Image Synthesis Based on a Model of Human Vision

Ross Brown

B.Comp. (Hons)
La Trobe University, Bendigo

Smart Devices Research Group
School of Software Engineering and Data Communications
Queensland University of Technology
G.P.O. Box 2434, Qld, 4001, Australia

Submitted as a requirement for the degree of Doctor of Philosophy
Queensland University of Technology
March 2003

Keywords

image synthesis, animation techniques, visual importance, visual attention, visual perception, adaptive rendering, ray-tracing, adaptive texturing, fuzzy logic, progressive rendering

Abstract

Modern computer graphics systems are able to construct renderings of such high quality that viewers are deceived into regarding the images as coming from a photographic source. Large amounts of computing resources are expended in this rendering process, using complex mathematical models of lighting and shading.

However, psychophysical experiments have revealed that viewers only regard certain informative regions within a presented image. Furthermore, it has been shown that these visually important regions contain low-level visual feature differences that attract the attention of the viewer.

This thesis will present a new approach to image synthesis that exploits these experimental findings by modulating the spatial quality of image regions by their visual importance. Efficiency gains are therefore reaped, without sacrificing much of the perceived quality of the image. Two tasks must be undertaken to achieve this goal. Firstly, the design of an appropriate region-based model of visual importance, and secondly, the modification of progressive rendering techniques to effect an importance-based rendering approach.

A rule-based fuzzy logic model is presented that computes, using spatial feature differences, the relative visual importance of regions in an image. This model improves upon previous work by incorporating threshold effects induced by global feature difference distributions and by using texture concentration measures.

A modified approach to progressive ray-tracing is also presented. This new approach uses the visual importance model to guide the progressive refinement of an image. In addition, this concept of visual importance has been incorporated into supersampling, texture mapping and computer animation techniques. Experimental results are presented, illustrating the efficiency gains reaped from using this method of progressive rendering.

This visual importance-based rendering approach is expected to have applications in the entertainment industry, where image fidelity may be sacrificed for efficiency purposes, as long as the overall visual impression of the scene is maintained. Different aspects of the approach should find many other applications in image compression, image retrieval, progressive data transmission and active robotic vision.

Publications

Brown R., Pham B., Maeder A., "A Fuzzy Model for Scene Decomposition Based on Preattentive Visual Features", *Proceedings of Human Vision and Electronic Imaging IV*, San Jose, USA, 1999, vol 3644, pp. 461-472.

Brown R., Pham B., Aidman E., Maeder A., "Efficient Image Rendering Using a Fuzzy Logic Model of Visual Attention", *Proceedings of Advances in Intelligent Systems: Theory and Applications (AISTA)*, Canberra, Australia, 2000, pp. 314-319.

Brown R., Pham B., Maeder A., "A Fuzzy Logic Model of Visual Importance for Efficient Image Synthesis", *Proceedings of the Tenth IEEE International Conference on Fuzzy Systems*, Melbourne, Australia, 2001, pp. 1400-1403.

Brown R., Pham B., Maeder A., "Visual Importance-biased Image Synthesis Animation", *Proceedings of the 1st International Conference on Computer Graphics and Interactive Techniques in Australia and South East Asia*, Melbourne, Australia, 2003, pp. 63-70.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

Acronyms

CCIR	Comité Consultatif International pout les Radiocommunications
CAD	Computer Aided Design
CIE	Commission Internationale de l'Éclairage
CPD	Cycles Per Degree
CSF	Contrast Sensitivity Function
DCM	Discontinuity Coherence Map
DCT	Discrete Cosine Transformation
DOF	Degree Of Fulfillment
FIT	Feature Integration Theory
GSM	Guided Search Model
HSV	Human Saturation Value
HVS	Human Visual System
JND	Just Noticeable Difference
JPG	Joint Photographic Experts Group
LGN	Lateral Geniculate Nucleus
MIP	Multum In Parvo
MPEG	Motion Picture Experts Group
MSE	Mean Square Error
OR	Optic Radiations
POR	Point Of Regard
RGB	Red Green Blue
ROI	Region Of Interest
VSTM	Visual Short Term Memory

Units

Cycles Per Degree	spatial cycles per visual degree subtended
deg. (°)	degrees in degrees or radians
Hz	cycles per second
ms	millisecond (1×10-3 seconds)

nm	nanometres (1×10-9 metres)
Pixels	picture elements in image-space
sec.	seconds
Texton	fundamental texture unit in Texton Theory
Texels	image elements in texture-space

Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: …… Date: ……..

Acknowledgements

There are many people I wish to thank.

First and foremost, I wish to thank my supervisor Prof. Binh Pham and co-supervisor Prof. Anthony Maeder. Your encouragement, strategic insight and shear belief in my work pulled me though many tough times.

Many thanks to my mother, brothers and sister, who have encouraged my academic pursuits to this day. Those times table drills finally paid off. You now have a doctor in the family!

Thanks also to the other Ph.D. students at QUT and Ballarat University, who have provided support, critiques, general high japes and company for coffee.

Thanks also to Sarah for support, hugs and soup when it was needed.

Finally, thanks to the little old lady who poked her head in the rear window of the car when I was a toddler, saying, “Oh, he’s got a big head! He’ll be a doctor someday!”

I guess she was right...

Chapter 1
Introduction

Three dimensional computer graphics, or *image synthesis* as it is formally titled, is the computerised generation of images from a synthetic *scene description*. Scene descriptions are comprised of complex mathematical models of geometry, lighting and surface properties.

The scene description may be gained from a number of sources including manufacturing and design data for Computer Aided Design (CAD) or experimental data for scientific visualisation. From media advertising to viewing oil exploration data, image synthesis techniques have led to enhancements in information presentation in a number of application areas, including:

- entertainment, such as movies and interactive computer games;
- data visualisation in design, science and business applications;
- virtual reality for education and training.

A process called *hidden surface removal* [46] is applied to the scene description data to produce an image revealing the content of the scene from a particular viewing position. An algorithm commonly used to perform hidden surface removal is *ray-tracing*, whereby a virtual light ray is fired through a pixel in the image plane into the scene description. The pixel is then coloured according to the first object intersected by that ray. The inherent image-space nature of ray tracing allows for high-fidelity *photo-realistic* rendering with complex lighting effects. Ray tracing has therefore become ubiquitous in the last three decades, due to the rise in use of high quality digital forms of image representation in many areas. Furthermore, these techniques have grown in sophistication, to the extent that it is now hard for viewers in certain circumstances to discern the difference between captured images and those generated by computer systems.

However, along with this sophistication has come the need for increasing amounts of computing resources to render these images within a reasonable time frame. Even with the present increase in processor speeds, it will be decades before scenes of photo-realistic quality are rendered in real-time [16]. Consider, for example, the computations involved for a typical 1000×1000 pixel image. Each of the million pixels in the final image must be coloured by systems that incorporate complex lighting calculations. These systems model the emanation of light from a source, its transport through a medium (usually air) and the interaction of this light with the surface to be viewed. The number of these calculations is then multiplied by the number of primitives involved in the complex geometric modelling of the surfaces in the scene. As a result, it is not unusual for an image to take a full day to be rendered [57].

The improvements in image synthesis realism have also brought about a concomitant increase in the structural complexity of the scenes to be rendered. Visualisation of large scale architectural and entertainment data sets involves the processing of millions of polygons. Along with this increase in scene description complexity is the latest trend towards computerised rendering of complete motion pictures [131]. A full-length motion picture quality animation requires thousands of frames to be rendered, with each frame having the presently mentioned computational overheads. Therefore, it is expected that the need to improve the efficiency of ray tracing algorithms will continue into at least the near future.

The main cost of ray tracing is the calculation of ray intersections with objects in the scene description. Techniques developed to ameliorate this computational cost fall into two main categories: reduction in the cost of firing a ray into the scene and/or reduction in the actual number of rays fired into a scene.

One solution that seeks to reduce the number of rays fired into the scene is *progressive rendering* [100]. Progressive rendering is the process of generating a synthetic image with a temporal fidelity gradient. That is, instead of rendering an image at the highest level supported by the viewing device, a low fidelity image is

rendered first and then subjected to further refinement until the desired image fidelity is reached. This progressive process thus facilitates speedier previewing of images. Progressive rendering concepts can be used to modify present image-based methods of rendering, such as ray tracing.

Progressive ray tracing initially performs a sparse ray sample across the scene, enabling a quick approximation of the scene to be rendered. This sampling is represented as a subdivision of the two dimensions of the scene, often as a quadtree data structure–this data structure being a recursive subdivision of the image into equally sized quadrants [141]. The image is then further sampled and subdivided at progressively higher levels of fidelity until the image is sampled at least once for every pixel in the image. At this point the subdivision of the pixel itself may occur. However, the value within the frame buffer for the pixel will be an average of the samples taken within its boundaries. This technique of subdividing pixels is known as *supersampling*, it is used to overcome the jagged visual effects that occur along edges in a rendered scene [46]. This supersampling can be made adaptive to features within the scene, due to the need to only sample heavily along contours in an image and less heavily in homogeneous areas [129].

Some ray-tracing techniques [110, 129] do not consider the limitations of the human visual system at all in their allocation of samples, leading to redundant samples being made. Work has been carried out to rectify this waste of sampling resources by allowing for low-level pattern sensitivities within the human visual system [15, 111]. This project, however, seeks to further address the efficiency problems in ray tracing by applying higher-level aspects of human visual attention to control the number of rays fired into the scene. In this newly developed approach, regions with a high level of *visual importance* will be sampled heavily compared to unimportant regions. Compared to other adaptive methods, this approach will bring additional computational cost savings, due to an overall lowering of the number of rays fired.

1.1 INCORPORATING VISUAL ATTENTION INTO PROGRESSIVE IMAGE SYNTHESIS TECHNIQUES

One of the major goals of the project is to modify progressive ray tracing algorithms to refine the visually important regions of an image first, so that the image presented to the viewer in its early stages is at the best possible perceptual quality. Supersampling techniques may also benefit from this approach by modulating the stop condition on the sub pixel refinement by the visual importance of the region. Those pixels in regions deemed to be visually unimportant are sampled less intensively. This approach to rendering reaps much needed efficiency benefits due to the savings in the number of supersamples made for each pixel. A predictive visual attention model therefore needs to be developed to ascertain the visual importance of the regions within the image. This is achieved by exploiting principles of visual attention derived from psychophysical research.

Models of human visual attention are believed to have their physiological basis in *feature detectors* in the early stages of the Human Visual System (HVS) [65]. These feature detectors highlight regions of the viewing field that contain edges and motion, attracting the attention of the viewer. Further corroborating evidence has emerged from psychophysical experiments that indicate the attention attracting capability of changes in visual features, such as luminance and hue [183].

A number of psychological and computational models of visual attention have been constructed to simulate the attention attracting ability of these visual features [72, 108, 128, 157]. Most current models hold that the HVS is in essence a two-stage system. The first *preattentive* stage processes the entire visual field in parallel for feature differences. This preattentive stage guides the later *attentive* stages of the HVS to regions of interest in the visual field [180]. These models assert that the HVS processes the visual field for differences in features, combining them together to form an *importance map* [97] or *saliency map* [83], as a quantification of the attention attracting power of a particular visual field region.

In order to facilitate the exploitation of visual attention principles in image synthesis, an appropriate model must be designed and implemented. Therefore, a further aim of this project is the development of a novel fuzzy logic system that models the attention attracting capability of differences in visual features. Fuzzy logic is used in this project due to its excellent capabilities with regards to modelling imprecise human thought processes. The qualitative nature of the rules governing visual attention in particular makes rule-based fuzzy logic control an appropriate modelling tool [7]. This fuzzy logic module has been integrated into the progressive image synthesis approach developed in this thesis. The role of the fuzzy logic module is to guide refinement processes by ordering data according to rules of visual importance.

The fuzzy logic model contains an implementation of two developed modules. Both modules calculate a visual importance values to control the progressive rendering process. The first module evaluates the visual importance of contours within an image, thereby aiding the progressive rendering process mention previously. The second module evaluates the visual importance of segmented regions within an image, to control the final pixel supersampling process.

In addition, the area of texture *resampling* [46] has benefited from the application of visual importance principles. In a similar manner to the image-space supersampling techniques developed, the size of the support of the texture filter is modulated by the visual importance value of the region. This removes superfluous texture samples in visually unimportant regions.

Computer animation techniques have been modified to incorporate visual importance. An extended version of the region importance model has been developed to allow for the visual importance effects induced by region motion. In a similar way to the still images, the animation frames have their supersampling rates modulated by the visual importance of image regions.

The results produced by the new approach have been analysed in both an objective and subjective manner. Objective L1 and L2 norm error ratios and difference images have been calculated to quantify the size and nature of the differences between the

normal and degraded images. Furthermore, subjective analysis of the quality of the images has been performed, as an indication of the perceptual quality of images rendered by the importance-biased approach.

1.2 RESEARCH QUESTIONS

The research questions investigated in this project are:

- Can the efficiency of present ray-tracing methods be improved by guiding the refinement process to those regions considered to be more visually important, and secondly, to modulate the termination of this refinement process by the visual importance of the image region?
- Can fuzzy logic be used to facilitate importance-based rendering by modelling the relationships between preattentive visual features in a scene description?
- Can texture mapping techniques benefit from the application of similar concepts, to reap efficiency gains by modulating resampling by the visual importance of the region being textured?
- Can motion importance be incorporated into the fuzzy logic visual importance model?
- Can animation rendering techniques be developed to reap efficiency gains from motion feature additions to the visual importance model?

1.3 ORGANISATION OF THESIS

The chapter contents of the thesis are:

- Chapter 2–*Physiology and Psychology of the Human Visual System*, provides a theoretical context for the construction of a visual attention system in both the physiological and psychological domains of research.
- Chapter 3–*A Visual Importance Model for Image Synthesis Efficiency*, details the fuzzy logic module developed to model visual attention. This includes analysis of previous computational attention models,

and the design of the membership functions, implication schemes and rule bases for the new application area.

