating the hypothesis, it has to be more expensive than using the Gradient or the Laplacian.

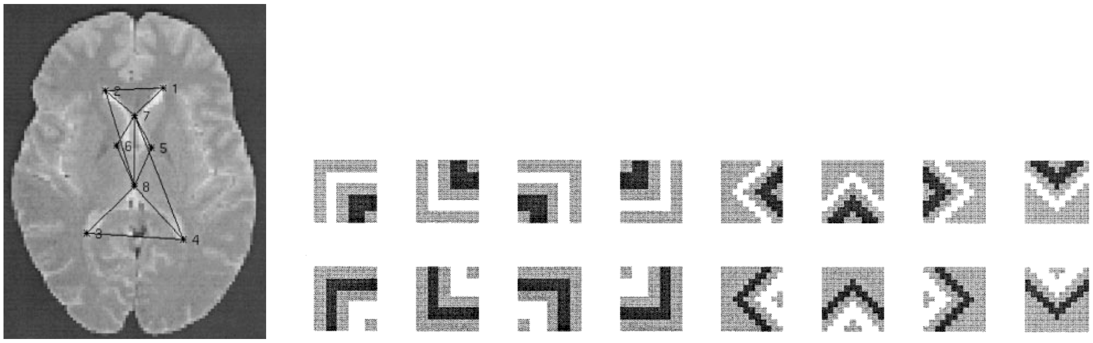

Figure 4.2. Landmarks (left) and masks used by Amit (right) (*Source: Amit, 1997)*

4.1.3 Features

Lowe’s (1999, 2004) Scale Invariant Feature Transform (SIFT) transformed an image into a set of feature vectors. These feature vectors were highly distinctive, translation, scale and relatively rotation invariant (e.g. for rotation: 60° camera rotation and 20° object rotation). First, it detected locations that were maxima or minima of a difference of Gaussians for rotation invariance. These maxima and minima were calculated at several levels of a Gaussian pyramid for scale invariance. At each level of the pyramid the gradients and orientations of each pixel were computed using pixel differences. Gradient values lower than 0.1 times the maximum gradient value were filtered for illumination invariance. In other words, only those pixels that were both minima / maxima at each level of the pyramid and their gradient passed the threshold would be considered as features (with a magnitude and orientation). SIFT keys for the sample images were matched with the keys from new images for recognition. They used Hough transforms and least squares to search for keys under the model pose. They tested their system with different objects (e.g. a book, a toy, sneakers). The most important characteristic about this approach is that it is very robust to occlusions as shown in the results. Invariance was dependent on the parameters they selected (e.g. the threshold for luminance invariance). Also, when testing their algorithm, object poses in the scene were similar in scale and position as the stored models.

4.1.4 Neural networks

Viola and Jones (2001) presented a robust and fast object detector. They first created a new representation of the image called the *integral image*. These integral images were based on features that consisted on the comparison of the sum of pixels from rectangular regions. Only the most informative features were selected. A cascade of classifiers had to learn these features per object. They tested their method with faces and reported satisfactory results for finding faces. The main problem with this method was that false positives was around 40%. Also, it is not clear that this approach generalizes to other objects. Faces are quite characteristic in a scene when using their features, but this would not be the case for finding objects. The method is efficient after training, but we have to consider that the offline learning stage is computationally expensive.

Riesenhuber and Poggio (1999, 2000, 2002) presented a model with five hierarchical levels of neurons that were connected through linear and non-linear MAX operations. (the strongest units determine the response of the system). The first level received input from the retina and was composed of simple neuron receptive fields that analyzed orientations. The next levels accounted for more complex features (e.g. junctions). The last level was composed of view-tuned neurons that achieved position and scale invariance. They only tested their model with paperclips achieving very good results, but it is difficult to evaluate their method only based on these results, paperclips are a very simple kind of objects (they are easy to analyze with orientation detectors). In a later study Schneider and Riesenhuber (2002) tested the model with cars and reported significantly less satisfactory results.

Amit (2000) presented a parallel neural network for visual selection. This network was trained to detect candidate locations for object recognition and it had layers similar to those found in the visual cortex. Objects were represented as composed of features localized at different locations with respect to an object centre. Simple features (edges and conjunctions) were detected in lower levels, while higher levels carried out disjunctions over regions. Detection was accomplished by first constructing a graph of features and finding the candidate regions on the image through a Hough transform. The Hough transform also accounted for size and rotation invariance. Visual attention was accomplished by priming the locations containing the object features. The system was successfully tested with faces, paperclips and numbers. As any neural network, performance depends on training data. Also, features were object dependent, more features, more computation.

Another neural network approach for object recognition (characters in this case) is the one developed by Olshausen et al. (1998), described in chapter 3.

4.2 3D object recognition

There exist two main representation schemes: Object centered and viewer centred. In the object centered representation positions are referenced based on the coordinates of the object of study. In the viewer centered, locations are referenced subject to the viewer (Marr and Nishihara, 1978). Object centered representations may be more compact than viewer centered representations. They also may be used in a viewer-centred coordinate system (Johnson and Hebert, 1999). This overview of 3D object recognition methods roughly follows the classifications of Bennamoun and Mamic (2002) and Besl and Jain (1985).

Techniques using object centred representations

The world model (*W*) is a set of triples (object, translation, rotation) (Besl and Jain, 1985): $W = \{(A_i, \alpha_i, \theta_i)\}_{i=0}^{N_{obj}}$

For a number of objects N_{obj}, each object A_i is in position α_i with orientation θ_i.

4.2.1 Boundary-based representations.

One of the oldest techniques is the *wire-frame representations* (Besl and Jain, 1985) that were based on vertices and edges. These representations have the disadvantage that they can be ambiguous for an object's volume or surface area (Figure 4.3).

Another approach is the *surface boundary representation* (Figure 4.4a) in which an object was approximated by surfaces that determine the object (Besl and Jain, 1985). The object was composed of a set of surfaces, their intersections, and a graph that stored the connectivity of the surface (Bennamoun, 2002).

More sophisticated methods stored features such as the length of segments, angles and probabilities for each segment of the model for being part of the object (Bhanu and Faugeras, 1984). Smoothing images with Gaussian filters have also been used to find discontinuities and extract boundaries (Ponce and Brady, 1987; Fan, 1990).

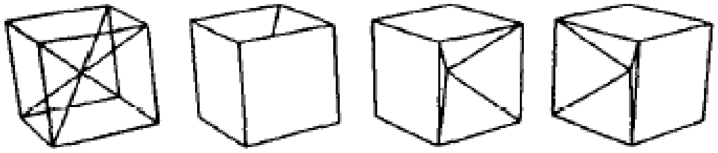

Figure 4.3 Wire-frame model with three different interpretations *(Source: Besl and Jain, 1985)*

Other authors used invariants for object recognition (Figure 4.4b). “Invariants are properties of geometric configurations which remain unchanged under an appropriate class of transformations” (Zisserman et al., 1995), the invariant *I(P)* of a transformation from the coordinates *P'=PT* is: $I(P') = |T|^{w} I(P)$

P' is the transformed configuration and *w* is the weight, if *w=0* the invariant is called a scalar invariant and does not change under transformations.

Invariant methods look for invariance for projections. Zisseman et al. (1995) used three constructions for the invariants in 2D: five lines, a conic plus two lines, and a conic pair. For example for the five lines, two invariants were obtained:

$$I_1 = \frac{|N_{431}||N_{521}|}{|N_{421}||N_{531}|} \qquad I_2 = \frac{|N_{421}||N_{532}|}{|N_{432}||N_{521}|}$$

N_{ijk} is a 3x3 matrix [l_i, l_j, l_k] where the three lines l_i, l_j and l_k have the form: *ax+by+c=0*, and $|N_{ijk}|$ is the determinant. Zisserman et al. (1995) computed the invariants of the target image and compared them to the invariants of the object stored in the objects library. If it was in the library, a hypothesis was produced, which was verified in a later stage. Originally the method was proposed for 2D scenes. An extension of the method was applied for 3D scenes. Zisserman et al. showed several examples to illustrate their model but they did not report test results for their method such as accuracy or efficiency.

Invariants are still very popular in object recognition. The reason for this popularity is that they can be used as indexing functions without any information about the model or its pose. The problem with invariants is that different invariants are needed for different geometric shapes. Sometimes there may be no way of calculating invariants for different structures, this limits the domain of application. Invariance can also be ambiguous for representing different objects. They are good for recognizing simple parts but do not cover the assembling of shapes into more complex objects.

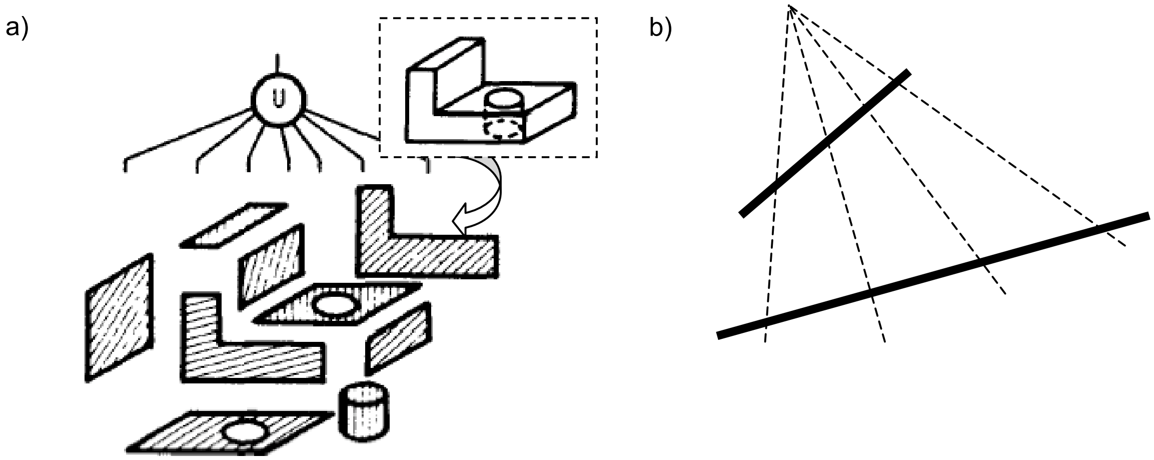

Figure 4.4. a) Surface boundary representation of top object. *(Source: Requina & Voelcher, 1983)*
b) Invariants: Five coplanar points form the cross-ratio invariant:

$$I = \frac{(A-C)(B-D)}{(A-D)(B-C)} = \frac{(A'-C')(B'-D')}{(A'-D')(B'-C')}$$

4.2.2 Curve-based representations

The most popular strategy for curve-based representations in object recognition are B-splines. Splines are piecewise polynomials in an interval *[a,b]* connected smoothly by *knots* (joining points). A spline is of order *k+1* if it has continuous derivatives up to order k-1 and is represented by a polynomial of order ≤ k+1 in each interval delimited by two increasing knots. A spline *s(x)* is defined as (Bennamoun, 2002):

$$s(x) = \sum_{i=0}^{k} c_{i,j}(x-\lambda_j)^i \qquad j = 0,\ldots,g$$

Where λ_j ($\lambda_0 = a$ and $\lambda_k = b$) are the knots and $c_{i,j}$ are the spline coefficients. A spline can also be constructed by means of $g+k+1$ basis spline functions N_i^{k+1} (B-splines) with c_i coefficients for the basis functions (Bennamoun, 2002):

$$s(x) = \sum_{i=-k}^{g} c_i N_i^{k+1}(x)$$

Under this definition, B-splines result to be symmetrical, bell-shaped and can be described by means of a *(k+1)* convolution of a rectangular pulse (Figure 4.5).