- Chapter 5–*Adaptive Image Synthesis Using a Visual Importance Model*, deals with the algorithms and supporting data representations for an attention modulated progressive ray tracing system. A prototype implementation is described and objective evaluation results are listed and discussed.
- Chapter 6–*Incorporating Texture Importance into Adaptive Rendering*, describes the processing of textures using the visual attention model and the early detection of texture contours for the progressive renderer. A prototype implementation is described and objective evaluation results are listed and discussed.
- Chapter 7–*Adaptive Image Synthesis Animation*, details the further application of the fuzzy logic model in the area of computer animation. The chapter first details the addition of motion effects to the region-based visual attention model. The chapter then goes on to describe the incorporation of this model into techniques for generating computer animations. The chapter ends with a theoretical evaluation of the new approach.
- Chapter 8–*Overall Approach Evaluation*, analyses data from subjective evaluations recorded during the viewing of images generated by the two progressive image synthesis systems. The remainder of the chapter contains an integrated discussion of the results for all components of the research carried out.
- Chapter 9–Discussion and Conclusions, contains a discussion of the achievements of the thesis and a description of future work in the area of visual attention applications in image synthesis.

1.4 MAIN CONTRIBUTIONS

The main contributions of this research can be divided into the two areas of visual importance modelling and adaptive rendering.

The major contributions to visual importance modelling are:

- the unification of existing visual attention theory into the construction of a computationally efficient and novel region-based model of the HVS attention system, using a rule-based fuzzy logic approach;
- the development of a fuzzy region and contour importance model for integration into progressive image synthesis approaches incorporating adaptive membership functions to account for global effects, complete difference threshold modelling of visual importance and the use of contour importance information from segmented regions;
- the development of a fuzzy logic-based region motion model incorporating motion direction, global effects, abrupt onset effects, gross non-rigid region motion and region-internal motion.

The main contributions in the area of image synthesis are:

- the development of a progressive image synthesis approach incorporating a region-based visual attention model for still images controlling the order, speed and termination of the refinement process, with extensions to texture resampling;
- the development of novel region segmentation techniques for motion detection incorporating object ID information;
- the implementation of the approaches within the framework of the Renderman™ standard, reaping large rendering efficiency gains while keeping perceptual distortion to a minimum;
- the objective and subjective analysis of the rendering approach using error ratios, difference images and subjective image quality experiments.

Chapter 2

Physiology and Psychology of the Human Visual System

Relevant physiological components of the HVS are described in this chapter, delineating the path from the eye to the visual cortex. The chapter then continues with a taxonomy of eye movements and their controlling mechanisms. Key physiological and anatomical constructs, sensitive to particular scene features, are highlighted in order to show their contribution to visual attention processes. These constructs, known as feature detectors, form the basis for the psychological and computational visual attention models detailed in later chapters. In particular, this chapter notes the effects of feature differences on visual attention and seeks to provide a theoretical basis for a new visual attention model. Various aspects of present psychological visual attention models are extracted and analysed to facilitate the development of the fuzzy logic-based visual attention model presented in Chapter 3. Analysis is concentrated on the low-level visual features that contribute to visual importance, eliciting general rules governing their effects. The chapter concludes with a discussion of proposed feature importance hierarchies and summarises the theoretical background to be used in later sections of the thesis.

2.1 HUMAN VISUAL SYSTEM PHYSIOLOGY

The visual perception of an environment by a human being emerges from a complex set of interacting physiological components. Putting it simply, light enters the eye, exciting specialised cells that transmit information via the optic nerve to a posterior region of the brain called the *visual cortex*. Using connections to the visual cortex, the rest of the brain processes this information to provide object recognition and stimulus response. Great progress has been made in understanding the physiology and functionality of the early portions of the primary visual cortex. However, knowledge of the relationship between the rest of the brain and the early stages of the visual cortex is still sketchy at best. A large amount of research effort is now being expended in the task of unravelling these physical connections and their related functionality.

The scope of this thesis only requires a physiological understanding of the early parts of the visual system. The other regions beyond the early visual cortex are outside the terms of reference of this thesis. Therefore, this chapter describes only the major HVS components: the eye, the primary visual cortex and the connection between the two, the optic nerve.

2.1.1 Physiology of the Human Eye and Optic Nerve

A common analogy used for the eye is the *camera*: a *darkened chamber with the image focused on its rear surface*[1]. Similar to a camera, light is reflected from a scene and focused through a system of optics onto the back surface of the eye, known as the *retina* (refer to Figure 2.1).

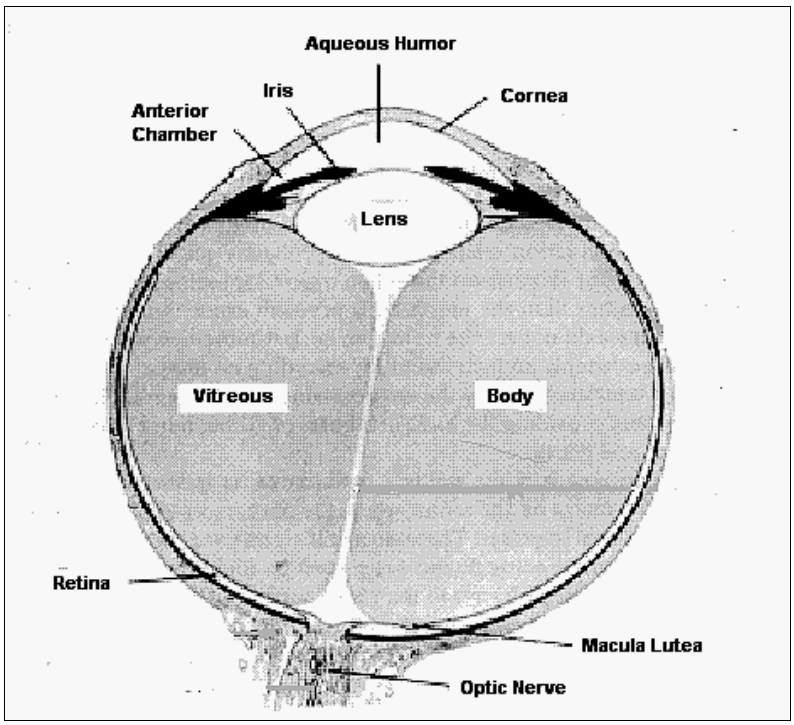


Figure 2.1 Illustration of the cross section of the right eye. Note the location of the Cornea, Lens, Retina, Optic Nerve and the position of the Fovea in the small indentation named the Macula Lutea (adapted from [145]).

The retina is densely coated with *photoreceptive* (light sensitive) cells. These cells are categorised as either *cones* or *rods*, so named because of their respective shapes. The cones are divided into three classes, with each class being sensitive to long (red),

[1] Unless otherwise noted, the concepts in this section are referenced from [145].

medium (green) and short (blue) wavelengths. The cones, numbering approximately six million, are mostly concentrated in an area named the *fovea*. The fovea is located in a small indentation on the retina called the *macula lutea*, in the centre of the HVS visual field. The fovea provides the viewer with a high detail colour view of our surroundings, at higher (*photopic*) light levels. The rods, numbering approximately 120 million, are situated outside of the fovea, providing monochromatic, low detail peripheral vision at low (*scotopic*) light levels. The rods also facilitate high sensitivity to motion in the peripheral field of view. A rod or cone emits an electrical impulse when it absorbs a threshold number of photons of visible light (wavelength 400-700nm). This impulse passes through a layer of neurons before being transmitted along the optic nerve.

This retinal layer of neurons is made up of four main classes of cells: *horizontal*, *bipolar*, *amacrine* and *ganglion* cells. The ganglion cells are connected to the bipolar cells, forming the final layer of the retina. The *axons* (transmission extension) of the ganglion cells make up the *optic nerve* that transmits electrical impulses from the retina to the *visual cortex*. The number of ganglion cells is much less than the number of photoreceptor cells in the retina. This numerical difference suggests a great deal of the photoreceptor output is condensed before being transmitted along the optic nerve. This compression of information happens in two ways. Firstly, the eye only transmits changes in the scene being viewed, causing the image to fade if the eye remains stationary while viewing an invariant image [184]. Secondly, cells in the retina carry out preprocessing of certain features in the scene, such as colour and motion, transforming the scene into a more efficient representation for further processing by the HVS.

Each ganglion cell has a defined receptive field, due to its interconnections with the other neurons in the retina. These regions of the retina, which are roughly circular in shape, affect the firing rate of the ganglion cell when stimulated [20]. Among the many different types of receptive cell field, is the *concentric field*. These fields respond to stimuli at the centre of the field, or in the periphery. The fields that respond to onset of stimuli in the centre and offset in the surround are named *centre-*

on fields. Receptive fields that respond to onset of stimuli in the surround, or the offset of stimuli in the centre, are named *centre-off fields* (refer to Figure 2.2).

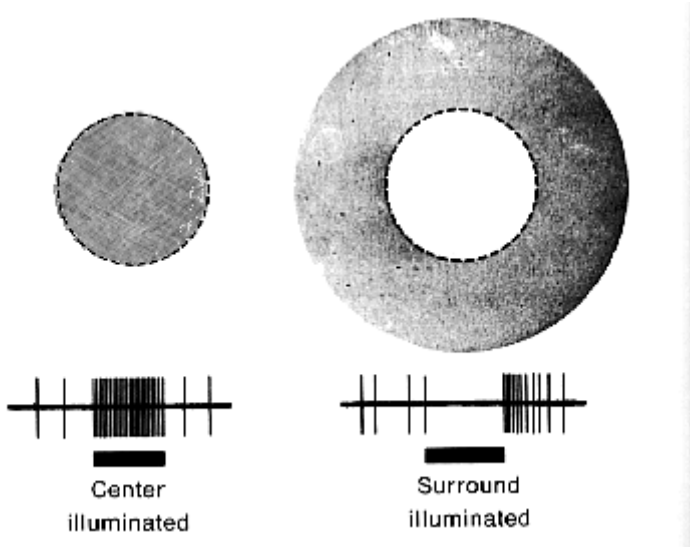


Figure 2.2 Diagram of a typical centre-surround antagonistic receptive field of a single ganglion cell, and the neural pulse sequence which results from illumination within the receptive field [145].

These ganglion cells can be further categorised into X and Y cells. The X cells give sustained responses when a sinusoidal grating is held stationary in front of the receptive field. The Y cells give only a transient response to stationary sinusoidal gratings. Therefore, in order to gain a sustained response from a Y cell, the grating must continually be in motion. The size of these receptive fields for the ganglion cells increases with eccentricity from the fovea. The pooling of receptor output is thus increased with distance from the fovea, suggesting a lowering of the perceived level of detail. The receptive fields of the X and Y cells are also distributed differently over the retina. The X cell receptive fields are concentrated in the foveal area, while the Y cell receptive fields are concentrated in the periphery of the retina.

Ganglion cells also transmit colour information through a colour opponency system. A colour-opponent cell of this type is excited by stimuli of one colour and inhibited by those of another. In monkeys, the centre-on and centre-off X cells are divided into four colour-opponent classes-the most common being the two types of green-red cells. These colour cells contain a centre that is sensitive to green and the surround to red, or vice-versa. Less common are the two types of blue-yellow cells [20] that

have a similar functionality with regards to blue and yellow stimuli. Analogues to these cells are considered to exist in the HVS.

All the above processing takes place in the retina. The motion, colour opponency and other information in the visual field is transmitted along the optic nerve to the visual cortex, illustrated in Figure 2.3.

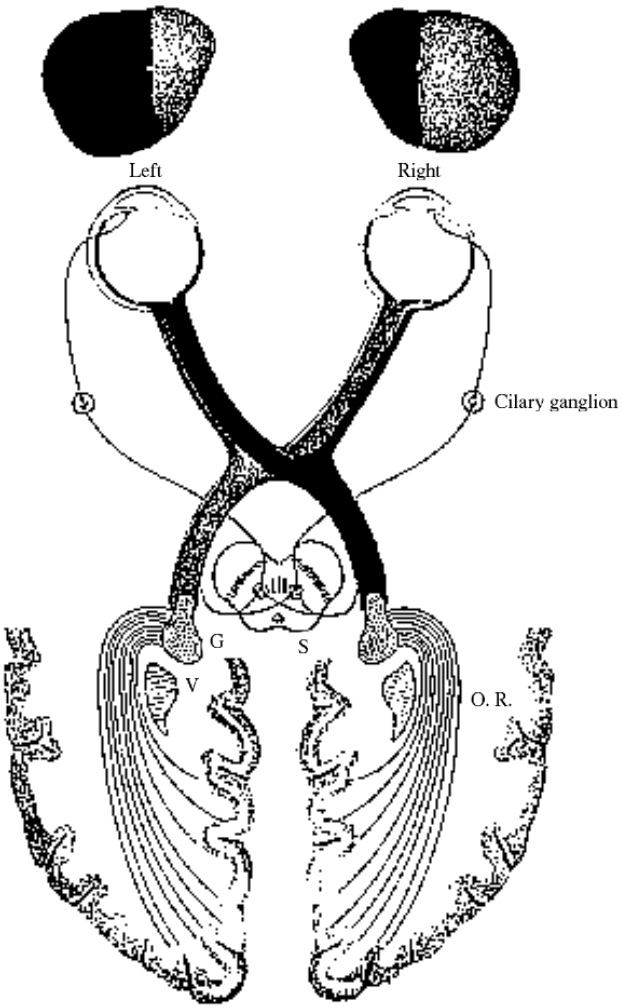


Figure 2.3 Diagram showing the visual pathways from the retina to the visual cortex illustrating the decussation (crossing over) of the fibres from the nasal half of the retina at the optic chiasm. G: lateral geniculate nucleus, S: superior colliculus; III: oculomotor nucleus; V: posterior horn of lateral ventricle; OR: optic radiations [145].

Of special note in Figure 2.3 is the *Lateral Geniculate Nucleus* (LGN), the synaptic termination of the optic nerve (labelled with a G). The cells in the LGN exhibit similar centre-surround characteristics to the receptive cells in the retinal network, for example, red-green opponency [65].

A few points should be noted at this point in the visual pathway. The presence of primitive visual feature detectors so early in the visual system supports the notion of the HVS being hardwired to detect such features in parallel across the visual field. This parallel mechanism of the HVS is explored further in Section 2.2, with regards to psychological models of visual attention. Furthermore, Y receptor cells appear to have a major role in the detection of motion, which is more noticeable in the periphery of human vision where these receptive fields are situated. Finally, the opponent colour sensitive receptors may also give rise to the high levels of contrast noticeable on the edge between contiguous opponent-coloured image areas [69]. This high peripheral sensitivity to scene motion is a strong attractor of visual attention, and along with the concept of colour (hue) contrast, is a component common to all of the leading visual attention models discussed in Section 2.2.

2.1.2 The Visual Cortex

From the LGN, signals continue along the optic radiations to the next major HVS construct, the primary visual cortex. The primary visual cortex (area V1) occupies a large component of the posterior location of the brain. Extensive research has been carried out into mapping the functionality and architecture of the monkey visual cortex, for example [64, 65, 133], which is assumed to map closely to the HVS. The visual cortex is architecturally and functionally hierarchical in nature. In this architecture, LGN concentric field cells converge to *simple cortical cells*, which in turn converge to *complex cells*, which finally converge to *hypercomplex cells*. These classes of visual cortex cells are now considered individually.

Simple cells have receptive fields located in one eye only, with spatially distinct on and off areas separated by parallel straight lines. Large proportions of these simple cells contain opponent colour properties-the largest of any of the cells described here. A line stimulus at a preferred angle, size, shape and retinal position for the given cell produces an optimal response.

The complex cell, similar to the simple cell, responds optimally to lines at particular orientations. It differs from the latter in being unaffected by the position of the

stimulus within the receptive field, if it lies inside or moves inside the receptive field, then the stimulus evokes an optimal response. While the complex cells are position insensitive, half of these complex cells show asymmetry in their response to movement in opposite directions, while others show little or no preference. Complex cells may be optimally responsive to edges, slits or dark bars, with the orientation increment being 5-10°.

The hypercomplex cells respond to movement of objects at an optimal angle, with antagonistic regions above and below the receptive field. They detect any change in the direction of the contours, that is, they detect curvature.

Of the complex and hypercomplex cells, only a low percentage of 10% or less exhibit colour specivity-in keeping with tests showing a degradation of colour perception compared to luminance perception [166]. The cells are retinotopically organised: movement along this section of the cortex represents a similar movement along the surface of the retina. In the periphery the receptive field topography is coarser, with the receptive fields being larger in area. This physiological analysis indicates that two major functions of the cortex are contour analysis and binocular convergence.

A number of points should be made here. Firstly, the complex and hypercomplex cells are sensitive to certain image features: edges, movement and to a lesser extent colour. The retinotopic arrangement of the receptive fields within the primary visual cortex shows that these features are processed in parallel at all locations in the image. Assuming that this physiological structure has a direct perceptual correlate, it can be inferred that the functional structure of the primary visual cortex supports the importance of edges, movement and colour in the human perception of visual stimuli. Any reasonable HVS attention model would need to incorporate these visual features.

Figure 2.4 proposes a schematic summary of the physiological contents of the primary visual cortex. The diagram highlights two major pathways that are surmised to exist in the higher visual areas.

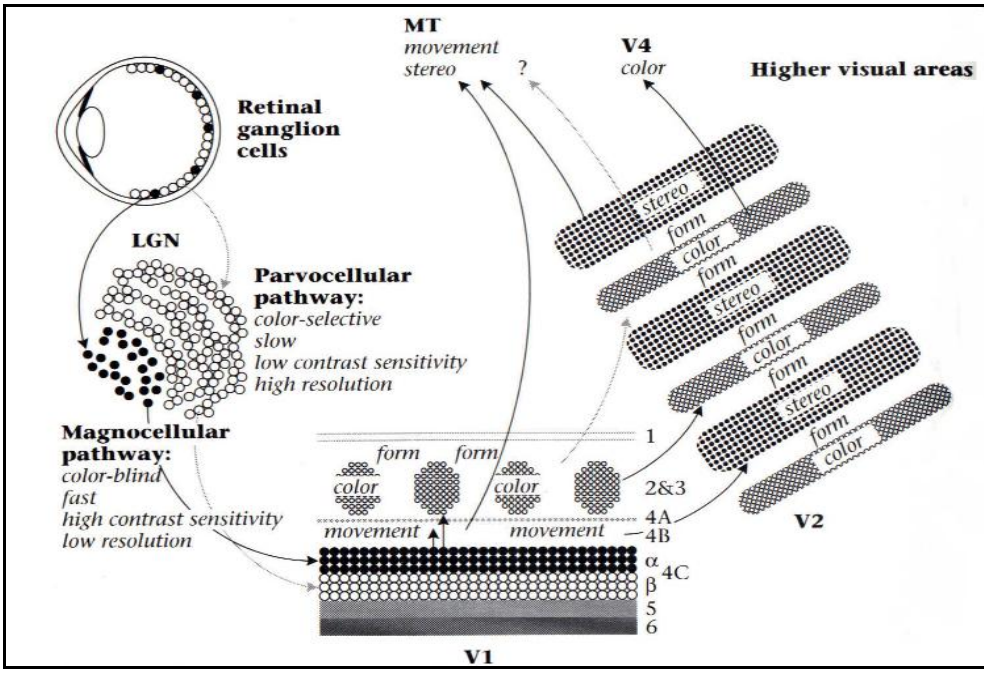


Figure 2.4 An anatomical/perceptual model of the visual cortex. In this speculative model, visual streams within the cortex are identified with specific perceptual features. The anatomical streams are identified using anatomical markers; the perceptual properties are associated with the streams by applying the neuron doctrine [92].

The broad division is into MT and V4 areas. The MT area is considered to be primarily associated with motion perception. The MT area is also characterised as being fast, drawing input as it does from the *magnocellular* pathway from the LGN. The V4 area is thought to deal with colour and form perception, and is slower, due to it drawing its input from the *parvocellular* pathway from the LGN. Due to this structural organisation, it can be postulated that the motion and depth features are faster to manifest perceptually, and as such, are stronger in attracting attention.

Furthermore, Livingstone and Hubel [92] note the agreement between their physiological work and perceptual experiments, that is, the magnocellular pathway aids depth perception, with luminance changes being reported as being perceived more quickly than hue changes. They also describe the colour-blind nature of motion detection. As motion is such a strong attractor, it could be inferred that luminance is a stronger attractor than hue, with motion being stronger than both. They also discuss the previous findings of Gestalt psychologists, where figure

ground and illusory borders disappear after the equi-luminance of colours is established. Furthermore, perspective and depth from shading is also removed under equiluminance.

They surmise that the magnocellular functionality could form the majority of what is required for day to day living in some lower animals. They go on to state that the parvocellular system, which is only well developed in primates, is possibly for the perception of much more detailed information, such as form perception. The parvocellular sections are thus a later evolutionary development for the more detailed and leisurely analysis of objects, while the simpler magnocellular system is more able to quickly detect threats and make depth perception calculations for manoeuvres.

Research has also investigated information theoretic aspects of the physiological organisation of the early stages of the visual cortex. Linkser has found that a form of weighted synaptic neural network using parameterised weight update techniques organises itself into structures exhibiting strikingly similar functionality to that of early stages of the HVS [89-91]. From the input layer onwards of the network a number of visual cortex structures: centre surround cells, orientation specific cells and orientation bands occur in a similar manner to the macaque monkey visual cortex. Linkser observes that this network organisation globally minimises the energy levels of each of the simulated synaptic cells. This value matches results obtained for simulated annealing experiments, and seems to indicate the physical efficiency of the visual cortex organisation.

Further to this, work by Daugman has developed an image encoding model which uses a set of translated, scaled and oriented *Gabor* wavelets to mimic the orientation selective receptors in the HVS [32]. A shortened spatial form of this function is the following:

$$G(x, y) = \exp(-\pi [(x - x_0)^2 \alpha^2 + (y - y_0)^2 \beta^2]) \exp(-2\pi i[u_0 (x - x_0) + v_0 (y - y_0)]) \quad (2.1)$$

where:

x_0, y_0 are the x, y coordinate position parameters for the function;

α, β are the scaling parameters;

u_0, v_0 are the modulation parameters.

When $\alpha \neq \beta$ there is a further degree of freedom that enables a rotation of the function out of the principal axes–not shown here for clarity. The actual spatial shape of the function is shown in Figure 2.5 and in Figure 2.6:

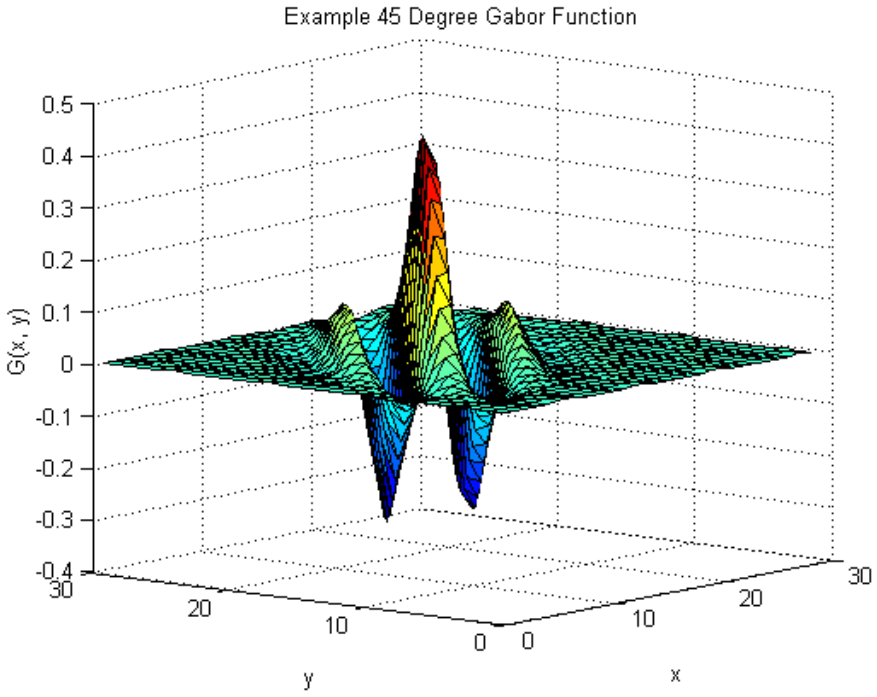


Figure 2.5 Example 3D Gabor function plot for a 45 degree oriented Gabor function.

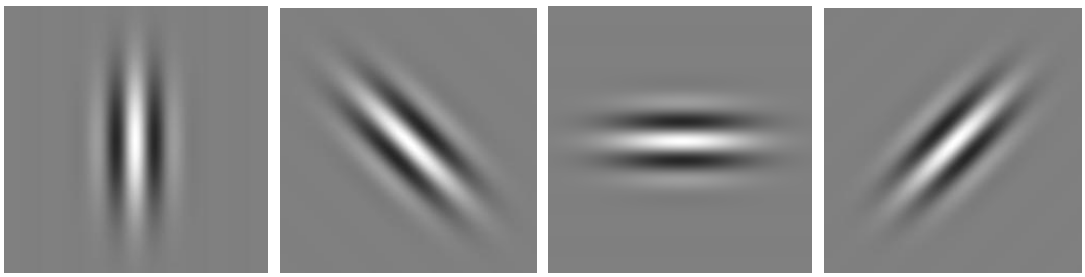


Figure 2.6 Examples of oriented Gabor functions rendered as luminance gradients, with from left to right 0, 45, 90, 135 degree orientations respectively.

These orientation functions are orthogonal to each other, and as such when convoluted with an image, encode the data with minimal mutual information between each of the functions. Using this scheme, Daugman has developed an efficient neural network encoding system for images, which is able to significantly reduce the entropy levels in an example image from a pixel-based high entropy representation to a low entropy Gabor representation [32]. As a corollary of this, the image can be compressed at varying levels by accessing weights within the Gabor wavelet representation. The representation can thus recover the image at varying image qualities from the network, by accessing an internal weighting scheme of the Gabor wavelet representation.

This research of Linkser and Daugman gives empirical support to an inherent efficiency in the visual cortex transformation of the image data on the retina to orientation specific Gabor-like functions. This may be considered an evolutionary adaptation for efficiency of visual tasks in survival scenarios [32]. This efficiency capability is discussed in Section 2.2.

Even though some of this cortical organisation is speculative, it does corroborate with psychophysical experimental results and lends physiological support to importance hierarchies within the set of basic visual features. Based upon this evidence, a simple feature importance hierarchy can be formed with motion at the top, followed by luminance and colour. In addition, psychophysical evidence indicates that the luminance-based features, such as motion, are the strongest attractors of attention [177]. This final point has repercussions with respects to the design of visual attention models, in particular, the relative weighted contributions of these features to the visual saliency of a spatial region. This is discussed further, with regards to visual attention, at the end of this chapter in Section 2.4.3.

2.1.3 The Generation and Execution of Eye Movements

The physiological systems described previously facilitate the generation of an internal mental image. The human eye must execute a series of movements in order

to maintain the stability of this image, so that the viewer has a temporally and spatially continuous visual perception of his/her surroundings.