B-splines have several properties that make them attractive for object recognition (Cohen and Wang, 1994):

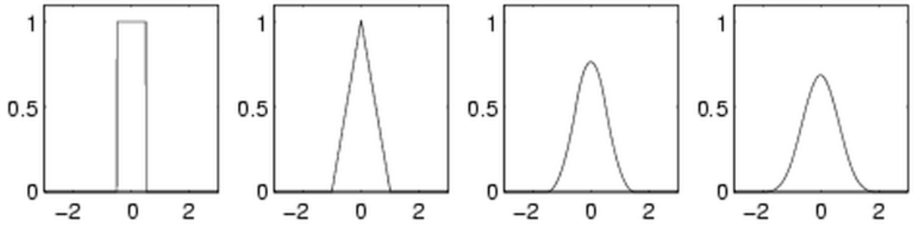

Figure 4.5. From left to right: Splines of first, second, third and fourth order. *(Source: Learning with kernels, Scholkopf and Smola, MIT Press, 2002)*

1. With a set of knots or control points we can describe a curve, splines are a continuous and smooth concatenation of curve segments.
2. They are invariant to affine transformations.
3. A projection of a B-spline is another B-spline with control points related to the original control points by the same transformation.
4. B-splines define globally a curve, but they also have local flexibility to model complex curves.

Cohen and Wang (1994) estimated the best B-spline based on a minimum mean-square error estimation and a Bayesian model for deciding the best order ($k+1$) and control points (knots, k) of the spline. Figure 4.6 shows a spline model of an airplane. They tested their model for the classification of silhouettes at different scales and orientations.

Splines fail in what is called “the knot problem”, where splines have more degrees of freedom than the objects they represent. For a particular surface, there is a number of different combinations. Also, a problem with splines is the initial fitting to data due to the non-linear nature of the optimization problem.

The main problem with boundary-based and curve-based representations is the loss of information when reducing surfaces to curves and lines. This affects robustness and accuracy. Many of these techniques require a growing contour to fit the surface and a seed from the user.

Figure 4.6. B-Spline model of an airplane. *(Source: Das et al. ,1996)*

4.2.3 Axial descriptions

Generalized surfaces, such as generalized cones and generalized cylinders, have been also used for object recognition. A generalized cone and a generalized cylinder consist of a space curve (spine) and a cross-section having the following equations:

$x(u,v) = p + vy(u)$ for the generalized cone and

$x(u,v) = vp + y(u)$ for the generalized cylinder

p is as fixed point (the vertex for the cone) and *y(u)* is the director curve for the cylindrical coordinates *(u, v)*.

Brooks' (1981, 1987) ACRONYM system consisted of three graphs: an object graph had the representation for the object, a representation graph contained constraints for the object models, and a prediction graph. Generalized cones were used to represent parts of objects in the object graph. The arcs on the object graph accounted for the parts relationships. The representation graph contained a hierarchy of constraints, such that any node was more restrictive than its father. Shapes in the image were predicted as ribbons (the 2D analog to generalized cones) and ellipses. ACRONYM predicted what shapes may appear in the image in five phases: (1) identify contours, (2) examine the orientation of the generalized cone with respect to the camera, (3) predict relations between contours, (4) predict shape, and (5) construct constraints. Matches were correspondences between the model representation and the predicted appearances (ribbons and ellipses). Brooks tested his system for recognizing types of planes in aerial images with good results. A flaw of Brooks' approach is that not every object can be decomposed into idealized geometric parts. An additional limitation is that an interaction with the user for building the models and restriction graphs is needed, which may be labour intensive. Finally, the use of generalized cones provides a restricted domain of application.

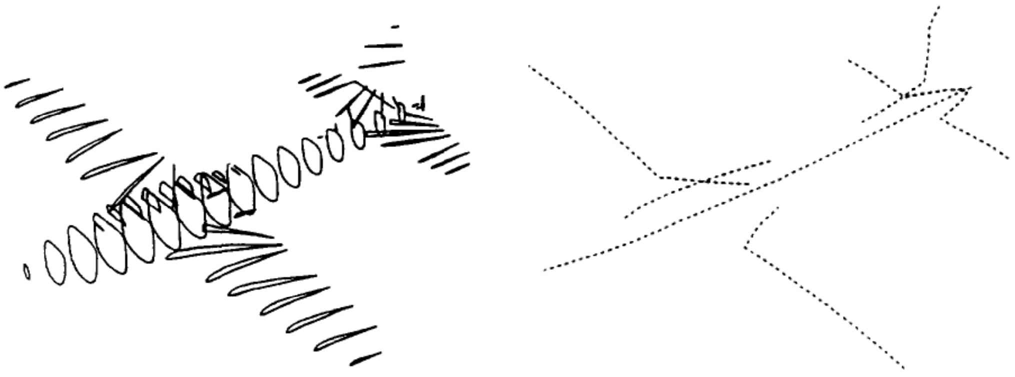

Figure 4.7. Generalized cylinders: Cross-sections (left) and axis (right) of the airplane represented with B-splines in figure 4.5. (*Source: Das et al. ,1996)*

Das et al. (1996) developed a system based on a hierarchical CAD (Computer Aided Design) model. Representations were constructed using two different representations: convex/concave edges and generalized cylinders. These representations were applied to parts of the objects and were used for feature extractions. Objects were recognized by their components characterized by these volumetric primitives in a hierarchy that went from the components in the bottom to a fine description of the whole object in the top. Figure 4.7 illustrates the cross-sections and axis of generalized cylinders applied to a plane. Das et al. also tested their system for the classification of planes from aerial images.

Generalized surfaces can describe objects with only a few parameters, but have problems for representing more complex objects as the ones present in nature. It is difficult to extract generalized cylinders or cones for non-symmetric or non-elongated objects. Another problem is their capabilities with occluded objects.

4.2.4 Surface descriptions

Surface representations use a series of surfaces for the representation of objects. A surface in 3D can be described as (Besl and Jain, 1985):

$$S = \{(x,y,z) : F(x,y,z) = 0\}$$

If the gradient ∇F exists, is continuous and nonzero at every (x,y,z), then S is a smooth surface, e.g. for planar surfaces:

$$F(x,y,z) = Ax + By + Cz + D$$

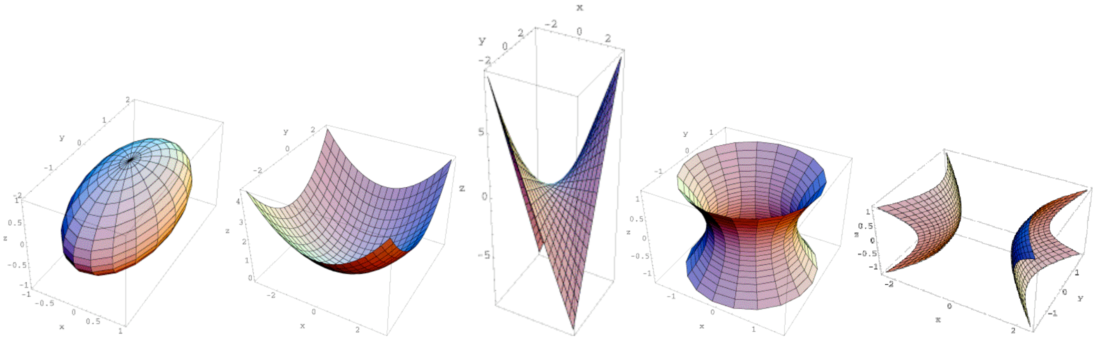

Figure 4.8. From left to right: ellipsoid, paraboloid, hyperbolic paraboloid and hyperboloids of one and two sheets

A, B and *C* account for the direction of the normal to the surface and *D* for the distance to the origin.

Quadrics have also been used for object recognition, quadric surfaces have 9 degrees of freedom:

$$F(x,y,z) = Ax^2 + By^2 + Cz^2 + Gxy + Hyz + Izx + Ux + Vy + Wz + D$$

Depending on the values of the parameters that describe the surface, quadrics can take six different forms (Besl and Jain, 1985, Figure 4.8): (1) ellipsoid (*A>0, B>0, D=-1*), (2) elliptic paraboloid (*A>0, B>0. W=-1)*, (3) hyperbolic paraboloid *(A>0, B<0, W=-1)*, (4) hyperboloid of one sheet *(A>0, C<0, D=-1)*, (5) hyperboloid of two sheets *(A>0, B<0, C<0, D=-1)*, (6) quadric cone *(A>0, B>0, C<0).* Faugeras (1993) used quadrics to fit surface patches of objects.

Gaussian images have been used as surface descriptors. Area and orientation of a polyhedron's face can be represented by point masses on a sphere (Horn, 1984), this sphere is known as the Gaussian sphere, the total mass of this sphere is equal to the total surface of the polyhedron. The Extended Gaussian Image or EGI (Horn, 1984, Figure 4.9a) represents an object surface area in its surface normal, this distribution is then matched with the ones of the objects in the database. In EGI, the normal vectors are weighted by the surface area. EGI is not affected by translations nor rotations and is position invariant since it does not take into account any positional data. This advantage can also be a drawback for unique representations. Kang and Ikeuchi (1993) introduced the Complex Extended Gaussian Image (CEGI, Figure 4.9b) to solve two problems with EGI: The lack of distance information and that in EGI the representation is unique only for convex object.

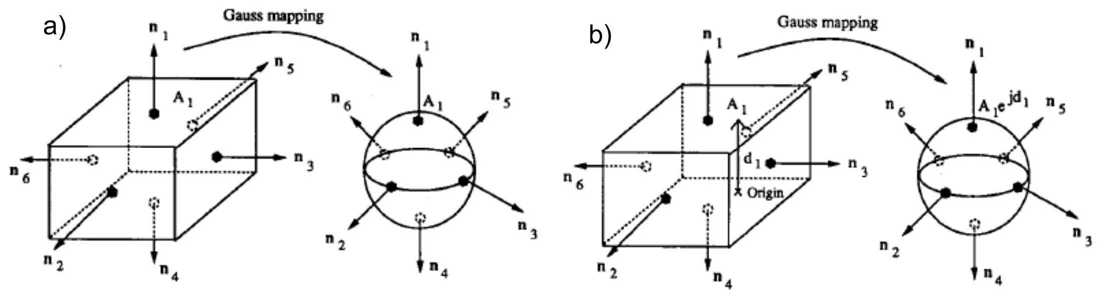

Figure 4.9. Gauss mapping of a cube with EGI (a) and CEGI (b). Only shown the weight for n_1, note that in the area is factored by a phase corresponding to the distance from the centre of the cube to the surface (*Source: Kang and Ikeuchi, 1993*).