Sharp and Phillips [145], Jacob [76] and Bruce and Green [20] list the following eye movements during normal HVS function:

- *Convergence* is the motion of both eyes relative to each other, usually for the generation of a single binocular image.
- *Rolling* of the eyes is an involuntary rotational motion around an axis passing through the fovea and the pupil, and is used for minor correction of the roll caused by head motion when viewing a scene.
- *Saccades* are a sudden, rapid (up to 700° per sec.) movement of the eyes. It takes approximately 100-300ms to initiate a saccade, and about 30-120ms to complete the saccade (depending upon the angle traversed). These motions are also ballistic, that is, they cannot be changed. The high speed of these saccades thus serves to minimise time spent in flight, as most of the time is spent fixating the chosen targets [136].
- *Pursuit motion* is a much smoother, slower movement than a saccade, and is enacted to maintain the foveal positioning of a moving object. Pursuit movements cannot be induced voluntarily, they require a moving object within the field of vision.
- *Nystagmus* is a pattern of eye movements that occur as a response to the turning of the head (acceleration detected by the inner-ear) or the viewing of a moving repetitive pattern (the train window phenomenon). It consists of a smooth pursuit motion in one direction to follow a position in the scene, followed by a fast motion in the opposite direction to select a new position;
- *Drift* and *microsaccades* involuntarily occur during fixations, they consist of slow drifts followed by very small saccades (microsaccades) that apparently have a drift correcting function.
- *Physiological nystagmus* is a high-frequency oscillation (tremor) of the eye that serves to continuously shift the image on the retina, thus

calling fresh retinal receptors into operation. If an image is artificially fixed on the retina, it disappears. This temporal attenuation of the input signal is countered by physiological nystagmus, where every point of the retinal image is moved the approximate distance between two adjacent foveal cones in 0.1 sec. Physiological nystagmus occurs during a fixation period, is involuntary and generally moves the eye less than 1°.

Of this list, saccadic eye movements are of special interest. Their ballistic nature indicates the possible existence of a mechanism for the calculation of saccadic eye movements. It is believed that a mechanism in the HVS facilitates the preparation of eye saccades to explore the most conspicuous, and therefore potentially important areas. This is considered to be the most effective method for the human visual system to explore any natural environment [45]. The ballistic nature of saccadic eye movements is simulated in psychological and computational visual attention models by calculating a master map of potential fixation locations, where the most conspicuous region becomes the next fixation region in the sequence of eye movements [155].

2.2 PSYCHOLOGICAL THEORIES OF VISUAL ATTENTION

Much psychovisual and psychophysical experimentation has been performed to elicit, from a perceptual standpoint, the image features that are important to humans. As a result, a number of models have been developed that seek to explain visual feature interrelationships, and their effects on viewer eye movements.

The human brain is limited in the amount of processing resources it can apply to the visual perception of its surroundings. As explained in previous sections, the acuity of the visual field degrades from high levels in the central foveal region to coarser levels in the visual field periphery. The centre of attention is where the object recognition capabilities of the HVS are at their optimum. Therefore, a search pattern is enacted when a human is presented with a complex scene, in order to bring the fovea to bear upon regions in the scene being viewed. Research has shown that these eye movements are attracted to areas of the screen that contain large amounts of

relevant information [21, 144, 184]. This search pattern enables these informative areas of the scene to be displayed across the fovea, to facilitate object identification processes.

The process of moving the focus of attention has been likened to a *searchlight* with an adjustable beam, or a zoom lens [48]. The searchlight is applied after a general impression has been gained of the scene. This general impression is generated by an early preattentive stage of the HVS, and is applied to the whole visual field as a parallel process. Soon after, the focus of visual attention is applied serially, to identify the objects in the most important areas of the scene [48].

2.2.1 Parallel and Serial Stages of Vision

Since the sixties, human vision researchers have divided human vision processes into early parallel, and later serial stages [113]. This has been concluded from visual experiments that involve search and identification of targets under controlled conditions. For example, a search task may be to find the line oriented at 45° in the example on the left in Figure 2.7. Note the ease of this search task. Whereas a similar search task on the right in Figure 2.7 is more difficult, due to the heterogeneous nature of the background distractors. The first search task was relatively easy, due to the visual phenomenon called *pop-out*. This pop-out occurs due to large differences between the visual features that make up the target, and its surrounding distractors.

Physiological correlates for this phenomenon are also indicated in cellular recordings of the cat striate cortex (These pop-out cells differ from the edge detectors uncovered by the work of Hubel and Weisel [65]). These results are expected to be analogous to human physiology, due to behavioural experiments indicating that cats perceive pop-out in a similar manner to primates [81].

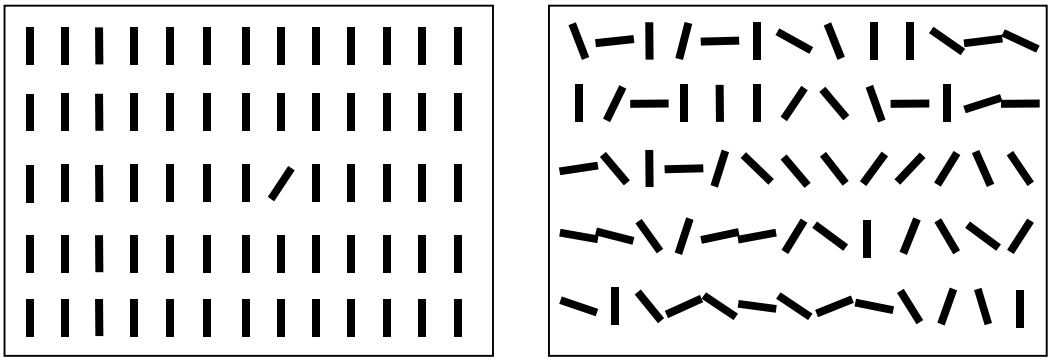


Figure 2.7 Examples of a parallel search task (left) and serial search task (right).

It is believed that experiments similar to the one presented in Figure 2.7 indicate two *preattentive* and *attentive* processes working within the application of visual search by the HVS. It is believed that an early preattentive stage of vision processes the scene for large feature differences and causes certain objects to pop-out. Research into the human visual system indicates that this preattentive stage has three main attributes [154]:

- Preattentive processing is unlimited in capacity, reaction time is unaffected by the number of distractors in an appropriate visual search with easily detectable targets.
- Preattentive processing is spatially parallel, operating simultaneously at various locations across the visual field. Earlier sections of this chapter detailed physiological constructs that are sensitive to scene features, for example, edges and motion. It is believed that the preattentive stage of human vision processes the scene in parallel for these features, alerting later stages of the HVS to their existence for further processing.
- Preattentive processing operates independently of conscious control. However, it is admitted that the interaction between preattentive and attentive processes has not been resolved fully.

The latter search task in Figure 2.7 was much harder, due to the heterogeneity of the surrounding distractors. This heterogeneity caused only a small spatial difference between the target and distractor visual features, inhibiting the ability of the preattentive process to highlight the target. The target in this case is a conjunction of

a number of features, and therefore requires top-down processes to aid the search for the target. Therefore, a slower serial process is executed, where the focus of attention is moved to every object in the scene, thereby applying the hyperacuity of the fovea to the task of finding the target matching the required features.

On the basis of results from visual search experimentation, similar to the examples in Figure 2.7, psychophysical researchers have categorised visual search tasks as *parallel* or *serial* [158]. Parallel search tasks are characterised by a relatively small response time gradient, with respect to the number of surrounding distractors. An example of a graph for a parallel search task is shown in Figure 2.8. Serial search tasks are characterised by a monotonically steep gradient, with respect to the number of surrounding distractors. A serial graph example is also shown in Figure 2.8.

The flat response times of parallel search tasks and the monotonic increasing response times of the serial search tasks are considered to be a definite border between the two categories. From the search dichotomy, it is then argued that there must be a similarly clear demarcation between the parallel preattentive and serial attentive stages of human vision [158]

However, some researchers disagree with this categorisation. Wolfe describes the strict dichotomy of parallel and serial stages of human vision as being a "useful, but potentially dangerous fiction" [177].

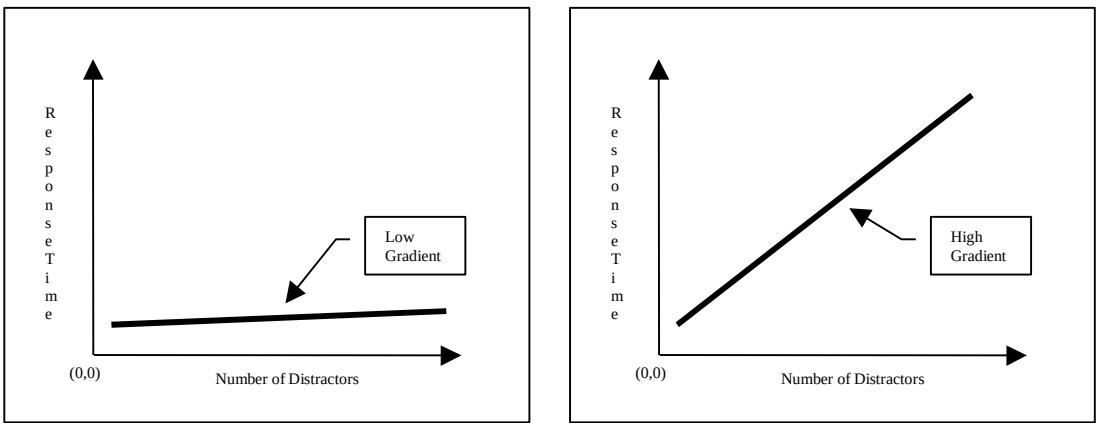


Figure 2.8 Examples of response time graphs for parallel search tasks (left) and serial search tasks (right).

Wolfe offers four reasons why he believes this strict dichotomy is incorrect:

- Inferring mechanisms from slopes is not that easy. The patterns of results can be produced by a variety of limited-capacity parallel methods, even mimicking the 2:1 slope that is considered characteristic of serial search.
- Strict serial search involves a number of unfounded assumptions. Firstly, a 2:1 serial slope prediction assumes with no-target searches that each potential target is only visited once. Secondly, they do not allow for errors, where the search is terminated before looking at all the possible targets. It is also assumed that only one item is checked at a time.
- The models assume a fixed dwell time for each item.

There is also a continuum of response times between the so-called lower pre-attentive stage to the higher-level object recognition tasks. There is no indication of a point where the graph of number of distractors × response times jumps discretely from having a flat to steep gradient. With these points in mind, the only general rule that can be gained from these graphs is that an increase in target/distractor similarity produces a concomitant increase in search times.

Wolfe [177] proposes a different nomenclature when describing HVS search mechanisms. He suggests the use of the terms *efficient search* and *inefficient search*. The terms are deliberately general in scope, to allow description of the continuous relationship between targets and distractors. This continuous relationship does not suggest the complete abandonment of the concepts of parallel and serial searches. What it does mean is that there is no evidence of a clean break between the two stages. The interconnections between visual units situated in the brain [92, 133] infer a similar continuous relationship. It would seem to contradict the physical structure of the brain to have a neat border between these two stages of human vision, and it is therefore likely that a more complex relationship exists between the two.

A number of psychovisual theories and models have been postulated to explain the relationship between these two stages of human vision and how eye movements are generated. Visual models seek to simulate two major influences on visual attention: *bottom-up* and *top-down*. Bottom-up influences consider visual perception to be a stimulus driven amalgam of lower-level scene components [158]. Top-down influences refer to the perception of the whole scene preceding that of its separate parts, including conscious task-oriented influences [82].

Much work has been carried out into developing bottom-up theories of visual attention, with two of the more popular theories being *Feature Integration Theory* (FIT) [155-159] and the *Guided Search Model* (GSM) [25, 176, 177, 179, 180]. FIT and Guided Search consider the early preattentive stage to be sensitive to low-level visual feature differences. These feature sensitive processes generate a mastermap of possible feature difference locations to be analysed by later attentive stages. This mastermap is then processed by the attentive stages of vision to move the fovea to informative scene areas. GSM improves on FIT by incorporating top-down influences into the process of segmenting the scene into potential targets for later examination by attentive processes.

Research has also uncovered influences on the deployment of eye movements and the response time of visual search experiments. These again can be broadly categorised as top-down and bottom-up [133]. Bottom-up influences emerge from the nature of the scene and its visual feature contents, while the top-down influences issue from higher cognitive areas of the brain [44], for example, previous experiences, visual search task nature etc.

In the light of FIT, there has been a large body of research devoted to uncovering a taxonomy of low-level scene features that are processed by the preattentive stage of human vision. As has been shown in Section 2.1 of this chapter, physiological evidence for these feature detectors has been found, but psychological experimentation has found a richer set of features which can be preattentively perceived: colour, motion, edges, contrast, curvature, depth cues and possibly even learnt features. Some work has been carried out on quantifying the effects of these

features upon visual search strategies used by the HVS. Characteristics explored so far include:

- global effects of feature differences upon local features differences [122, 123];
- surprisal probability, postulated to model the attention attracting capability of image changes [144];
- relative feature weights [128];
- interference effects have been noticed between luminance and hue, within the phenomenon of preattentive texture segmentation [22, 23].

What is lacking is a more complete quantification of feature interrelationships. Quantifying these visual feature interactions would facilitate the creation of computational FIT models, which would more closely simulate preattentive processes in human vision.

2.3 PSYCHOLOGICAL MODELS OF HUMAN VISUAL ATTENTION

There are two major influences in HVS models of visual attention; *bottom-up* and *top-down* [133]. The bottom-up vision model states that a *perceived image is sequentially formed by building up individual features of the image until the entire scene is recognised* [82]. With the top-down approach *an immediate overall impression, a gestalt of the entire scene, is initially formed with the individual features filled in later* [82].

Top-down models, for example, *Scan Path Theory*, have been proposed to explain eye movements in a natural scene [120, 121, 148-150]. Scan Path Theory proposes that eye movements in recognition tasks are stored internally as a loop of alternating features and eye movement instructions, drawn from the objects in the scene. It is believed that eye movements are closely coupled to the immediate task at hand [135]. For example, in natural language processing there has been recent evidence of eye-movements and fixations reflecting the instantaneous parsing of a spoken sentence [153]. Gale [48] notes that the scan path concept does have its critics.

These researchers recognise the existence of scan paths, but do not attribute any functionality to them.

The critics of bottom-up models suggest that the processing power needed to bind low-level features together for object recognition would be prohibitive [82]. A number of discoveries suggest otherwise. Firstly, the feature detectors in early stages of human vision indicate the detection of simple features at a preattentive stage in the HVS [65], [82]. Secondly, there is physiological and psychological evidence for object recognition processes being optimum in the centre of vision in the fovea, due to the high visual acuity available [133]. This allows the early feature detection stage to be low in physiological complexity, leaving more of the complexity that is related to high levels of visual acuity in the fovea, a relatively small region of the visual field. Support for the bottom-up theories also comes from the successful implementation of computational preattentive vision systems [109, 116, 160], thus refuting criticisms of the prohibitive computational complexity of such systems.

Upon reflection, the general consensus is that the top-down and bottom-up processes seem to interact in a complex task specific manner to influence visual search [44, 82, 183]. It would therefore be wise to consider the HVS to be an amalgam of such bottom-up and top-down processes in any future research. However, this project will concentrate on analysing and developing a model of the bottom-up components of human vision, in particular, bottom-up visual attention. This is due to the general, nontask-oriented nature of the application area intended for the visual attention model.

2.3.1 Feature Integration Theory

One of the first and more popular bottom-up theories about human vision is *Feature Integration Theory* (FIT) [155-159]. FIT forms the basis for a number of the computational models of human vision now being developed [107-109, 116-118, 160]. A schematic diagram of this theory is shown in Figure 2.9.

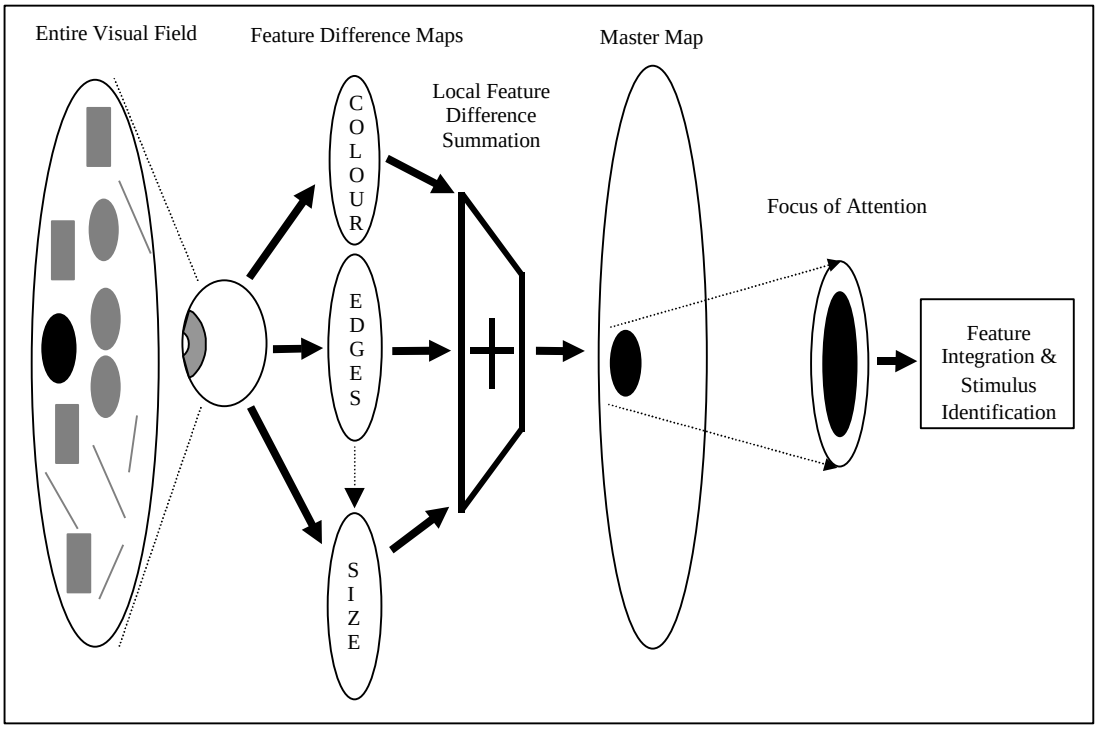


Figure 2.9 Diagram of Feature Integration Theory illustrating an example of pop-out with the black circle being unique in the hue feature dimension [53]

FIT postulates that features in a scene are registered *early, automatically and in parallel across the visual field, while objects are identified separately and only at a later stage, which requires focused attention* [158]. Groups of features, which are identified by functionally separate perceptual systems, are called *feature dimensions*, for example, colour and orientation. A feature is considered to take a particular value within this dimension, for example, red within the dimension of colour [158]. *Focal attention* utilises the increased visual acuity available within the fovea to integrate the features of the object for recognition purposes.

FIT considers that only differences in single features are processed preattentively. Conjunctions of feature differences are processed by the serial attentive mechanism. In a conjunction search the task becomes a top-down conscious search, due to the lack of target pop-out to guide the attention mechanism. The following experimental evidence supports the FIT model of visual attention:

- In visual search, single feature differences are detectable preattentively (for example, a red line amongst blue lines), while

conjunctions (for example, a 45 degree red line amongst 45 degree blue and horizontal reds) require focal attention to find.

- Preattentive *texture segregation* occurs due to spatial discontinuities in separable features and not conjunctions of features.
- *Illusory conjunctions*, predicted by FIT, are caused by visual overload or brief viewing. In the case of object recognition, for example, mistaken identification can occur due to lack of viewing time.
- Identity and location differentiation is indicated by the ability to detect the presence of feature differences without necessarily knowing the location, although it is easy to eventually home in on the location of the single feature differences. Conjunctions usually require attention to be identified.
- Unattended stimuli are registered at only the feature level. The interference they provide only comes from single feature differences, and not from conjunctions of a number of features.
- Anecdotal medical evidence is present in people with *visual agnosia*. Visual agnosia causes people to perceive objects as a number of separate features, finding it hard, or even impossible, to combine these features into a single object for recognition purposes [158].

The attentive feature integration process is considered to work in two ways. In the majority of cases the integration process occurs through the application of focal attention to identify the object. Otherwise, it is thought that top-down conscious processes are applied when the focus of attention cannot be deployed, due to overloading or brief exposure. In the latter, the rate of illusory conjunctions is high, but in familiar environments this can still be a useful technique for recognition. For example, during a game of sport a ball can be recognised more efficiently with this approach, due to the likelihood of a moving target being the ball in this environmental context.

Treisman has modified her feature integration theory to incorporate a separate map for each feature within each dimension, for example, red and green colour feature maps, or vertical and horizontal orientation feature maps. "Mutual inhibition within

each of these maps lessens the activation for elements that share the same feature value with many other elements" [157]. This allows some feature conjunction searches to exhibit the efficient parallel processing characteristics of single feature difference searches [159]. It is postulated that this inhibition of distractors facilitates the discarding of distractors in visual search tasks, where the target is known, allowing top-down influence on the segmentation of the visual field [163].

The FIT model, attractive in its simplicity and application to a wide range of visual input tasks, fails to explain a number of human vision phenomena. The main problem is using FIT to explain the speed with which humans are able to perceive natural visual phenomena. These natural scenes are made up of objects with conjunctions of many features, and yet we are able to perceive many objects at near parallel search speeds. Wolfe has modified and extended FIT to incorporate mechanisms to explain these contradictory visual search results, naming it the Guided Search Model (GSM)[179].

2.3.2 Guided Search

Wolfe [25, 176-181] has produced evidence of visual search tasks involving conjunctions that indicate the subjects were able to preattentively reduce the search to objects of single colours or shape. FIT would expect a viewer to be reduced to serial search in these circumstances. Wolfe provides a modification to FIT to account for these differences.

GSM differs by its modelling of interactions carried out between the early preattentive stage of vision and the later serial stage. This interaction allows for conjunction searches to be accomplished at a faster speed than the serial search suggested by FIT. This is due to the parallel stage guiding the serial stage by grouping the scene into areas of similar colour or lines of similar orientation. In the GSM the parallel stage provides more detailed processing than suggested in FIT, by processing each feature dimension to highlight areas that may contain the target. This information is not perfect, so each search task is not considered parallel, but the division of the screen into like feature dimensions does produce more efficient serial search tasks. This partially explains the almost continuous change between parallel

and serial search tasks noted in Section 2.2.1. It should be noted that this model addition applies to directed viewing tasks, with prescribed targets consisting of feature conjunctions. However, it is worth analysing, due to its insight into possible models of preattentive feature processes for other viewing scenarios.

For each feature dimension the parallel stage identifies those elements that are closest to the target value for that dimension, and that differ from the other elements in the display. Information from each feature dimension is summed for each element to produce an overall activation map, which has a value for each location representing the likelihood that the position is a target. There is also some physiological evidence for an activation map in the HVS that registers particularly salient regions of the viewing field [54]. When the serial stage is ready to start processing a new element, it chooses the one with the highest activation in the activation map. Only the serial stage is capable of initiating a response. Until the target is processed by the serial stage, no response will be made. Once an element has been processed by the serial stage and found to be a non-target, it is eliminated from further consideration.

Figure 2.10 depicts how the GSM accounts for efficient conjunction search tasks. The parallel stage has partitioned the scene into areas based on colour and orientation. This is summed together to form an activation map providing information to the serial search stage on likely locations to investigate for targets. This improves the efficiency of the search task, but not to the level of a parallel search task.

A simulation of the GSM has been tested, producing satisfactory results for the vision search task categories of feature search, conjunction search and triple conjunction search. The guided search model produces similar response times to humans performing the same tasks, in the process exceeding the capabilities of FIT [179].

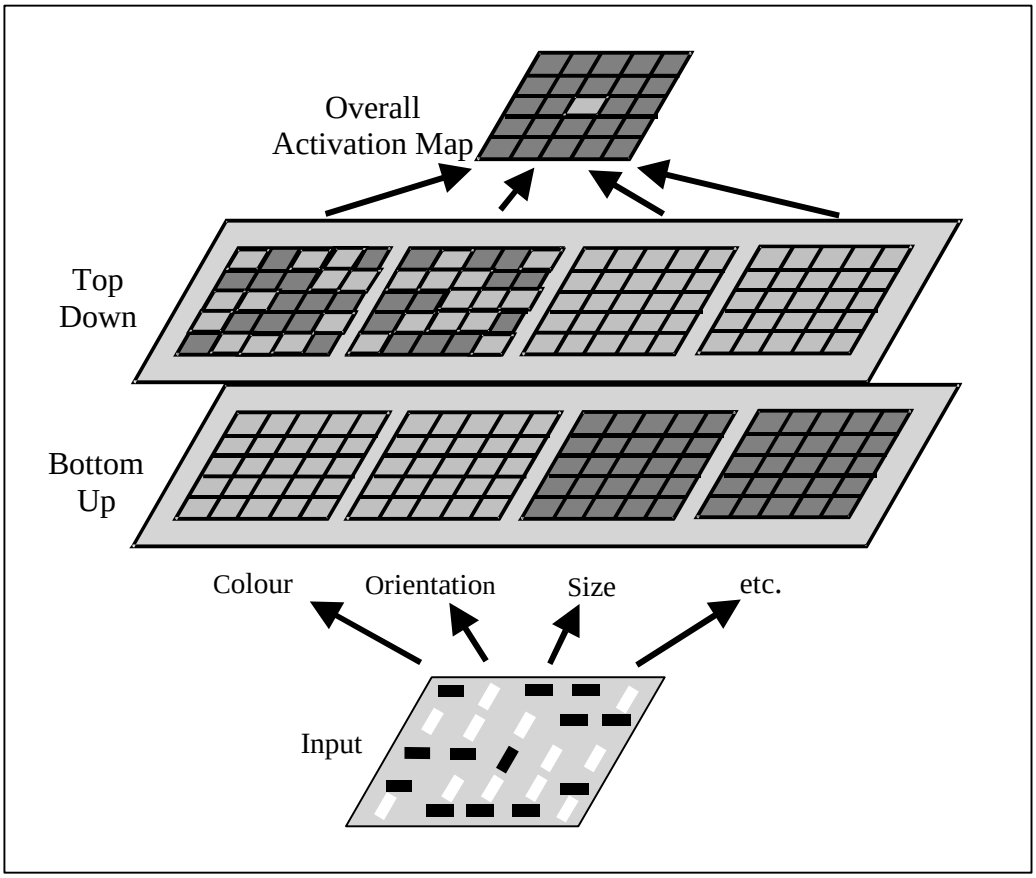


Figure 2.10 Example of explanation by Guided Search for near parallel conjunction searches. The top-down feature maps contain local activations for specific orientation and colour features. These are summed together to produce the final activation map, guiding the viewer to the conjunction target [25].

The GSM also accounts for subject to subject variations with the addition of random noise to the parallel stage. This is normally distributed across the final activation map, while the variance is a subject by subject based parameter. This reduces the reliability of the parallel stage, leaving the decision for movement to the serial stage, mimicking some of the imperfections in the decision making process of the parallel stage.

Guided Search seeks to account for central tendencies of eye movements in visual search by introducing a simulated fovea by a complex log transformation of centre surround inputs to mimic the V1 area of the visual cortex. This provides a basis for eye movement generation in the GSM, although they state that the model is quite far from mimicking eye movements with real images [180].