In CEGI, distances were included in the weight, being this weight a complex number that accounted for the surface area and distance between the centre of the object and the area. With CEGI it was also possible to estimate the orientation (by means of its distribution) and translation (comparing complex weight phases) of an object with respect to the stored model in the database. Kang and Ikeuchi tested their method with polyhedra, torus and ellipsoids. They obtained successful classifications, except for translations.

The problem with representations based in Gaussian images is that the representation of objects is not unique (especially in EGI). Another flaw with these representations is that they fail when dealing with occlusions.

Recognition has been applied dividing a surface into patches and then match these patches with different types of base patches (Besl and Jain, 1985). Classification of these patches was made regarding their Gaussian curvature and mean curvature arising eight fundamental types of surfaces. The problems regarding surface patches is that they are computationally expensive and have difficulties treating with partially occluded objects and noisy images.

An interesting method is the dynamic balloons of Chen and Medioni (1995). A dynamic balloon was constructed by means of a triangular mesh whose vertices are linked through springs. This balloon was inflated until the vertices lay on the object surface (Fig 4.10). The initial triangular mesh was an icosahedron inside the object (Fig 4.10 left). Vertices moved only in the direction normal to mesh surface using an algorithm for calculating line-surface intersection (Fig 4.10 centre). At the same time the mesh was grown and further divided in new triangles until one or more vertices arrived to the surface. Then, growing was stopped in this place but continued in the remaining directions to fit the object (Fig 4.10 right). The model of a telephone and an automotive part was presented, but no recognition was performed. The disadvantages of this model are: there is the possibility of mesh intersections for some shapes; the technique required a wireframe representation before recognizing an object; it required the selection of a point inside the object manually; and, finally, the system would fail for objects composed of separated parts.

Physical methods as the Finite Element Method have been applied for object recognition by Pentland and Sclaroff (1991). The finite element method computed forces that can be applied to an object for motion and deformation, that were described in terms of nodal displacements *U*:

$$M\ddot{U} + C\dot{U} + KU = R \qquad (1)$$

Where *U* is a 3n × 1 vector of the displacements of the *n* nodal points relative to the object's centre of mass, $\dot{U}$ and $\ddot{U}$ are the first and second derivatives. *M, C* and *K* are 3n × 3n matrices accounting for the object's mass, damping and stiffness. *R* is a 3n × 1 vector that represents the forces acting in the nodes. To achieve object recognition only the vectors of generalized displacements $\tilde{U}_k$ for each model were necessary to be stored. To match an object as one of the *p* models stored, the following was maximized:

$$\varepsilon_k = \frac{\tilde{U} \cdot \tilde{U}_k}{\|\tilde{U}\|\|\tilde{U}_k\|} \qquad k = 1,\dots,p$$

Where $\tilde{U}$ corresponds to the object to recognize and $\tilde{U}_k$ corresponds to each model, $\tilde{U}$, $\tilde{U}_k$ are linear transformations of the *U* vectors to avoid the drawback of the high computational cost (*$O(n^3)$*) of the finite element element method:

$$\tilde{U} = P^{-1}U$$

Pentland and Sclaroff (1991) calculated the *P* matrix using the theory of free vibration modes of the equilibrium equation:

$$KU = R$$

For this they used the eigenvectors of the finite elements that best fitted the surface. This model was also translation, rotation and scale invariant. The authors reported 92.5% correct classification of faces. Accuracy was maintained until a level of 8% of noise. The main problem of this method is the need to segment data into blobs. Also, the orientation of the object has to be estimated in a range of ±15°.

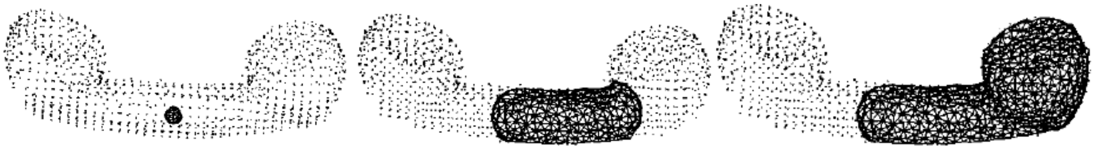

Figure 4.10. Dynamic inflating balloon. The algorithm starts with an icosahedron inside the object (left), then this balloon is growing (centre) until the mesh arrive to the surface of the object, stopping there but growing in other directions (right). *(Source: Chen & Medioni, 1995)*

The Three Dimensional Object Recognition Based on Super-Segments (TOSS) proposed by Stein and Medioni (1992) used two primitives that are encoded in super-segments. A super-segment is a group of connected segments and is represented by a number of segments, angles between two consecutive segments and angles between consecutive normals (Figure 4.11a). Super-segments are invariant with respect to rotation, translation and scale. The primitives that Stein and Medioni (1992) used are the 3-D curve and splashes (Figure 4.11c). The 3-D curve accounts for surface and depth orientation discontinuities examining zero crossings and extremal values (Figure 4.11b). These 3D-curves were fitted with super-segments. Splashes are basic features that represent a surface patch consisting of a Gaussian map in the vicinity of the centre of the patch (Figure 4.11c). These two primitives for every represented object were stored in a hash table, the index for this table was based on a codification of splashes in function of super-segments. To recognize an object in the scene, first, the primitives were computed, then encoded and several hypotheses were generated. The hypotheses that did not correspond to a rigid transformation were removed, by calculating the least squares match between the model primitives values and the ones corresponding to the scene object. Hypothesis that did not conform to a consistent cluster were also removed. The main characteristics of this method are its robustness to occlusions and its performance (*O(kN)*) for *N* models. Stein and Medioni tested their system with different objects. TOSS performance was very slow for calculating the splashes. Splashes calculation for moderately complex objects took a huge amount of computation (e.g. 1h 20 min for a Mozart bust). Another problem of TOSS is that detection is also very slow when there are similar objects in the scene or in noise situations.

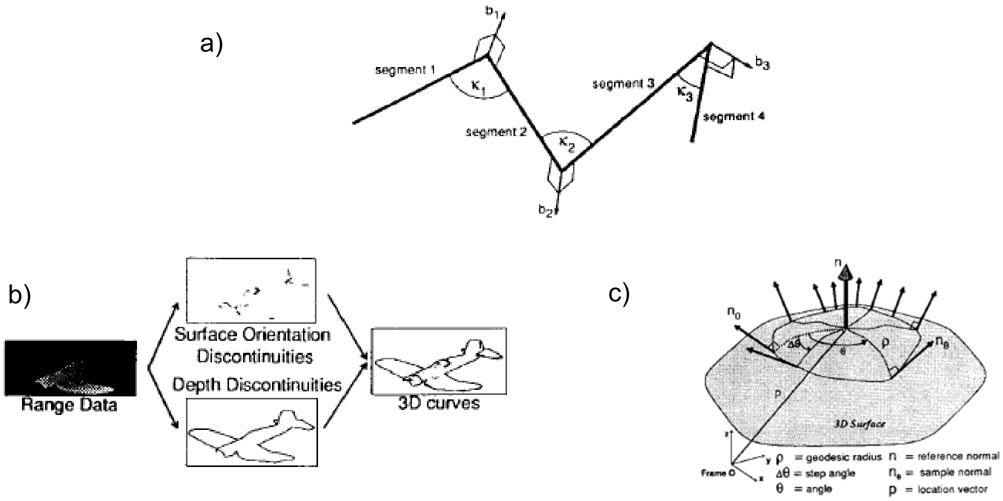

Figure 4.11. a) Super-segments are characterized by the number of segments (here 4), the angles (k_i) and the binormals (b_j) between two consecutive segments, **b)** Extraction of 3-D curves for surface and depth, **c)** Scheme of a Splash.

A recent method was the spin images representation (Johnson and Hebert, 1997, 1999). Spin images relate relative locations of points in an object to a point in the same object, these points are oriented points: 3-D points with an associated direction (Figure 4.12). Given a tangent plane P to this point and a line L parallel to the normal of this plane, an oriented point describes a 2-D basis (p, n). The two basis coordinates are (α, β), given a point x, α is the perpendicular distance to the line L, and β is the perpendicular distance to the plane P (Figure 4.10a). For projecting 3-D points in a basis (p, n) Johnson and Herbert (1997) used a function called spin-map:

$$S_O(x) \to (\alpha, \beta) = (\sqrt{\|x-p\|^2 - (n \cdot (x-p))^2}, n \cdot (x-p))$$

Each oriented point O has a unique spin-map S_O.

For calculating the spin images, a mesh of the surface was performed, for an oriented point O and every point x, the spin map coordinates with respect to O was computed, these spin maps were accumulated into discrete bins to give a spin image, bilinear interpolation was applied in order to reduce noise. Spin images from different orientation points formed a spin image stack. For object recognition, a match between the model's spin images and the one from the object was performed. For this match Johnson and Herbert (1997) used correlation, while other authors (de Alarcon et al., 2002) have used Euclidean distances.

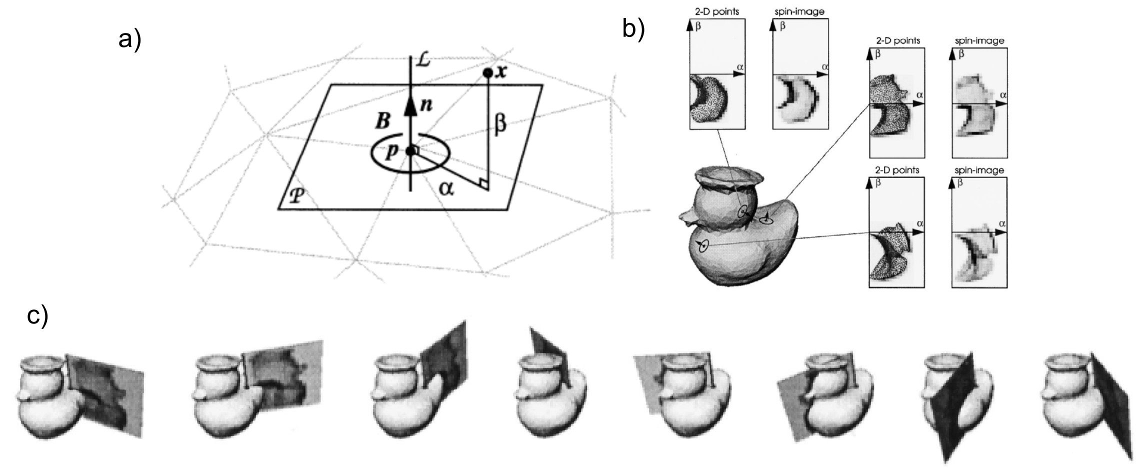

Figure 4.12 a) Orientation points, **b)** spin images of orientation points of a duck, **c)** A spin image can be seen as a sheet spinning around an orientation point. *(Source: Johnson and Herbert, 1997)*

Spin images are robust to occlusion and noise and are translation and rotation invariant but not scale independent. As a result, a normalization of the spin images is usually performed. Furthermore, spin images need huge amounts of storage and can be computational expensive depending on the rendering and the objects. To solve this problem Johnson and Herbert (1999) presented a variation using Principal Component Analysis (PCA), the distance between two spin images in spin image space is the same as the distance between these two spin images in eigenspace. De Alarcon et al., (2002) used Self Organizing Maps (SOM) as an indexing mechanism for fast retrieval from 3-D databases. Johnson and Herbert tested their model in complex scenes and occlusion. Some objects were not recognized due to insufficient data for matching. The system also had problems with occlusion and clutter, that manifested in too many false positives. De Alarcon et al. tested their system with three databases: Aircraft, molecular and mixture. The use of spin images can give rise to very big databases, affecting not only the storage capacity, but also efficiency. For this reason, the authors attempted to find indexing mechanisms for retrieving data from spin image databases that give more complex algorithms and more computation.