However, guided search cannot explain the factors of distance and search asymmetries arising from experimentation in visual search. The spatial relationships within a scene are important, as nearby objects have more of an effect on each other than far away objects. Search asymmetries have also been discovered, for example, it is harder to find a short line surrounded by long lines than a long line surrounded by short lines. These asymmetries have yet to be accounted for by both FIT and GSM.

2.3.3 Texton Theory

Another popular human vision model is *Texton Theory*, proposed by Julesz [77-80]. His research has dealt with the ability of the HVS to instantaneously segregate dissimilar textures.

Three heuristics define the main structure of Texton Theory:

- Human vision operates in two distinct modes of preattentive and attentive vision (the distinct separation of parallel and serial stages of human vision is a moot point, see Section 2.2.1).
- The preattentive stage of the HVS is sensitive to texture components called *textons*. These textons consist of elongated blobs (rectangles, ellipses, line segments with specific colour, angular orientations, widths and lengths), line-segment terminators and crossings of line segments.
- Preattentive vision directs attentive vision to the locations where differences in the density (number) of textons occur, but ignores the positional relationships between textons.

Physiological support for elongated blobs being a texton comes from evidence that monkeys have retinal areas specifically for the recognition of elongated blobs [65], [80]. Julesz concedes that the texton elongated blobs and the blob receptors discovered in the retina of monkeys are not isomorphic. This is due, in part, to the psychological nature of this work, his findings do not necessarily map directly to neurophysiological discoveries.

Experiments performed by Julesz show that humans are able to preattentively (within 150msec [79]) discriminate textures, if there is a difference in the textons, or a difference in the first order statistic of the texture. The first order statistic is simply the probability of a randomly thrown dot landing on or off an arbitrary texture unit. This translates to, for example, a difference in the size of the texture units.

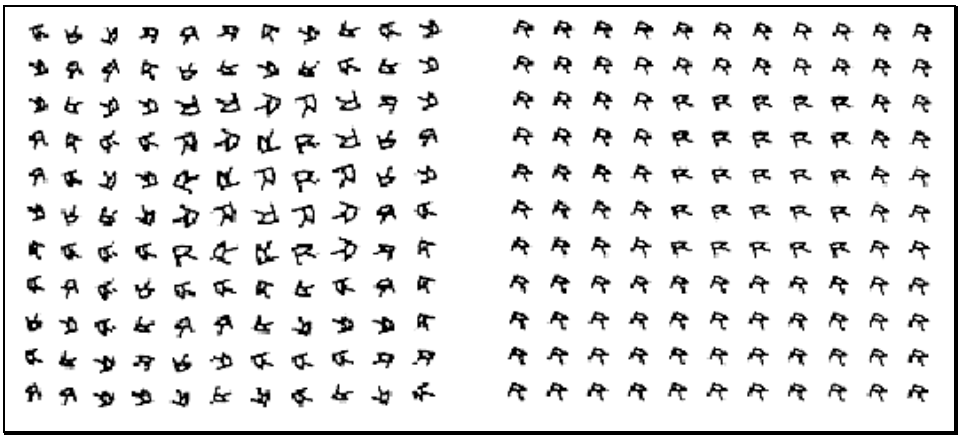


Figure 2.11 Examples of a preattentively separable texture with different first-order statistics and differences in element size (left), and a preattentively distinguishable texture with different second-order statistics and different element orientations (right) [77].

Effortless texture recognition can occur with the same first order statistics, but different second order statistics. The second order statistic is the probability of a 2gon (dipole or needle) being randomly thrown on the texture and having one or both of its ends land on or off a texture unit. For example, this may be a similarity in size, but a difference in orientation of the texture units (refer to the diagram on the right in Figure 2.11).

Textures with differences in third or higher order statistics, having same first and second order statistics, are not processed by the pre-attentive stage. These texture differences are processed by the serial attentive stage.

Julesz proposes that the output of local feature analysers is linearly averaged over the whole image. This reasoning comes from the observation that *areas covered by rectangles containing only single Xs, when linearly averaged, have the same*

contribution as rectangles containing zero or two Xs when these occur with 0.5 probability and one assumes that the simple cortical units are themselves linear [77].

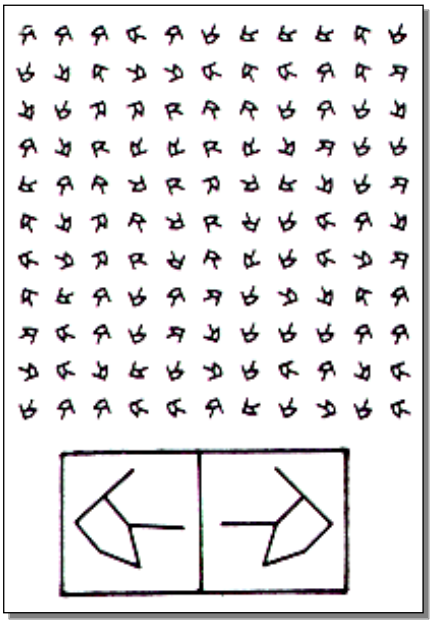


Figure 2.12 A preattentively indistinguishable texture pair with identical second-order, but different third and higher-order statistics composed of randomly thrown similar micropatterns and their mirror images [77].

Related to the linear averaging is the observation that only the differences in the densities of the textons cause pre-attentive texture division. The positional relationships of the textons remain unnoticed in the pre-attentive phase [80].

Texton Theory provides a simple and robust statistical model for the preattentive segregation of texture areas. Its constructs are similar to FIT and Guided Search, with a parallel stage providing guidance to a later serial stage. The fundamental textons are similar to the fundamental features in FIT [158] and Guided Search [179]. It has to be noted, never the less, that it is restricted in its applications due to the artificial nature of the textures, and that few, if any, computational models of human vision use Texton Theory as a basis. However, some of the general principles of texture density may still benefit a visual attention model [152]. This is due to the possibility of developing computationally efficient methods for ascertaining texture density using measures other than the textons described in the work of Julesz.

2.3.4 Stimulus Similarity

Duncan and Humphreys [38] suggest a general theory based upon target/non-target similarity, whereby search efficiency decreases with increasing target/non-target similarity and increases with decreasing target/non-target similarity. For example, a blue target circle will stand out against a background of red non-target circles. However, if the target circle is gradually turned to red, then the efficiency of searching for the target circle progressively decreases. Stimulus similarity theory contains three major components:

- a preattentive parallel stage of perceptual description, producing a structured hierarchical representation of the input across the visual field at several levels of spatial scale;
- a process of selection by matching input descriptions against an internal template of the information needed in current behaviour;
- a process entering selected information into *Visual Short-Term Memory* (VSTM) through a relatively weighted competitive process.

Their theory explains some anomalous results from conjunction search experimentation, not explained by FIT, drawn from their own experimentation and a review of visual search literature. They state that the evidence indicates feature search and conjunction search are essentially the same process (similar to conclusions by Wolfe [177]). It has to be noted that the complexity of the features in the preattentive stage of Similarity Theory precludes it from being used in entirety for the computational modelling of visual attention.

2.3.5 Comparison of FIT, GSM, Texton Theory and Similarity Theory

FIT [158] is the theoretical basis for a number of computational vision models [50, 107, 116]. On the other hand, the GSM [177] is similar, and has been implemented on a computer and tested with simple psychophysical experimental stimuli. Yet the GSM is not mentioned nearly as often as a basis for vision models, possibly due to its inclusion of top-down factors, which are hard to model in many computing application areas. The major relevant difference of GSM to FIT is the additive effect

of the feature maps to the saliency map, which produces a high excitation level where there is a possible target, thus explaining some fast feature conjunction searches.

Treisman offers more evidence for her distractor feature map inhibition from experimentation and neurophysiology [159]. She does note, however, that the evidence in her favour is thin. Still, FIT remains the leading psychological theory for visual search/attention.

Although the Texton Theory model neatly describes texture segmentation, it is of little utility due to the incompatibility of Textons [79] as visual features with the features used in this project. While the complexity of the similarity model [38] of preattentive structures is too loosely defined for application to this project, their general rule of target and distractor similarity is a useful qualitative description of the pop-out of certain visual field regions, due to the closeness of the features of a region to its surrounds. This qualitative rule will be applied to the different feature dimensions considered in this project.

The feature processing that occurs in the early stages of both Guided Search and FIT is relevant to the project, as these principles can be exploited to calculate the visual importance of the regions in a rendered scene by processing the differences in visual attributes of a scene description, for example, colour, size etc. Furthermore, the concept of a master activation map summing feature differences has definite utility within computational modelling applications [180]. This activation map, or importance map as it has also been termed [97], represents the level of visual saliency of regions within the image. Therefore, the activation map may be used to modulate directly the spatial quality of an image in progressive rendering techniques. This concept is developed more fully in Chapter 3, Chapter 4 and Chapter 5.

2.4 INFLUENCES ON EYE MOVEMENTS

Regardless of the viewing conditions, experimentation has uncovered strong correlations between the eye movements of viewers while viewing natural images

[21, 184]. From this research, general observations can be made about the viewing of natural images [126]:

- The distribution of fixations over a viewed image is not even, but is skewed towards particular regions within an image. The regions fixated are correlated strongly across different viewers, within similar viewing conditions and task scenarios for both still [21, 96, 184] and motion images [151, 174]. This correlation is especially strong for motion images, with up to 90% correlation between viewers of motion videos, with the regions fixated being only 6% of the area of the image [151].
- Viewers when freely regarding images tend to regard certain informative, different or unusual regions of an image due to cognitive reasons [2] (for example incongruous objects) [93], or by the presence of particularly informative contour features [8, 96]. In general, these regions stand out from the background of the image due to contrasting features, and therefore attract the attention of the viewer. Regions may also attract attention due to the presence of high edge concentrations, indicating detailed information about the scene [144].
- With an unlimited viewing time, certain salient regions are repeatedly regarded by the viewer [149, 184]. In addition, research indicates that there is a strong tendency for a person to regard the interesting regions of an image in the same order, through repeated viewing. These observations form the basis of Scan Path Theory [120, 121]. However, this is only for a particular viewer, the correlation of fixation ordering is not so strong across different viewers, despite the strong correlation of overall fixation locations.

These fixations on regions within a viewed image have a more complex relationship to the application of visual attention than would be expected. Researchers have discovered that people are able to regard stimuli in the periphery without orienting their eye to the stimuli [132]. This phenomenon is known as *covert* attention, and is characterised by having less of an ability to perform search tasks due to the degraded

acuity in the peripheral regions of the HVS. When searching natural scenes there is a need for high levels of visual acuity to recognised detailed objects, thus requiring movement and fixation of the fovea upon objects in the viewing field.

The movement of the eye to attend to an object is known as *overt* attention. Related to overt attention is the concept known as the *mandatory shift hypothesis*. This hypothesis states that while attention may move without a related eye movement, an eye movement will always be preceded by a relocation of visual attention [63, 147].

From the evidence presented, it can be concluded that the correlation of eye movements across different viewers lends support for a similar correlation of visual attention across different viewers regarding the same images. Therefore, it is reasonable to also assume support for comparable fixation generation processes across viewers. In order to model these processes effectively, there is a need to identify more precisely the various influences on eye movements. Broadly speaking, these can be divided into *top-down* and *bottom-up* categories.

Top-down influences are products of higher order cognitive systems in the brain, and tend to be under attentional control. Bottom-up influences are reflexive responses to the image being examined, and tend not to be under attentional control [133]. As has been stated before, the top-down and bottom-up systems tend to interact in a complex manner, as deduced from the continuity of search task response times from parallel to serial [177]. Yantis and Jonides have evidence that the attention grabbing effect of peripheral feature onsets depends on the amount of top-down focus exerted on other areas of the image. The more the person is concentrating on another area of the scene, the more likely the onset of a stimulus will be ignored, showing that conscious top-down control of reactions to feature-based stimuli do occur [183]. An example of this interaction is the process of searching for a particular person in a crowd. The top-down processes visualise the characteristics to look for, this in turn influences the bottom-up processes which alert a person to some of these features [43]. An example of this is the searching of a crowd of people for a person who is wearing a red peaked cap. This goes some of the way to explaining the ability of humans to perform efficient visual search in highly complex environments.

It must be noted however, that the discrimination capabilities of the HVS change depending on the scene being presented. Recent work has uncovered the ability of the HVS to discriminate the presence of quite complex objects within a natural scene, without the application of attention [88]. It is noted that this does not hold for the simple scenes used in psychophysical tests, for example, discrimination between different alphabetic letters. From experiments, it appears that the HVS can categorise a natural scene within 27msec, which is easily under the time taken to orient attention. Control experiments also established that the effects were not induced by training, and that the effect does not work with the simple stimuli of psychophysical experiments.

These results are explained by a number of hypotheses. First, the HVS is given, at a very early stage, a gist of the contents of the scene [88]–also supported by other research [177], and referred to later on in this thesis. This gist is used by the HVS to very quickly understand the presented scene. Secondly, it has been noted that the visual cortex responds more strongly and more efficiently to the presentation of a natural scene than when presented with one of the artificial scenes used in psychophysical research [17]. Thus, the visual cortex is possibly organised for efficient understanding of the sparse informative components of natural scenes as compared to artificial scenes. Overall this indicates that attention is not necessarily the gate to higher levels of consciousness, and that, in the case of natural scenes, the later object recognition systems of the human brain have some access to the visual field without the use of attention. This indicates a more complex relationship between the attentive and non-attentive components of the HVS than has been previously thought.

2.4.1 Top-down Influences

As stated before, these influences are generally attentional and can be related to task nature, previous experience, context effects and physiological effects.

Yarbus [184] researched various aspects of eye movements during the viewing of stationary images. He noted that subjects would regard only certain regions of an

image, despite having an unrestricted viewing time. Experiments showed that eye movements were remarkably consistent across viewers, especially when the context and the viewing task were the same. Results also indicated that visual attention was drawn towards humans in the test images, especially faces and hands. Yarbus believed that this was due to the faces containing useful information about the context of the image, for example, the emotional state of the character. The horizon in an image was shown to be another strong attractor, possibly due to training effects induced by humans constantly having a surrounding horizon when outdoors, as useful information is often present along horizons.

However, the eye movements of the viewer were modified by the nature of the viewing task, with the viewer concentrating on regions that provided relevant information for the task at hand. During the viewing of an image titled An Unexpected Visitor (see Figure 2.13), the viewer was set the task of discerning the material circumstances of the people in the image. For this task the viewer took particular notice of the clothing of the women and the furniture in the room, whereas for the task of estimating the age of the people in the scene, the viewer concentrated on the faces of the people (see Figure 2.14) [184].

Figure 2.13 An Unexpected Visitor, a test image used by Yarbus in his eye movement experiments [184].

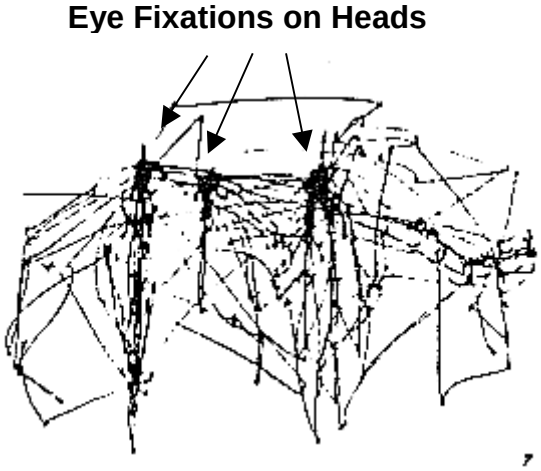


Figure 2.14 An example of the eye movements of one subject during free (uninstructed) viewing of the image in Figure 2.13 for three minutes [184]. Note the concentration of fixations upon the faces of the major people in the scene, as highlighted by the arrows.

Senders supports these task related conclusions, stating that the movements of the point of regard could be deterministic or statistical [144]. Deterministic processes look at the place with the greatest uncertainty, whereas statistical processes choose the place to look with a probability proportional to the level of uncertainty presented there. This uncertainty is essentially related to the task at hand. For example, a fighter pilot will examine different instruments depending upon whether he/she is landing, taking off or pursuing an enemy.

Previous experiences become a major factor in HVS search patterns. Gale notes that experienced industrial inspectors will fixate their gaze automatically to areas where targets are most likely to occur [48]. He later states that experienced radiologists will enact a wider search pattern than a naive subject. Giving pre-task instructions to the viewer can also induce an effect. The search pattern is modified to include likely locations for objects relevant to the task. Senders notes that with more training a subject will enact more fixations on a similar scene, with less time being spent on each fixation, indicating less analysis of scene regions due to prior knowledge of their contents [143].

Context effects are related to previous experience. Researchers have performed experiments in which incongruous items are inserted into an image, such as a pay

phone in a lounge room scene. These items were found to influence eye movements, due to top-down processes being surprised at the appearance of the incongruous object [2, 8].

Some success has been achieved in modelling task-related top-down search factors [121, 135, 144]. It has to be said that some of the above top-down effects are difficult to quantify in any computational vision model, but this does not diminish their effect upon human eye movements. For example, the experience factor has repercussions for computational vision models. As the user grows used to the environment, they may change their search method and diverge from a programmed average search model. Many of these top-down issues will only be effectively dealt with when science has a more effective understanding of higher-order brain processes.

Senders generalises the top-down influences on the patterns of eye movements with the statement, *…the eye moves from one POR (Point Of Regard) to another in order to minimise the total relevant uncertainty of the observer about the scene* [144]. From this statement it can be inferred that top-down modelling of visual attention should incorporate task-oriented components to mimic the visual search behaviour of human viewers.

2.4.2 Bottom-up Influences

Bottom-up influences are relatively easier to model than top-down influences, as they are based upon better understood physiological mechanisms in the HVS. Bottom-up influences proceed from the image presented to the HVS. They include such phenomena as priming, inhibition of return and feature-based pop-out.

Maljkovic and Nakayama describe a process of subconscious location priming that occurs with colour and spatial frequency features [101]. These features can cause an involuntary fixation if the feature has caused pop-out previously at the same location. This effect is thought to last for 30 seconds as a decaying memory of that pop-out inducing feature.

Inhibition of return subconsciously occurs when a stimulus attracts attention to a particular location. The HVS will tend to fixate the location once, and then ignore the stimulus until it changes. Kwak and Egeth [85] in their experimentation note that inhibition of return occurs only with location, no other feature will cause this phenomenon to occur.

The pop-out phenomenon occurs in the preattentive stage of the HVS, through a large local difference in image features [123]. Researchers in the fields of psychology and physiology have defined a group of image features that facilitate this pop-out phenomenon [65, 79, 92, 101, 122, 123, 155, 179, 180]. Wolfe, in his review of vision research, distils the list of fundamental preattentive features to: colour, orientation, curvature, vernier offset, size, motion, shape, depth cues and gloss [177]. He suggests that this list is by no means complete and that some of the features are still questionable.

What many researchers in the field agree on is the lack of effective models to quantify the influences of preattentive features and their interrelationships [127, 144, 177]. Some work has been carried out on the quantification of feature pop-out using subjective evaluation of the pop-out level against a varying background [122, 123], and other probabilistic models have been proposed to explain top-down effects [144]. However, present computational models of preattentive vision use arbitrary formula to quantify the contribution of each image feature [58, 99, 116, 126]. Computational preattentive feature models will be more thoroughly investigated in Chapter 3. Despite the lack of quantitative models, there is a body of descriptive knowledge characterising the attentional effects of visual features. This descriptive information for edges, hue, depth, size, location and motion is now explained.

Edges are considered one of the fundamental features within an image [103]. With regards to eye movements, an absolute high density of edges attracts attention, while a low density of edges does not attract attention as much [144]. Another major influence is the sharpness of the edges. Sharper edges are more likely to attract visual attention than blurred edges [144]. Experiments have also shown that pop-out occurs with a local difference in orientation between the edge(s) and the local

surrounding edges, modulated by the variability of the orientation in the background [122, 123, 177]. Pop-out also occurs due to the uniqueness of the target amongst distractors [177].

Hue pop-out occurs with a local difference in hue between one region and another. This is either suppressed or enhanced by the variability of the hues in the background [122, 123]. That is, if the hue difference is dissimilar to hue differences in other regions, then the region stands out strongly. If the region hue difference is similar to those surrounding it, then the mutual inhibition incurred by the other differences suppress the pop-out of the region [177]. Opponent colours are considered to cause the highest contrast, for example, red against green and blue against yellow, as well as complementary colours, for example, orange and blue [33, 69]. However, these results must be approached with caution, as they have not been rigorously tested under laboratory conditions. It is sufficient to say, though, that hue category differences do aid the phenomenon of pop-out.

With luminance, some models suggest a pop-out influence similar to colour [177]. They also suggest that white and black are achromatic opponent colours, so they exhibit a high contrast value when placed next to each other [33, 69].

The perception of depth is caused by a number of feature differences: edge orientation cues, texture induced slant, shading effects and binocular disparity [180]. However, it should be noted that although ocular information is available from the LGN onwards, some of the above depth features are composed of other features. This evidence indicates that depth perception occurs later in the visual system, and so its preattentive nature remains a paradox. Wolfe suggests that the preattentive stage of human vision might extend into higher visual processing areas than the visual cortex [180].

Size is considered by many visual attention models to contribute to the importance of an object, however it is still inconclusive as to whether it contributes to the psychophysical phenomenon of pop-out [177]. Some models use an absolute measure of size as a model of the importance of an object [58, 127]. With visual

search, in a similar fashion to hue, the task becomes efficient if the relative local differences in size are large enough, although no mention is made about the variability of background distractors. It should be noted that size is related to spatial frequency, as the change in size of an object changes the local spatial frequency of the contrast [177].

Humans tend to first look towards the central 25% of a computer screen, due to expectation of a properly framed scene, for example, news broadcast sequences [127]. Location is listed by Wolfe as a separate visual feature [177], and is considered unique in its effects on priming of pop-out [101].

Motion is considered an important influence on eye movements. Stelmach notes the high correlation between viewer eye-movements during the viewing of various television scenes [151]. With motion intensive video (the experiment used a hockey game) 90% of the viewers looked at the major cluster of eye positions. This compares to 40% for low motion intensity scenes (the experiment used a weather report). Motion it seems is a strong attractor of attention for all viewers.

2.4.3 Feature Hierarchies

The relevant literature indicates a substantial amount of research performed on the identification and characterisation of low-level visual features. However, little work has been carried out into the interaction between these features. What has been performed is still at essentially a qualitative level.

Evidence suggests that feature relationships can change due to top-down experience effects. Koch [83] notes that the weighting of different visual features is slightly plastic. With training, certain features will more strongly attract attention than others [38]. Experiments with macaques by Bichot et al. [9] show a training effect on saccades, indicating feature weight plasticity related to the tasks performed. Bichot et al. interpret their results as being a process for establishing habits or skills. They propose that top-level establishment of skills can influence the weightings of low-level features. The authors stress, however, that this does not change the general

relationships between features, for example, motion will still strongly attract attention, only the amplitude of its effect may change with training.

This plasticity inhibits the discovery of a fixed set of weights describing the importance of each of the features to the visual system. However, the results of Bichot et al present a case for application specific feature hierarchies, base upon the features important to the task at hand.

Despite the lack of a hard and fast weighting scheme for features, there is evidence for a gross ordering of features into a hierarchy. The issue of what constitutes a visual feature complicates the ordering, and whether the list previously described is not a collection of sub-features that are preattentively discerned. So far, the only work carried out has been with motion, hue and luminance.

Motion is generally regarded as the most conspicuous feature. Most models consider motion, and its related feature temporal change, to be the most attractive visual feature [128, 151, 185]. The physiological evidence detailed in Section 2.1.1 confirms this assumption.

Lohse has identified Hue as being more attractive to people than achromatic information [94]. Yet in the case of texture segmentation, it has been found that luminance information is dominant over hue [22, 23, 60]. The research by Lohse involved eye movement tracking of viewers of advertising in yellow pages. Two factors may confound these results. The hue luminance values were not controlled for, thus the attracting ability of the hue could be enhanced or suppressed by the brightness of the hue. Secondly, the values of the local differences and mutual inhibition were not analysed. The achromatic information in the image will have inhibited the effects of any luminance contrast in the image. There would need to be a comparison with equiluminant hues and a similar spatial distribution of hue/luminance differences.

The work of Callaghan [22, 23] and Healy [60] is performed under these conditions, and shows luminance to be a dominating factor in texture segmentation. Their results show that a luminance texture difference can interfere with texture

segmentation by hue in the same spatial region. With texture segmentation considered to be an influence on visual attention, it can be considered that this is supporting evidence for luminance to be considered above hue in a proposed importance hierarchy. The physiological evidence noted in Section 2.1.2 further supports the dominance of luminance as a visual feature.

Other empirical work has been performed by Osberger [128] to ascertain the weighted contribution of image features towards region importance in an image. The experiments involved the tracking of the eye movements of 14 subjects while viewing a series of 136 still and 46 moving images. The weighting factors were calculated based upon a weighted average across segmented regions making up 10, 20, 30 and 40 percent of the image area. Each of the features within the regions had a correlation value derived based upon how many fixations occurred within the region. His results produced a hierarchy of feature weightings, with the following features being in order from smallest to largest weighting: image location, foreground/background differentiation, skin colour, shape, luminance contrast, hue and size.

The derived weights are probably not so important to this thesis due to implementation differences. On the other hand, the rough hierarchy is significant in the light of the physiological results described in Section 2.1.2. The hierarchy adds further evidence supporting luminance and foreground/background features (derived from luminance effects) as being greater in influence than hue values. The position of region location at the top of the hierarchy may be considered a result of viewer training, due to the continual viewing of properly framed images with the subject of the scene being in the centre.

Allowing for the aforementioned plastic nature of feature influences, what has been presented here is support for a hierarchy among visual features in their influence on bottom-up visual attention. What is also significant is that both the physiological and psychophysical schools of vision research have evidence to support similar hierarchies.

2.5 DISCUSSION

This chapter has described and analysed major physiological and psychological components involved in the concept of visual attention.

Physiological evidence has been presented for neural mechanisms that respond to visual features within the viewing field. Furthermore, evidence was shown supporting the concept of an importance map summarising the visual field, highlighting those regions worthy of further investigation by object recognition processes concentrated in the centre of the visual field. Physiological evidence supporting a hierarchy of feature importance was also presented.

Supporting evidence was then presented, from psychovisual and psychophysical experimentation, of preattentive and attentive processes in human visual attention, along with theories describing the relationships between them. Models of visual attention involving these two processes have been described and compared in some detail. The chapter was then completed with an investigation of both the top-down and bottom-up influences on visual search, including a list of preattentive image features. The lack of preattentive feature interrelationship models was identified. Some evidence for a gross hierarchy of motion, luminance and hue was presented, from a psychophysical perspective.

From this analysis, a new approach to rendering can be devised using the principles previously outlined. This approach will consist of two major components: the development of a visual attention module and the development of efficient rendering techniques that use the newly developed visual attention module.

The state of the art in visual importance modelling is reviewed in Chapter 3. From this review a new fuzzy logic-based bottom-up visual attention module is described within Chapter 4. The module is based upon the observations in this chapter, including: the attracting potential of local spatial differences in visual features, an appropriate list of features to be analysed in an image, the global effects of other feature differences in an image and possible feature importance hierarchies. Two

models have been developed to simulate the effects of contours and feature differences between regions. These differences are stored in an importance map, derived from similar master maps within the visual attention models described within this chapter. The importance map quantifies the relative visual importance of regions within the image for later use by a new adaptive rendering approach.