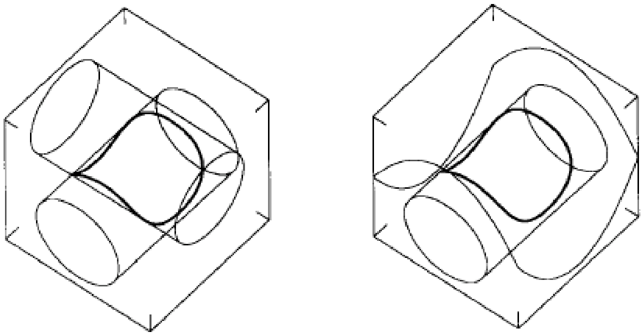

Figure 4.13. Representation of a space curve as an intersection of two cylinders (left) or as an intersection of a surface and a cylinder (right). *(Source: Taubin, 1991)*

Implicit algebraic surfaces have been used by Taubin (1991). 3-D objects can be considered as groups of surface patches that intersect in curve patches (Figure 4.13). Surfaces are implicit functions of three variables *x, y, z*, a space curve is the intersection of two surfaces, and a curve is a function of two variables *x, y*. They developed a method based on eigenvectors to calculate the value of the parameters of these functions that minimize the mean square distance between the input object and the model. The implicit functions used in this approach were cylinders, algebraic curves and superquadrics. Not all the objects can be modelled as intersection of surfaces patches, and not in only one way. This method's weaknesses are stability, uniqueness and efficiency.

4.2.5 Volumetric descriptions

Biederman's (1987) *geons* ("geometrical ions") were modeled by generalized cones. Geons are symmetrical volumes such as blocks, cylinders, spheres and wedges. Biederman proposed 36 geons with the following attributes: shape, degree of symmetry, axis (straight, curve) and constancy of size as it is being swept along the axis. A set of geons compose an object. Geons are related through a set of primitives whose basis are the relative sizes of the components, their orientation and the locations of their attachments. Biederman's geons have been supported with a series of psychophysical experiments.

Superquadrics (Barr, 1981) have been proposed as another method for recognizing objects (Pentland, 1986). Superquadrics are extensions of quadrics (spheres, cones, cylinders, ellipsoids). Superquadrics have the following expression when aligned along the coordinate axis and centred at the origin: $(\alpha x)^n + (\beta y)^n + (\gamma z)^n = k$

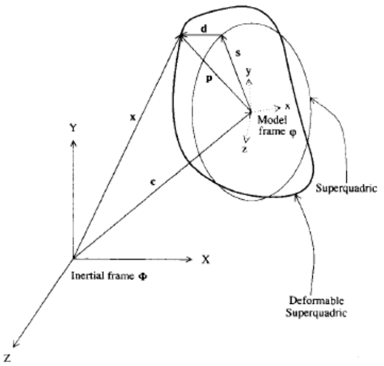

Figure 4.14. Geometry of the deformable superquadric*(Source: Terzopoulos and Metaxas, 1991)*

When $n=2$ we have a quadric. Superquadrics include superellipsoids, superhyperboloids of one piece, superhyperboloids of two pieces and supertoroids (Jaklic et al., 2000).

Terzopoulos and Metaxas (1991) constructed a model mixing superquadrics for their global deformable capabilities with splines for their locally deformable possibilities. They used superellipsoids defined by six parameters, which were controlled by laws of dynamics. Terzopoulos and Metaxas (1991) used the Lagrangian equations of motion for mass, damping and stiffness (Eq 1). Positions were expressed as:

$$x = c + Rp$$

where c is the origin of a model-centred reference frame whose orientation is given by the rotation matrix R, p has the positions of points relative to the model frame and we denoted by its reference shape and displacement:

$$p = s + d$$

For the reference shape s is where the superellipsoid was used (Figure 4.14):

$$s = a\begin{pmatrix} a_1 C_u^{\varepsilon_1} C_v^{\varepsilon_2} \\ a_2 C_u^{\varepsilon_1} S_v^{\varepsilon_2} \\ a_3 S_u^{\varepsilon_1} \end{pmatrix} \qquad -\frac{\pi}{2} \le u \le \frac{\pi}{2} \qquad -\pi \le v \le \pi \qquad \begin{matrix} C_w^{\varepsilon} = \mathrm{sgn}(\cos w)|\cos w|^{\varepsilon} \\ S_w^{\varepsilon} = \mathrm{sgn}(\sin w)|\sin w|^{\varepsilon} \end{matrix}$$

Where a is the scale parameter, a_1, a_2 and a_3 account for the deformation of the superellipsoid, ε_1 and ε_2 specify its nearness to cubic or sphere. Terzopoulos and Metaxas tested their system with a doll, and egg and a pestle. Their approach required a manual initialization to roughly approximate the object. Also, they needed to adjust several parameters for every object.

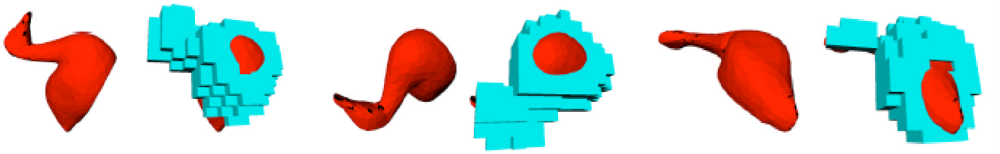

Figure 4.15. Object poses and voxels templates *(Source: Greenspan and Boulanger, 1999)*

Dickinson and Metaxas (1994, 1997) addressed a method for recognizing and localizing 3D objects. They also used superquadrics in a similar way to Terzopoulos and Metaxas (1991) for object modeling. The main difference is that for the local deformable properties they used the Lagrangian equations for motion (eq. 1) instead of splines. A graph of model faces from the object was constructed. For matching they first constructed a *region topology map* in which nodes represented regions (straight, convex, concave) and arcs specified regions adjacencies. A *face topology graph* had nodes containing a set of face probabilities associated with every region. An *aspect hypothesis hierarchy* was generated from the face topology graph. Hypotheses were pruned in relation to their probability, occluded faces and quantity faces. This last step would act as an attention mechanism. Dickinson and Metaxas tested their system with a table lamp and blocks. No accuracy or efficiency results were reported. The problem with this approach is that objects have to be composed of volumetric parts. The main drawback when using superquadrics is that they are symmetric and for this they can be applied to a small domain of objects. They also have problems handling occlusions, Dickinson and Metaxas (1994) tried to avoid this problem with the face topology graph.

Other approaches have used voxels as basic elements that conforms an object. Voxels (volume elements) are small volume primitives, generally cubics of parallelepipeds (Bennamoun and Mamic, 2002). Greenspan and Boulanger (1999) developed a method applying voxels. Taking a template of the object, a set of templates was constructed at various rotations specified by the vertices of an icosahedron. These set of templates were transformed to voxel space (Figure 4.15). A binary decision tree was constructed per template set. This tree was composed of voxels that reference a true branch containing the templates that included the voxel and a false branch with the templates that did not contain the voxel. Three trees were constructed based on heuristics that consider balance (voxels that are in half of the templates), maximum information (voxels that are common to the maximun number of templates) and minimum information (voxels that are common to the minimun number of templates). Recognition was performed by transforming the object to voxel space, randomly selecting one and traversing the binary tree until a leaf node was found. The correlation operator was used to consider the template as a possible solution. This process continued for every voxel. They tested their method with a toy boat and a carved duck and reported efficiency and robustness to occlusion and noise (till 10% signal to noise ratio and occlusion). This technique also requires large amounts of memory.

Constructive Solid Geometry Representation (CSG, Besl and Jain, 1985) used volumetric primitives as blocks, cylinders or spheres that are nodes of binary a tree. The parents of those nodes were operators, such as union, intersection and difference (Figure 4.16). The main advantage of this representation is its simplicity in representing complex scenes with several primitives, but it is computationally expensive and as the previous methods, involving primitives for cylinders, cones, etc. have problems representing natural objects (as faces).

Techiques using viewer-centered representations

Usually viewer-centered representations use various 2D views instead of 3D views and are based on psychophysical findings (Bennamoun and Mamic, 2002), those psychophysical findings conclude that objects are recognized as a set of 2D views.

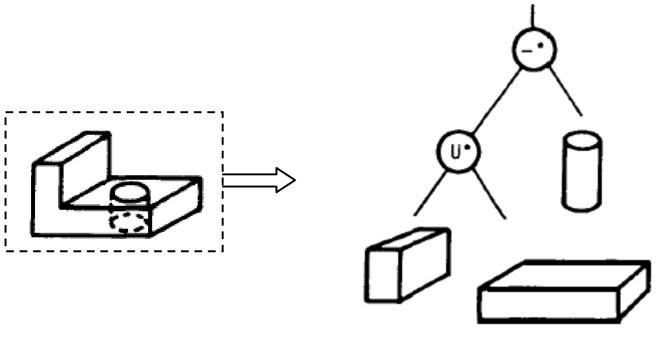

Figure 4.16. CSG (right) of an object (left) *(Source: Besl and Jain, 1985)*

4.2.6 Silhouettes

Bergevin and Levine (1993) developed the Primal Access Recognition of Visual Objects (PARVO) system. PARVO was based in the Recognition By Components psychophysical theory (Biederman, 1987) in which the human visual system can construct coarse descriptions from single views. PARVO was composed of seven modules (Figure 4.16a) and considered that objects were constituted of *parts* (simple volumes) and the connections among them. PARVO first extracted those parts based on algorithms that extract silhouettes using inflection points, tangents and curvature discontinuities (Figure 4.17b). Each part was then classified as one of 22 solids based on three attributes (*labels*, Figure 4.17c) and a specification of the connections between them (Figure 4.17d). For recognition, a model was a graph in which each node is a part (label and aspect ratio) and links were connections between two parts with the relative sizes of the parts connected. Any extracted object from the scene was represented in the same way and matched against the models by means of a measure of similarity. This similarity measure depends on the proportion of the object parts associated to the model parts, the proportion of model parts associated with object parts, and the similarity in the connections between parts. PARVO was tested with several human-made objects. Not all objects can be extracted as composed of simple volumes, limiting the domain of the method. The system also is very dependent on the good extraction of segments.

Ulman and Basri (1991) showed that only a small number of 2D views are needed to recognize an object after applying transformations. Models were denoted by 2D pictures of the object. They showed that a 3D object can be represented by three 2D images of the object if it has sharp contours and for a small number of images if it has smooth boundaries. Occlusions were treated as a special case. If we have a viewed object P, a set of n models $M = \{M_1, \ldots, M_n\}$ and a group of transformations T, matching requires that P and its transformed model TM_i be as close as possible:

$$P = a_1M_1 + a_2M_2 + \ldots + a_nM_n$$

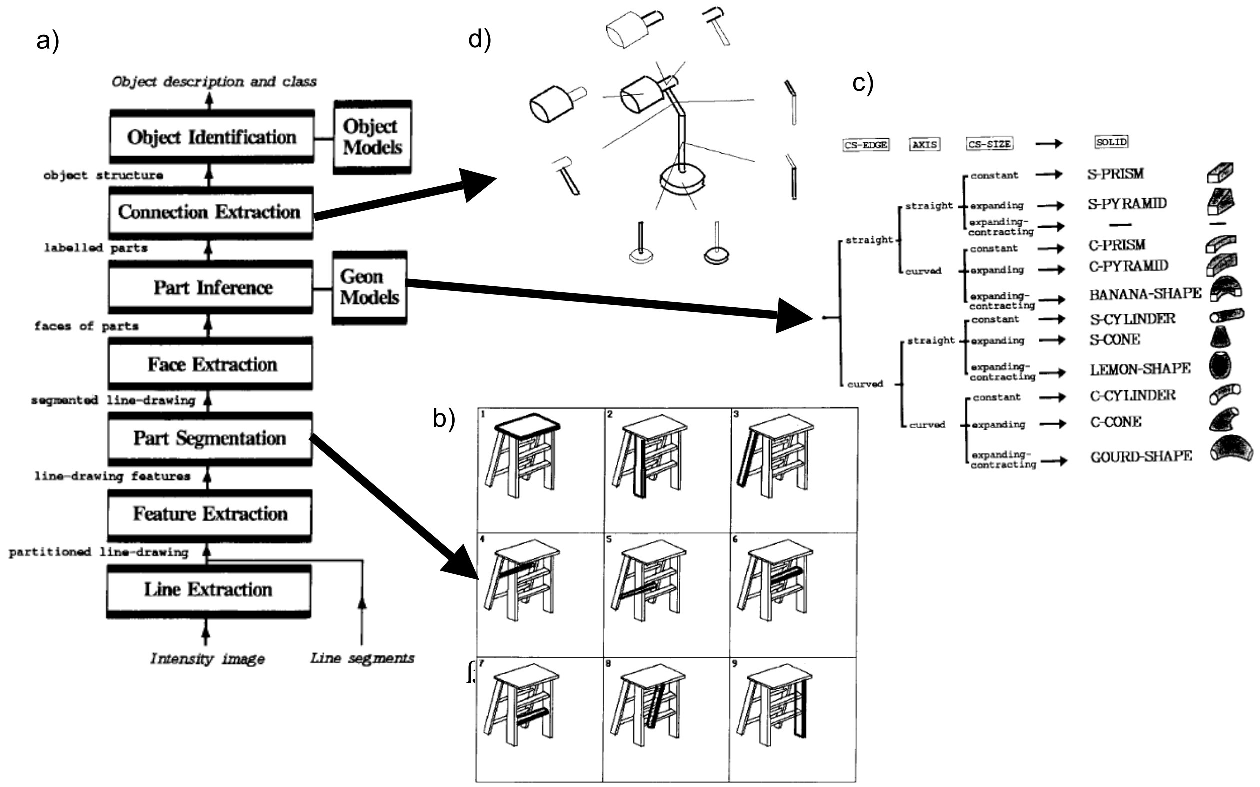

Figure 4.17. a) PARVO structure **b)** Part segmentation **c)** Part classification **d)** Connections
(Source: Bergevin and Levine, 1993)

The point is to find the coefficients $a_1, a_2, \ldots, a_n$ that minimizes a distance measure, and for this they proposed three methods: (1) to use a small number of corresponding features (lines and points), (2) varying the coefficients until they matched under some error measure and (3) linear mapping using a linear operator that mapped the 2D views space into a vector corresponding to the 3D object. They showed a couple of examples on how their model would work in modelling basic shapes and cars. They did not report recognition tests or efficiency and accuracy.

4.2.7 Principal Component Analysis (PCA)

Turk and Pentland (1991) proposed the use of principal component analysis (PCA) in an appearance based approach for object recognition. Murase and Nayar (1995) generalized the PCA approach to capture reflectance properties, pose and illumination. This system could learn objects by first taking several views of the object in different illumination conditions. This set of 2D images at different orientations captured by a sensor was normalized in scale and brightness. Then, the eigenvectors were computed

(Figure 4.18), considering only those ones that accounted for the majority of the variance, these eigenvectors would constitute the eigenspace of the image set. For object recognition, any new object was normalized and transformed to the eigenspace and then matched to the closest object by computing the distances in this eigenspace. For finding the closest object, the authors implemented two approaches: binary search in multiple dimensions and Radial Basis Functions networks. They tested their model with different objects. Learning took 12 hours. This method did not work under occlusions and can have problems regarding the size of the database.

To solve the problem of occlusions, Ohba and Ikeuchi (1997) proposed the use of *eigen windows* (Figure 4.19). Each 2D appearance was separated into small windows, and for each window, the eigenvectors were computed. Matching was performed with these eigen windows. The problem that arises using eigen windows is that they can become very large thus consuming large amounts of memory and that matching can be predisposed to errors. To solve these drawbacks the authors developed a method for selecting the optimal set of windows based on detectability (ease of detection, e.g. a window that contains a corner is more detectable than one containing a planar region), uniqueness (remove windows that are too similar) and reliability (stability of the eigen window when the object is viewed from several views). They tested their method with different objects. This method is not robust under variations in luminance.

4.2.8 Probability distributions

Pope and Lowe (2000) associated a probability distribution to object models over its possible appearances. Objects were described by their appearance rather than by their shape, for appearance they used several features from edge segments to perceptual groupings and regions. In a first stage, models were constructed by learning training images that corresponded to multiple views of the object, these training images were clustered in views and generalized to form a model view (Figure 4.20a). Probability distributions were assigned to the possible appearances of the object in this learning phase, high probability would correspond to the object most usual view (Figure 4.20b).

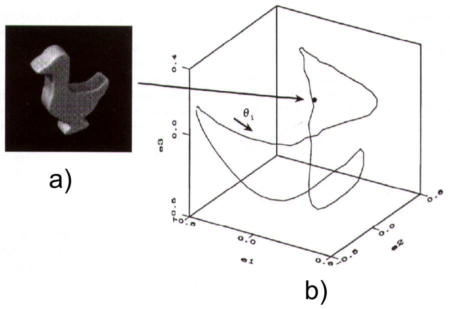

Figure 4.18. Object (a) and its representation in eigenspace (b)
(Source: Murase and Nayar, 1995)

A model view was represented by a graph, in this graph, nodes represented features and arcs represent relations among features. Recognition was performed by first matching features between the object and the models, and match quality was carried out by computing the probability of a hypothesis following the Bayes theorem:

$$P(H \mid E,T) = \frac{P(E \mid T,H)P(T \mid H)}{P(E,T)} P(H)$$

E is the set of pairings $<e_1, e_2, \ldots>$, $e_j = k$ if the model feature j matches the feature k in the image, H is the hypothesis of the modeled view to be in the image and T is a transformation to transform the image feature into model coordinates.

They used the following measure for evaluating the goodness of the match:

$$g(E,T) = \log P(H) + \sum_j \log P(e_j \mid T,H) - \sum_j \log P(e_j)$$

Lowe and Pope tested successfully their system with different objects such as a bunny, a boat and a sneaker at different poses and under occlusions. The system required 19 hours to construct the model of the bunny. Although this method is computationally expensive, very good results were obtained.

4.2.9 Polyhedra

MORAL (Lanser et al., 1997) searched for an appropriate 2D model view that matches roughly the 3D pose of the object to recognize. Models are sets of 2D views under polyhedral approximations. First, associations were constructed between models features and image features taking into account the model views. After that, hypotheses were inferred considering if transformations applied to the image object and the model views gives similar features and have the same constraints.

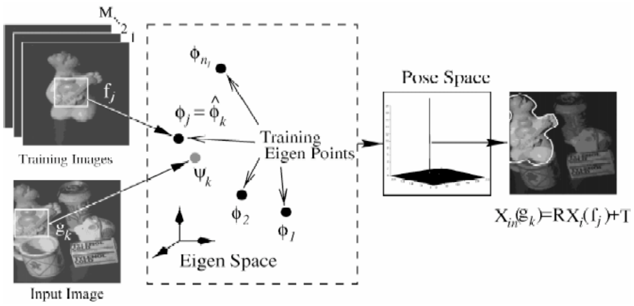

Figure 4.19. Eigen windows (Source: Ohba and Ikeuchi, 1997)

Systems based on polyhedra have the disadvantage that the range of objects that can be recognized is limited.

4.3 Discussion

Object recognition algorithms attempt to find an object accurately and efficiently in a scene. Object recognition can be bottom-up or top-down. Most of the models found in the literature are top-down. Usually in these approaches a database containing a representation of the object(s) to find. These models representations are matched with the representation of the objects in the scene

Models and object representations are mainly based on geometric and physics properties. Some use neural networks to achieve recognition. There is still much work to be done to obtain a method that locates and recognizes object with precision, accuracy and efficiency similar to humans.

Current methods usually have problems with handling translations, rotations, occlusions and noise. They also require large amounts of computation and memory. Some of them require some previous image processing or manual initialization.

It is difficult to perform a direct comparison of methods due to the fact that each study tested their algorithms with different objects. Some of them are starting to use common databases as aircrafts, molecular models, etc. Leibe and Scheile (2003) recently constructed a database of objects (ETH-80) and they tested different appearance and contour based methods for the categorization of objects. Methods were based in colour, texture, global and local shapes (contours).

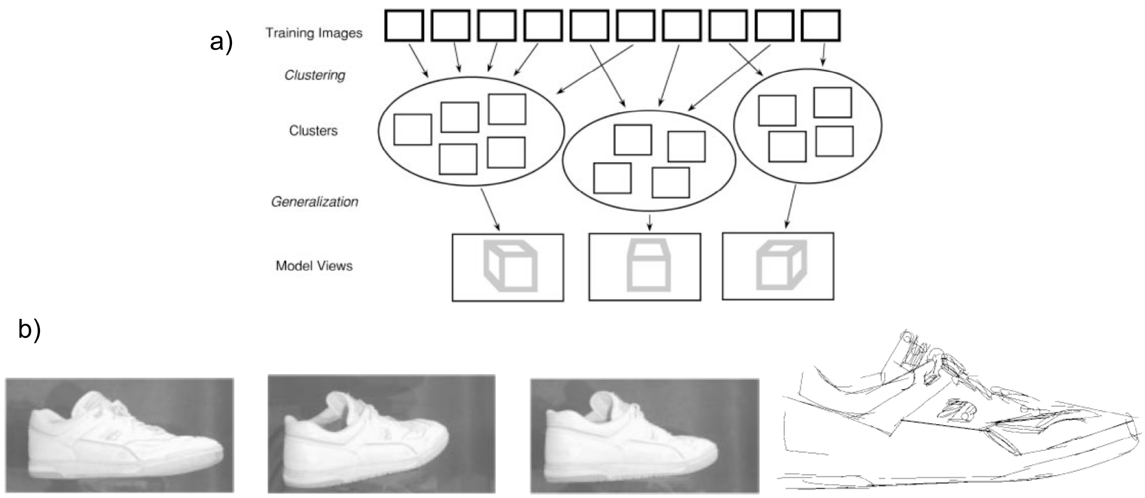

Figure 4.20. a) Constructing model views **b)** Three poses and model view
(Source: Lowe and Pope, 2000)

They compared the results of these algorithms and arrived at the conclusion that the most informative features are contours followed by global shape and texture, the worst performance was achieved by the colour method. But, it is already known that humans can recognize with high accuracy and efficiency objects in grey situations, and what seems the best feature for classifying objects is its shape, most of the object recognition systems use shape for classification or recognition. They tested and compared their method to texture and colour methods. Their shape-based method is superior to the texture and colour based, but they do not compare with other shape-based methods. Shape is a more important feature for recognizing an object than its colour or texture.

5 Conclusions

The actual systems for object recognition are usually geometric or physics-based. Most of them are computationally expensive and have problems when dealing with occlusions, natural images, size, illumination, translation, rotation or noise (the best ones are tolerant to levels of 10% noise). Some have to consider these variations as separate conditions.

Computational models from object recognition rely on matching between the model representation and the objects in the scene. For this matching they need to apply additional methods for size, orientation and luminance invariance. Recent neuroscience studies argue that the human visual system does not perform such a matching (Tanaka, 1996).

Instead of analyzing the whole scene as the current methods for object recognition, the human visual system selects candidate parts of the visual field for analysis. With this strategy, the human visual system recognizes objects in all these different situations in milliseconds with an accuracy close to 100%. The human visual system has no competitor in efficiency, accuracy and variations.

A lot of studies about the human visual system deal with visual search. A visual search task is to find an object (target) among a set of distractors. The target is usually different to the distractors on one or more features. Several theories have appeared to explain the results obtained from visual search tasks. The Feature Integration Theory proposed by Treisman and Gelade (1980) was the pioneer. Today it is an obsolete theory that cannot explain new visual search results, but several important conclusions for other theories have be extracted. After Treisman and Gelade, other theories have appeared. All of them seem to agree on four main aspects:

- Basic features for visual search include colour, orientation, size, etc.
- Grouping of features and objects
- Parallelization of feature/objects processing inside these groups
- Efficiency goes down when the distractors become more similar to the target

Grouping of objects is important for visual search and grouping of features is important for object recognition. There is evidence that the processes encoding the relation between objects are different than the ones involving the grouping of features and are found in different hemispheres, not only that, they seem to work in parallel. These two processes could be primed in a different way depending on the task to develop (task-based selection), this priming would come from attention (Humphreys, 1999). But, the strategy to apply to both groupings would be the same.

A strategy for object recognition and visual search could incorporate a modeling of the visual system for those tasks involving attention. Some models have been tested successfully already using this approach (Grossberg et al., 1998; Deco and Zhil, 2001). A good way to test such a system would be with visual search tasks. Then, a comparison with human performance could be performed.

To be biologically plausible, the model should have a hierarchy of the different areas thought to participate in visual search and object recognition, from bottom to top: V1, V2, V3, V4 and IT. Also, the model should incorporate a winner-take-all strategy of attention to select candidate locations/objects.

Some basic low level features can be extracted from different studies: edges and bars (thought to involve V1), contrast (V1, V2), colour differences (V1, V2) and constancy (mainly V4), size and scale (V1-IT) and motion (V1-MT).

It is already known how V1 neurons respond to different features and they have been modeled with a difference of Gaussians or a Gabor filter. Less is known about the other areas. But, in all areas, studies have found differences in response to attention vs not-attention condtions. Also, receptive fields have increasing size as we go up in the hierarchy, this increase is not linear. Finally, the input and output from each area are known, so, some kind of integration from the different areas is performed.

V4 has been modelled as concentrically organized (Wilson and Wilkinson, 1998) based on results from psychophysics. This organization explains the response to simple contours as curves. TEO neurons (PIT) can codify intermediate features as basic combinations of shape and orientation, color and orientation, etc.

Finally, TE (AIT) can codify complex features and objects (Tanaka, 1996). For combinatorial explosion reasons, it is not possible to have one neuron per view of one object or even per object. It has been proposed that the codification of objects is distributed along the brain, different combinations of neurons codify different objects. One conclusion we can extract from Tanaka's group data is that TEO respond to differences in size, but TE is quite invariant to this property. For this, it seems that TE would accomplish size invariance. Another conclusion is that TE's columnar organization can account for a continuum of features and also for position invariance. Finally, an object can be encoded by the combination of several TE columns representing different complex features.

The organization of these layers is hierarchical, from bottom to top: V1-V2-V4-TEO-TE. Neuron receptive fields increase going up in the hierarchy. For feature extraction and modelling, the path would be from bottom to top. First, basic features would be extracted, these features would be combined at each layer until a representation of the object is formed at the top of the hierarchy. This representation of the object at the top could be stored in a dynamic memory. For object recognition, this memory could have the representation of the object in an object-centred frame. For visual search, the memory would have the spatial relation as well as features of the target and the items in the display. When performing visual search tasks and object recognition, the dynamic memory would be at the top, and it would interact with TE and then go from top to bottom in the hierarchy.

In this hierarchy, we need a strategy to analyze the scene. This strategy is attention. Attention would prime the objects in the scene more similar to the object we want to search or recognize. Models of attention seem to converge in a top-down fashion over the hierarchy with some sort of winner-take-all strategy. The Selective Tuning Model (Tsotsos et al., 1995) would fit this task of selection.

6 References

Ahmand S (1991). VISIT: An efficient computational model of human visual attention. *Thesis TR-91-049. University of Illinois.*

Albright T.D., Desimone R. & Gross C.G. (1984). Columnar organization of directionally selective cells in visual area MT of the macaque. *Journal of Neurophysiology 51(1):16-31.*

Amit Y (1997). Graphical shape templates for automatic anatomy detection with applications to MRI brain scans. *IEEE Transactions on Medical Imaging 16:28-40.*

Amit Y (2000). A Neural network architecture for visual selection. *Neural computation 12:1141-1164.*

Amit Y (2002). 2D object recognition. *MIT Press.*

Andersen RA, Asanuma C, Essick G & Siegel RM (1990). Cortico-cortical connections of anatomically and physiologically defined subdivisions within the inferior parietal lobule. *Journal of Computational Neurology 296:65-113.*

Anderson C & van Essen D (1987). Shifter Circuits: a computational strategy for dynamic aspects of visual processing, *Proceedings of the National Academy of Sciences USA 84: 6297-6301.*

Atkinson RC, Homlgren JE & Juola JF (1969). Processing time as influenced by the number of elements in a visual display. *Perception and Psychophysics 6 : 321-326.*

Barr AH (1981). Superquadrics and Angle-preserving Transformation. *IEEE Transactions on Computer Graphics and Applications 1:11-23.*

Barrow H & Tenenbaum J (1978). Recovering Intrinsic Scene Characteristics from Images. *Computer Vision images. A Hanson and E Riseman eds:271-282.*

Bazier JS (1982). Receptive field properties of V3 neurons in monkey. *Invest Ophthalmology Vision Sciences 23(1):87-95.*

Beergevin R & Levine MD (1993). Generic Object Recognition: Building and Matching Coarse Descriptions from Line Drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence 15(1):19-36.*

Behrman M & Tipper S (1994). Object-based visual attention: evidence from unilateral neglect. *Attention and performance. Conscious and nonconscious processing and cognitive functioning. Urmita & Moscovitch ed, MIT Press:351-375.*

Besl PJ & Ramesh CJ (1985). Three-dimensional object recognition. *Computing Surveys 17(1):75-145*

Bhanu B & Faugeras OD (1984). Shape matching of two-dimensional objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6(2):137-156.*

Biederman I (1987). Recognition-by-components : A theory of human image understanding. *Psychological review 94(2) :115-147.*

Boussaoud D, Urgerleider LC & Desimone R (1990). Pathways for motion analysis: cortical connections of the lateral intraparietal area (area LIP) in the macaque. *Journal of Compuational Neurology 296:462-495*

Bowmaker JK & Dartnall HJ (1980). Visual pigments of rods and cones in a human retina. *Journal of Physiology 298:501-11.*

Broadbent DE (1958). Perception and Communication. *London: Pergamon.*

Broadbent DE (1971). Decision and Strees. *London: Academic Press.*

Broadbent DE (1982). Task combination and selective intake of information. *Acta Psychologica 50:253-290.*

Brooks RA (1981). Symbolic Reasoning Among 3-D Models and 2-D Images. *Artificial Intelligence 17(1-3):285–348.*

Brooks RA (1987). Mdel-based three-dimensional interpretations of two-dimensional images. *In Readings in Computer Vision. Fischler and Firschein editors. Moran-Kaufman:360-369.*

Burdesen C (1990). A theory of visual attention. *Psychological Review 97: 523-547.*

Bushnell MC, Goldberg ME & Robinson DL. Behavioral enhancement of visual responses in monkey cerebral cortex I. Modulation in posterior parietal cortex related to selective visual attention. *Journal of Neurophysiology 46:755-772*

Bennamoun M & Mamic GJ (2002). Object Recognition, fundamentals and case studies. *Springer.*

Cave KR (1999). The FeatureGate model of visual selection. *Psychological Research 62:182-194*

Chelazzi L, Miller EK, Duncan J & Desimone R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature 363: 345 – 347.*

Chelazzi L, Duncan J, Miller EK & Desimone R (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology 80:2918-2940*

Chen Y & Medioni G (1995). Description of Complex Objects from Multiple Ranges Images Using an Inflating Balloon Model. *Computer Vision and Image Understanding 61(3):325-334*

Cohen FS & Wang JY (1994). Modeling Image Curves Using Invariant 3-D Object Curve Models-A Path to 3-D Recognition and Shape Estimation from Image Contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1):1-12.*

Das S, Bhanu B & Ho CC (1996). Generic object recognition using multiple representations. *Image and Vision Computing 14:323-338.*

De Alarcon PA, Pascual-Montano AD & Carazo JM (2002). Spin Images and Neural Networks for Efficient Content-Based Retrieval in 3D Object Databases. *International Conference on Image and Video Retrieval: 225-234.*

Deco G & Rolls ET (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research 44:621-642.*

Deco G & Zhil J (2001). Top-down selective visual attention: A neurodynamical approach. *Visual Cognition 8(1):119-140.*

Deutsch JA & Deutsch D (1963). Attention: some theoretical considerations*. Psychological review 70:80-90.*

Dean P. (1976). Effects of inferotemporal lesions on the behavior of monkeys. *Psychological Bulletin 83 :41-71.*

Desimone R, Albright TD, Gross CG & Bruce C (1984). Stimulus-selective properties of inferior temporal neurons in the macaque monkey. *Journal of Neuroscience 4: 2051-2062.*

Desimone R & Duncan J (1995). Neural mechanisms of selective visual attention. *Annual Reviews of Neuroscience 18, 193-222.*

Desimone R, Schein SJ, Moran J & Ungerleider LG (1985). Many V4 cells exhibit length, width, orientation, direction of motion and spatial frequency selectivity. *Vision Research 125(3):441-52.*

Dickinson SJ & Metaxas D (1994). Integrating qualitative and quantitative recovery. *International Journal of Computer Vision 13(3):1-20.*

Dickinson SJ & Metaxas D (1997). Integration qualitative and quantitative object representations in the recovery and tracking of 3D shape. *in .), Computational and Psychophysical Mechanisms of Visual Coding. L. Harris and M. Jenkin (eds, Cambridge University Press: 221--248.*

Dubner R & Zeki SM (1971). Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Research 35(2):528-32.*

Duncan J (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review 87:272-300.*

Duncan J (1993). Similarity between concurrent visual discriminations: dimensions and objects. *Perception & Psychophysics 54:425-430.*

Duncan J & Humphreys G (1989). Visual search and stimulus similarity. *Psychological physics 57:117-120.*

Duncan J & Humphreys G (1992). Beyond the Search Surface: Visual Search and Attentional Engagements. *Journal of Experimental Psychology 18(2):578-588*

Duncan J & Humphreys G, Ward R (1997). Competitive brain activity in visual attention. *Current Opinion in Neurobiology 7:255-261.*

Duncan J, Ward R & Shapiro K (1994). Direct measurement of attentional dwell time in human vision. *Nature 369:313-315.*

Driver J & Baylis G (1998). Attention and visual object segmentation. *The attentive brain, R. Parasumaran ed, MIT Press: 299-325.*

Egly R, Rafal R, Driver J & Starrveled Y (1994). Covert orienting in the split brain reveals hemispheric specialization for object-based attention. *Psychological Science 5:380-383.*

Fan T (1990). Describing and Recognising 3-D objects using surface properties. *Springer-Verlag.*

Faugeras O (1993). Three-Dimensional Computer Vision: A Geometric Viewpoint. *MIT Press.*

Felleman DJ & van Essen DC (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex 1(1): 1-47.*

Ferrera VP & Lisberger SG (1995). Attention and target selection for smooth pursuit eye movements. *Journal of Neuroscience 15(11):7472-7484.*

Fischer B & Boch R (1985). Enhanced activation of neurons in prelunate cortex before visually guided saccades of trained rhesus monkey. *Experimental Brain Research 44:129-137.*

Fujita I, Tanaka K, Ito M & Cheng K (1992). Columns for visual features in monkey inferotemporal cortex. *Nature 360:343-346.*

Geiger D, Gupta A, Costa LA & Vlontzos J (1995). Dynamic programming for detecting, tracking and matching deformable contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17:294-302.*

Gilmore GC (1985). Aging and similarity grouping in visual search. *Journal of Gerontology 40: 586-592.*

Goldberg ME & Wurtz RH (1972). Activity of superior colliculus in behaving monkey. Effect of attention on neuronal responses. *Journal of Neurophysiology 35:560-574* .

He ZJ & Nakayama K (1995). Visual attention to surfaces in 3-D space. *Proceedings of the National Academy of Sciences USA 92:11155-11159.*

Heywood CA & Cowey A (1987). On the role of cortical area V4 in the discrimination of hue and pattern in macaque monkeys. *Journal of Neuroscience 7:2601-2617.*

Heywood CA, Shields C & Cowey A (1988). The involvement of the temporal lobes in colour discrimination. *Experimental Brain Research 71(2):437-41.*

Holmes G & Horax G (1919). Disturbances of spatial orientation and visual attention, with loss of stereoscopic vision. *Archives of Neurology and Psychiatry 1:385-407.*

Hopf JM, Boelmans K, Schoenfeld MA, Luck SJ & Heinze HJ (2004). Attention to features precedes attention to locations in visual search: evidence from electromagnetic brain responses in humans. *Journal of Neurosciece 24(8):1822-1832)*

Greenspan M & Boulanger P (1999). Efficient and reliable template set matching for 3D object recognition. *Proceedings of the 2nd International Conference on 3D Imaging and Modelling 3DIM:230-237.*

Gross C.G. (1972). Visual functions of the inferotemporal cortex. Handbook of Sensory Physiology, edited by R. Jung. Springer Verlag vol. III, p 451-482.

Grossberg S, Carpenter G & Lesher GV (1998). The what-and-where filter: a spatial mapping neural network for object recognition and image understanding, *Computer Vision and Image Understanding 69(1): 1-22.*

Grossberg S (1998). How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex. *Technical Report CAS/CNS-97-023.*

Grossberg S, Mingolla E & Ross WD (1994). A neural theory of attentive visual search: interactions of boundary, surface, spatial and object representations*. Psychological Review 101: 470-489.*

Hawken MJ & Parker AJ (1987). Spatial properties of Neurons in the Monkey Striate Cortex. *Procceedings of the Royal Society of London, Series B, Biological Sciencies, 231:251-288.*

Heywood CA, Gadotti A & Cowey A (1992). Cortical area V4 and its role in the perception of color. *Journal of Neuroscience 12(10):4056-65.*

He ZJ & Nakayama K (1992). Surface features in visual search. *Nature 359: 231-233.*

Heinke D & Humphreys GW (2003). Attention, Spatial Representation, and Visual Neglect: Simulating Emergent Attention and Spatial Memory in the Selective Attention for Identification Model (SAIM). *Psychological Review:29-87.*

Hoffman JE (1979). A two-stage model of visual search. *Perception and Psychophysics 25:319-327.*

Hoffman JE (1998). Visual attention and eye movements. *In Attention. Pashler ed., Psychology Press. 119-154.*

Horn B (1983). Extended Gaussian Images. *AI memo No 740, MIT, AI laboratory.*

Horowitz TS & Wolfe JM (1998). Visual search has no memory. *Nature 394: 575-577.*

Horowitz TS & Wolfe JM (2001). Search for multiple targets: Remember the targets, forget the search. *Perception & Psychophysics 63:272-285.*

Humphreys GW (1999). Neural representation of objects in space. *In Attention, space and action. Ed. By Humphreys GW, Duncan J and Treisman A. Oxford University Press.*

Humphreys GW, Quinlan PT & Riddoch MJ (1989). Grouping processes in visual search: Effects with single and combined-feature targets. *Journal of Experimental Psychology General 118: 258-279.*

Hubel DH & Wiesel TN (1959). Receptive fields of single neurons in the cat's visual cortex. *Journal of Physiology 148:574-591.*

Hubel DH & Wiesel TN (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology 160:106-54.*

Hubel DH & Wiesel TN (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology 195:215-243.*

Itti L, Koch C & Niebur E (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11):1254-1259.*

Jaklic A, Leonardis A & Solina F (2000). Segmentation and Recovery of Superquadrics. *Computational imaging and vision vol. 20, Kluwer, Dordrecth.*

James W (1890). The principles of psychology. *New York: Holt.*

Johnson AE & Hebert M (1997). Surface matching for object recognition in complex three-dimensional scenes. *Image and Vision Computing 16:635-651.*

Johnson AE & Hebert M (1999). Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 21(5):433-449.*

Kahneman D & Henik A (1981). Perceptual organization and attention. *Perceptual organization. Kubovy & Pomerantz ed. Erlabaum: 181-211.*

Kahneman D & Treisman A (1984). Changing views of attention and automaticity. *Varietys of attention. Parasumaran & Davies ed. Academic Press:29-61.*

Kahneman D, Treisman A & Gibbs BJ (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology 24:174-219.*

Kang SB & Ikeuchi K. The complex EGI (1993): A new representation for 3-D pose determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence 15(7):707-721.*

Kastner S, De Weerd P, Pinsk MA, Idette E, Desimone R & Ungerleider (2001). Modulation of sensory suppression: Implications for receptive field sizes in the human visual cortex. *Journal of Neurophysology 86:1398-1411.*

Kastner S & Ungerleider L (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience 23:315-341.*

Kastner S & Ungerleider L (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologica 39:1263-1276.*

Kinchla RA (1974). Detecting targets in multi-element arrays: A confusability model. *Perception and Psychophysics 15:149-158.*

Kobatake E & Tanaka K (1994). Neuronal selectivities to complex object features in the ventral pathway of the macaque cerebral cortex. *Journal of Neurophysiology 71: 856-867.*

Koch C & Ulman S (1985). Shifts in selective visual attention: towards an underlying neural circuitry. *Human Neurobiology 4:219-227.*

Lanser S, Zierl C, Munkelt O & Radig B (1997). MORAL – a vision based object recognition system for autonomous mobile systems. *7th International Conference on Computer Analysis of Images and Patterns:33-44.*

Leibe B & Schiele B (2003a). Analyzing appearance and contour based methods of object categorization. *Proc IEEE Conference on Computer Vision and Pattern Recognition.*

Leibe B & Schiele B (2003b). Interleaved object categorization and segmentation*. British Machine Vision Conference.*

Lennie P. (1998). Single units and visual cortical organization. *Perception 27(8):889-935.*

Livingstone M & Hubel D (1988). Segregation of form, colour, movement and depth: Anatomy, physiology and perception. *Science 240, p 740-749.*

Lowe (1999). Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision.*

Lowe (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision.*

Luck, SJ, Chelazzi L, Hillyard SA & Desimone R (1997). Neural mechanisms of Spatial Selective Attention in Areas V1, V2 and V4 of Macaque Visual Cortex. *Journal of Neurophysiology 77:24-42.*

Luck S & Vogel E (1997). The capacity of visual working memory for feature and conjunctions. *Nature 390:279-281.*

Malinowski P & Hubner R (2001). The effect of familiarity on visual-search performance: evidence for learned basic features. *Percept Psychophysics 63(3):458-63.*

von der Malsburg C. (1981). The correlation theory of brain function. *Internal Rpt. 81-2, Dept. of Neurobiology, Max-Planck-Institute for Biophysical Chemistry, Gottingen, Germany.*

Marcelja S (1980). Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America 70:1297-1300.*

Marr D & Nishihara HK (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B Series 23;200(1140):269-94.*

Martinez A, Anllo-Vento L, Sereno MI, Frank LR, Buxton RB, Dubowitz DJ, Wong EC, Hinrichs H, Heize HJ, Hillyard SA (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature neuroscience 2:354-369.*

Merigan WH & Pham HH (1998). 4 lesions in macaques affect both single- and multiple-viewpoint shape discriminations. *Visual Neuroscience 15: 359-367.*

McClelland JL & Rumelhart DE (1981). An interactive activation model of context effects in letter perception: Part I. Account of basic findings. *Psychological Review 88:375-407*

McElree B & Carrasco M. The temporal dynamics of visual search: Evidence for Parallel Processing in Feature and Conjunction searches*. Journal of Experimental Psychology: Human Perception and Performance 25: 1517-1539.*

Minkowski M (1920). *Arch Neurol. Psychiatr. 6:201*

Milner P (1974). A model for visual shape recognition. *Psychological Reviev 81: 521-535.*

Moran J & Desimone R (1985). Selective Attention Gates Visual Processing in the Extrastriate Cortex. *Science 229: 782-784.*

Motter BC (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli. *Journal of Nerophysiology 70:909-919.*

Motter BC (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience 14(4):2178-89*.

Mozer MC (1991). The perception of multiple objects. *MIT Press.*

Mozer MC & Sitton M (1998). Computational modeling of spatial attention. *Attention. Pashler Ed., UCL Press: 341-393.*

Murase H & Nayar SH (1995). Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision 14:5-24.*

Nakayama K & Silverman GH (1986). Serial and parallel processing of visual feature conjunctions. *Nature 320:264-265.*

Neisser U (1967). Cognitive psychology. *New York. Appleton-Century-Crofts.*

Neisser U (1979). The control of information pickup in selective looking. *Perception and it Development. Pick ed:201-219.*

Neisser U & Becklen R (1975). Selective looking: attending to visually specified events. *Cognitive Psychology 7:480-494.*

Niebur E, Koch C & Rosin C (1993). An oscillation-based model for the neural basis of attention. *Vision Research 33: 2789-2802.*

Niebur E & Koch C (1994). A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience 1(1):141-158.*

O'Connor DH, Fukui MM, Pinsk MA & Kastner S (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature neuroscience 5(11):1203-1209.*

O'Craven K, Downing P & Kanwisher N (1999). FMRI evidence for objects as the units of attentional selection. *Nature 401:584, 587.*

Ohba K & Ikeuchi K (1997). Detectability, uniqueness and reliability of Eigen Windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19(9):1043-1048.*

Olshausen BA, Anderson CH & van Essen DC (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience 13(11):4700-4719.*

Pashler, H. (1987). Detecting conjunctions of color and form: Reassessing the serial search hypothesis. *Perception & Psychophysics, 41:191-201.*

Pashler H. (1998). Attention. *Psychology Press.*

Pentland A (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence 28:293-331.*

Pentland A & Sclaroff S (1991). Closed-Form Solutions for Physically Based Shape Modeling and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13(7):715-729.*

Pessoa L, Kastner S, Ungerleider LG (2003). Neuroimaging studies of attention: From modulation of sensory processing to top-down control. *The Journal of Neuroscience 23(10):3990-3998.*

Phaf RH, van der Heijden AHC & Hudson PTW (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology 22:273-341.*

Poggio GF & Fischer B (`977). Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Journal of Neurophysiology 40: 1392-1405.*

Ponce J & Brady M (1987). Three Dismensional Machine Vision. *Kluwer International series in Engineering and Science.*

Pope AR & Lowe DG (2000). Probabilistic Models of Appearance for 3-D object recognition. *International Journal of Computer Vision 40(2):149-167.*

Posner MI, Snyder CRR & Davidson BJ (1980). Attention and the detection of signals. *Journal of Experimental Psychology 109:160-174.*

Postma EO, van den Herik HJ & Hudson PTW (1997). SCAN: A scalable model of attentional selection. *Neural networks 10(6):993-1015.*

Press WA, Knierim JJ & Van Essen DC (1994). Neuronal correlates of attention to texture patterns in macaque striate cortex. *Society for Neuroscience 20:838.*

Pyslyshyn ZW (1989). The role of location indexes in spatial perception: a sketch of the FINST spatial index model. *Cognition 32:65-97.*

Pylyshyn ZW & Storm RW (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision 3:179-197.*

Quinlan PT & Humphreys GW (1987). Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches. *Perception and Psychophysics, 41(5): 455-472.*

Ramón y Cajal S (1904). Textura del Sistema Nervioso del Hombre y los Vertebrados.

Ratcliff R (1978). A theory of memory retrieval. *Psychological Review 85: 59-108* .

Reynolds JH, Pasternak T & Desimone R (2000). Attention increases sensitivity of V4 neurons. *Neuron 26:703-714.*

Riesenhuber M & Poggio T (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience 2(11): 1019-1025.*

Riesenhuber M & Poggio T (2000). Models of object recognition. *Nature neuroscience 3:1199-1204.*

Riesenhuber M & Poggio T (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology 12:162-168.*

Rolls E.T. (1987). Information representation, processing, and storage in the brain: analysis at the single neuron level. *In The Neural and Molecular bases of learning, edited by –P Changeux and M Konishi. New York, Wiley, p 503-540.*

Rolls ET & Cowey A (1970). Topography of the retina and striate cortex and its relationship to visual acuity in rhesus monkeys and squirrel monkeys. *Experimental Brain Research 10, pp 298-310.*

Rolls ET & Deco G (2002). Computational Neuroscience of Vision. *Oxford Press.*

Sáenz M, Buracas GT & Boynton GM (2002). Global effects of feature-based attention in human visual cortex. *Nature neuroscience 5(7):631-632.*

Schein SJ & Marrocco RT, de Monasterio FM (1982). Is there a high concentration of color-selective cells in area V4 of monkey visual cortex?. *Journal of Neurophysiology 47(2):193-213*.

Schein SJ & Desimone R. (1990). Spectral properties of V4 neurons in the macaque. *Journal of Neuroscience 10(10):3369-89.*

Schneider WX (1995). VAM: neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action. *Visual Cognition 2, 331-375.*

Schneider R & Riesenhuber M (2002). A detailed look at scale and translation invariance in a hierarchical neural model of visual object recognition. *AI memo 2002-011. MIT.*

Scholl BJ (2001). Objects and attention: the state of the art. *Cognition 80:1-46.*

Shapley R & Perry VH (1986). Cat and monkey retinal ganglion cells and their visual functional roles. *Trends in Neurosciencies 9, p 229-235.*

Shokoufandeh A, Marsic I & Dickinson SJ (1999). *Image and Vision Computing 17:445-460.*

Stein F & Medioni G (1992). Structural Indexing: Efficient 3-D Object Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14(2):125-145.*

Tanaka K (1996). Inferotemporal cortex and object vision. *Annual Reviews in Neuroscience 19: 109-139.*

Tanaka K, Saito H, Fukada Y & Moriya M (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology 66(1):170-189.*

Taubin G (1991). Estimation of Planar Curves, Surfaces and Nonplanar Space Curves Defined by Implicit Equations with Applications to Edge and Range Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13(11):1115-1137.*

Tipper SP (1985). The negative priming effect: inhibitory priming by ignored objects. *Q J Exp Psychol A 37(4):571-590.*

Terzopoulos D & Metaxas D (1991). Dynamic 3D models with Local and Global Deformations: Deformable Superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13(7):703-714.*

Trappenberg T.P. (2002). *Fundamentals of Computational Neuroscience*. Oxford University Press.

Treisman A (1964). The effect of irrelevant material on the efficiency of selective listening. *American Journal of Psychology 77: 533-546.*

Treisman A (1993). The perception of features and objects*. In Attention: selection, awareness and control. Oxford: Clarendon Press: 5 –35*.

Treue S & Andersen RA (1996). Neural responses to velocity gradients in macaque cortical area MT. *Vision Neuroscience 13(4), 797-804.*

Treue S & Martínez-Trujillo (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature 399:575-579.*

Treisman A & Gelade G (1980). A feature-integration theory of attention. *Cognitive Psychology 12 : 97-136.*

Treisman A & Sato S (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance 16:459-478.*

Tsotsos JK (1988). A complexity level analysis of immediate vision. *International Journal of Computer Vision 2 (1): 303-320.*

Tsotsos JK (1990). Analyzing Vision at the Complexity Level. *Behavioral and Brain Sciences 13(3*):*423 –444.*

Tsotsos JK (1992). On the Relative Complexity of Passive vs Active Visual Search. *International Journal of Computer Vision 7(2):127 – 141.*

Tsotsos JK (1993). An Inhibitory Beam for Attentional Selection, in Spatial Vision. *Humans and Robots, ed. by L. Harris and M. Jenkin. Cambridge University Press: 313-331*.

Tsotsos JK, Culhane S, Wai W, Lai Y, Davis N & Nuflo F (1995). Modeling visual attention via selective tuning. *Artificial Intelligence 78(1-2)*:*507 – 547.*

Tsunoda K, Yamane Y, Nishizaki M & Tanifuji M (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature neuroscience 4(8):832-838* .

Turk MA & Pentland AP (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3(1)*, 71-86.

Ullman S & Basri R (1991). Recognition by Linear Combinations of Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13(10):992-1006.*

Usher M & Niebur E (1996). Modeling the temporal dynamic of IT neurons in visual search: A mechanism for top-down selective attention. *Journal of Cognitive Neuroscience 8(4):311-327*.

Van Essen DC & Anderson CH (1990). Information processing strategies and pathways in the primate retina and visual cortex. *Cerebral Cortex 3: 259-329.*

Van Essen DC & Zeki SM (1978).The topographic organization of rhesus monkey prestriate cortex. *Journal of Physiology 277:193-226.*

Viola P & Jones M (2001). Robust real-time object detection. *Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling.*

Walsh V (1999). How does the cortex construct color?. *PNAS 96(24):13594-13596.*

Wang G, Manabu T & Tanaka K (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research 32, pp 33-46*.

Webster MJ & Urgerleider LG (1998). Neuroanatomy of Visual Attention. *The attentive brain. Parasuraman ed, MIT Press.*

Wilkinson F, James TW, Wilson HR, Gati JS, Menon RS & Goodale A (2000). An fMRI study of the selective activation of human extrastriate form vision areas by radial and concentric gratings. *Current Biology 10:1455-1458.*

Wilson H, Wilkinson F, Asaad W (1997). Concentric orientation summation in human form vision. *Vision Research 37(17):2325-2330.*

Wilson H, Wilkinson F (1998). Detection of global structures in glass patterns: implications for form vision. *Vision Research 38:2933-2947.*

Wolfe JM & Cave RK (1989). Guided Search: An alternative to the Feature Integration theory for visual search. *Journal of Experimental Psychology: Human Perception and Performance 15: 419-443.*

Wolfe JM, Friedman-Hill SR, Stewart ML & O'Connell KM (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance 18(1):879-892.*

Wolfe JM (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review I:202-238.*

Wolfe H. (1998). Visual Search. *In Attention, Psychology Press:13-73*.

Woodman GF & Luck SJ (2003). Serial deployment of attention during visual search. *Journal of Experimental Psychology: Human Perception and Performance 29(1):121-138.*

Wurtz RH & Mohler CW (1976). Enhacement of visual responses in monkey striate cortex and frontal eye fields. *Journal of Neurophysiology 39:745-765.*

Xiao DK, Marcar VL, Raiguel S.E & Orban GA (1997). Selectivity of macaque MT/V5 neurons for surface orientation in depth specified by motion. *European Journal of Neuroscience 9(5), 956-964.*

Zeki S (1977). Simultaneous anatomical demonstration of the representation of the vertical and horizontal meridians in areas V2 and V3 of rhesus monkey visual cortex. *Proceedings of the Royal Society of London B Biol Sci. 195(1121):517-23.*

Zeki S (1983). Colour coding in the cerebral cortex: the responses of wavelength-selective and colour-coded cells in monkey visual cortex to changes in wavelength composition. *Neuroscience 9(4):767-81*.

Zeki S, Aglioti S, McKeefry D & Berlucchi G(1999). The neurological basis of conscious color perception in a blind patient*. PNAS 96: 14124-14129.*

Zisserman A, Forsyth D, Mundy C, Rotwell C, Liu J & Pillow N (1995). 3-D object recognition using invariance. *Artificial Intelligence 78:239-288.*