In Chapter 5, adaptive and progressive rendering techniques have been modified to accommodate the new visual attention module. This is used to direct and control the level of refinement applied to regions within an image. Furthermore, this approach is extended to texture mapping in Chapter 6, where the sampling of textures for antialiasing purposes is modulated by the importance of the region in the image. Chapter 7 then extends the visual importance module developed in Chapter 3, to accommodate motion and abrupt onset feature models. This is then applied to the task of efficient computer animation by deriving an importance value for the motion of segmented regions within the scene. Chapter 8 then describes the evaluation of the rendering system via both objective and subjective measures of image quality.

Chapter 3

Previous Computational Models of Visual Importance

In this chapter, a literature review of the state of the art in computational visual importance modelling is performed. Previous work is roughly divided into multiresolution and region-based models, which are derived from related psychological theories of human vision. The strengths and weaknesses of both approaches are described, and relevant aspects are drawn together to form a theoretical framework for the design of the fuzzy logic-based visual importance system detailed in Chapter 4. In addition, some basic principles of fuzzy logic control are illustrated.

Fuzzy logic control systems are particularly suited to this project, due to the uncertain nature of eye movement measurement and prediction. Eye movement prediction is an imprecise process due to the following factors:

- eye movements may or may not indicate the presence of visual attention, due to the phenomenon known as *covert attention*, where attention may not be correlated with the centre of the visual field [132];
- eye movement data is imprecise, due to the large error range of measurement instruments (0.5-1.0° viewing angle) [3, 146];
- eye movement data is noisy, making it hard to extract accurate fixation information [37, 128];
- bottom-up feature driven visual attention is based upon the fleeting experience of a naïve viewer regarding a scene, this disregards effects induced by task-oriented processes [48];
- few experiments have been carried out in order to determine a quantitative model of bottom-up influences on eye movements, leaving general rules as the staple descriptions of researchers [144].

Despite the above problems, the empirical evidence for consistent eye fixation locations across viewers regarding the same scene has been established [21, 151, 184]. It is precisely this sort of ill-specified control problem that fuzzy logic is able to handle, making it well suited to the project at hand.

Fuzzy logic also lends itself to this thesis from an engineering perspective. The use of fuzzy logic is widespread and is thus supported by a large body of research [7, 182]. Fuzzy logic also facilitates the engineering and tuning of rule sets in an easy and intuitive manner. Finally, the nature of region-based importance models lends itself to being modelled by a rule-based fuzzy logic system, due to the comparison of regions of similar features using comparison rules [35, 59]. These points add weight to the selection of fuzzy logic techniques to model the relationships between visual features and their perceived visual importance.

A number of application areas for computational visual attention models have been explored within the relevant literature. Active vision systems incorporating models of visual attention have been developed for robotics, in order to facilitate autonomous target acquisition for further analysis [135, 170, 190]. Scientific visualisation applications have also benefited from the application of visual attention models to assist with the segregation and understanding of multidimensional data [60, 139]. For example, two different groups of data may be represented using separate shapes that form preattentively discernable boundaries. Progressive image transmission also benefits from models of visual importance, to facilitate the transmission of informative regions of an image first [161, 188]. Image compression applications have also used models of visual attention to modify levels of compression to match the visual importance of regions in the image [98, 128, 189]. Finally, image synthesis systems have reaped similar benefits by modulating the pixel sampling rates by the visual importance of the pixel [186].

The ongoing research into modelling visual attention has been divided between multiresolution [83, 107] and region-based approaches [97, 127, 189].

The multiresolution approaches follow from multiresolution theories of human vision. These theories seek to explain experimental phenomena requiring the HVS to contain components sensitive to a range of spatial frequencies [166]. The physiological evidence of simple cells in the visual cortex also supports multiresolution vision theories, with receptive fields tuned to different spatial frequencies [64].

Region-based models derive support from psychophysical evidence for people regarding *objects* and not *locations* [103]. Furthermore, there is mounting evidence indicating that objects are formed from underlying features in the preattentive stage [177]. The preattentive arrangement of the viewing field into regions of *bundled* features as objects provides further support for the argument of region feature differences attracting attention, and not the feature difference locations themselves [178]. The following sections go on to analyse these two groups of models in more detail.

3.1 MULTIRESOLUTION VISUAL ATTENTION MODELS

Multiresolution visual attention systems are based on multiresolution theories of the human visual system and the related image processing methodologies [166]. These HVS-based models incorporate *feature detectors*, *saliency maps* and *models of human attention* drawn from relevant psychophysical and psychovisual research [83, 107, 160].

One of the earliest multiresolution attention models was developed by Koch et al. [83, 115-118], and further extended by Itti [70-74]. The system processes an image for four features: intensity, hue, edge orientations and movement, at multiple scales. The edges are detected by convolving the original image with *Gabor*[2] patches, due to their ability to simulate HVS pattern sensitivity [137]. The hues are processed using an HVS-based colour opponency system. HVS-like centre-surround receptive fields are simulated by using locations in the lowest scale map as the centre, with a larger region in a higher resolution feature map as the surround.

[2] These Gabor functions are described in Section 2.1.2.

The model developed by Koch et al. does not seek to completely model the interactions of visual features as they are combined into the saliency map. Simple linear summation is used to combine the feature maps together into the saliency map. Weights modelling the contribution of the feature maps are equal, with the exception of an arbitrary five times greater weight for motion. No empirical basis is given for the magnitude five [115, 116]. However, the additional weight added to motion is justified due to evidence for motion being a strong, if not the strongest, attention capturing scene feature [144, 151].

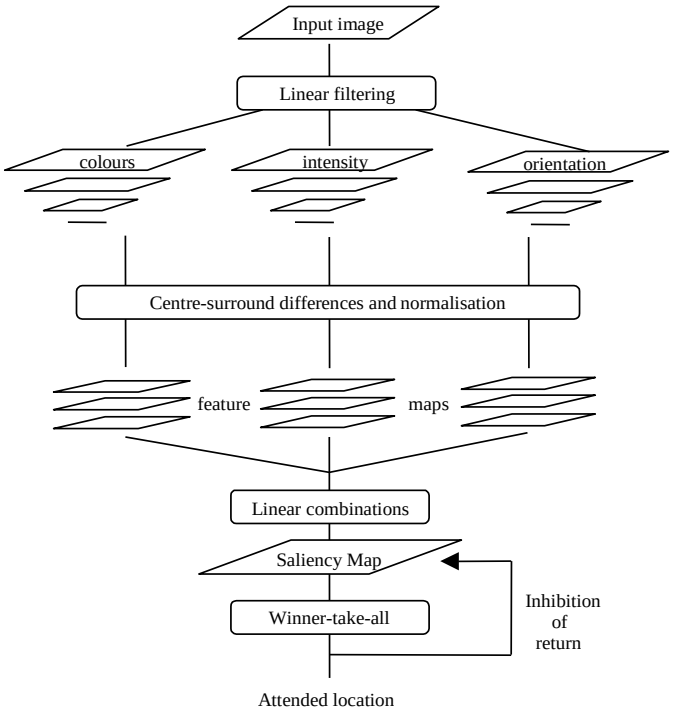


Figure 3.1 Diagram of Koch visual attention system architecture [74].

This saliency map is processed by a *winner-take-all* neural network, which fixates visual attention onto the highest peak for a prescribed interval of time. The fixated location in the map is then inhibited, with the next fixation of attention being the next highest peak remaining in the saliency map. An overview of the architecture of this system is shown in Figure 3.1.

Itti extends the work of Koch et al. by simulating the visual cortex phenomenon of lateral inhibition of surrounding receptive fields [70-74]. Perceptually, this produces a suppression effect upon the differences that invoke a saliency effect upon a particular region. For example, a bright target object in an image consisting of bright objects will appear less conspicuous due to the activation surrounding the target object. The following steps describes the *normalisation* process [70]:

1. Normalise all the feature maps to the same dynamic range, in order to eliminate across-modality amplitude differences due to dissimilar feature extraction mechanisms;
2. For each map, find its global maximum M and the average $\overline{m}$ of all the other local maxima;
3. Globally multiply each map by:

$$(M - \overline{m})^2 . \tag{3.1}$$

This process effectively suppresses maps with uniform activation, but enhances those maps with an *odd man out* activation peak. Evaluation of the method is performed in a qualitative manner. They report that the model is able to generate fixations and saccades similar to those generated by humans. These statements are made without any empirical evidence, although this has been stated as being a part of future work.

A similar approach developed by Milanese processes an image for contours (length, orientation, contrast, curvature) and regions (size, perimeter size, elongation, average grey level) [107-109]. These attributes each have an associated *retinotopic*[3] feature map, indicating the location of the listed features within the viewing field. The feature maps are then processed by a feature difference function or histogram analysis. To calculate any conspicuous areas the output is loaded into k conspicuity maps, one for each attribute. These pixel-based conspicuity maps are integrated together using a non-linear energy relaxation function, to produce a small coherent

[3] The term retinotopic refers to the coordinate system in the feature map being spatially related to a position on the retina. Therefore, a change of position in the feature map produces a related movement in the retina.

set of visually important regions. The system then thresholds the saliency map to produce a binary mask, indicating probable regions of interest.

The work has also been extended to incorporate motion effects using multiresolution image representations, which calculate temporal derivatives for pixels derived from frame to frame differences [108]. These are then converted into a motion mask and further refined through low to high resolutions to form a convex hull surrounding the moving region in the images. This alert mask is integrated with a saliency mask, determined for static components of an image, to form an overall bottom-up set of targets within the scene. An overview of the static saliency architecture of the system is shown in Figure 3.2.

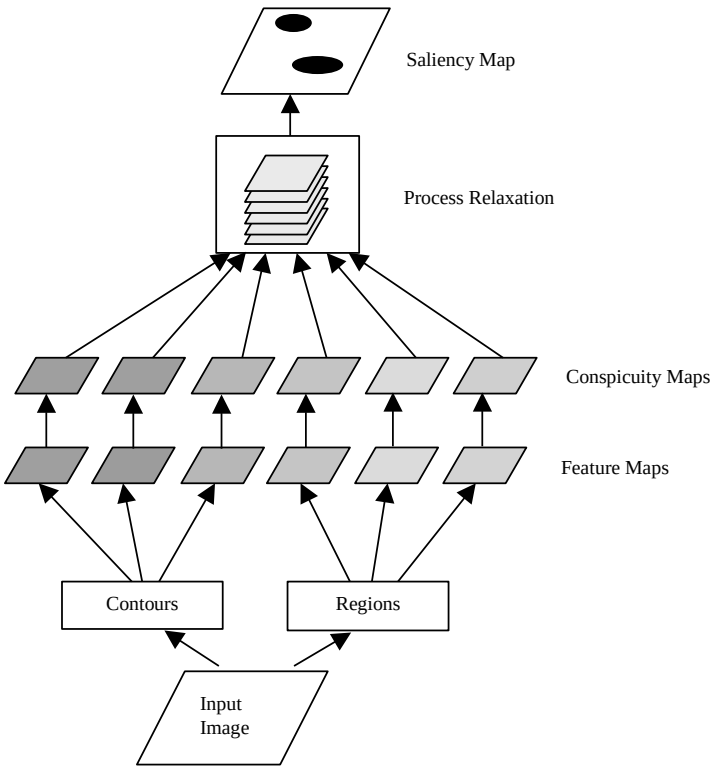


Figure 3.2 Diagram of Milanese visual attention system architecture [109].

Milanese states that the final image areas selected are the same as those selected by the HVS, but does not offer evidence of a thorough evaluation of the system. The system does not process colour, although this may be due to the application area being industrial inspection, and not human vision simulation. The non-linear

relaxation feature integration method is not based on HVS characteristics, but instead is designed to meet arbitrary criteria for creation of a small number of regions of interest, again due to its application area being industrial inspection. However, his feature integration method does offer an improvement over simple linear averaging of the conspicuity maps. The relaxation method filters out the small spatial differences by modelling competition between conspicuity maps, suppressing the saliency of local image differences within highly activated conspicuity maps. Finally, the system does not produce a graded importance map for each area of the image. Instead, it produces a binary mask indicating areas for the later attention system to process. This masking of the input image is appropriate for the industrial inspection applications listed, but is not useful for graded control of image synthesis techniques according to visual importance.

Another multiresolution model of attention is a progressive image display system by Zabrodsky and Peleg, which presents the most important regions of an image first-giving a recognisable image in a short space of time [188]. The system encodes an image using a multiresolution *Gaussian pyramid*, a *Laplacian pyramid* (to encode spatial frequency differences between the levels of the Gaussian pyramid) and a *Difference of Laplacian* E*nergy pyramid* (to encode the motion differences between two frames for a video sequence). A quadtree then encodes a path through the image pyramid from lowest to highest resolution. The quadtree traversal dictates the order of the image regions sent over the transmission link, with the most visually interesting areas being sent first. An attention function is used to determine spatial contrast and temporal changes, and an inhibition term, to prevent the focus of attention from being chosen at the same resolution all the time. The nodes of the tree are then sent to the receiving end with at first a low spatial frequency general outline, followed by the rest of the image sent in order of visual saliency. Therefore, the image is recognisable at an early stage of the transmission. As the system uses a specifically defined model of attention, and bases information on previously sent data, the receiver can be sent the sample value without a need for its location to be sent. It should be noted that the attention model used in this system is simple, being only based upon contrast levels and temporal changes in video sequences, and is therefore not sensitive to other features, such as size, hue etc.

While these multiresolution approaches do give a biologically plausible model of HVS bottom-up visual attention, they do not model the object-based viewing bias within the HVS. This behavioural evidence suggests that a more useful model of visual attention may be region or object-based. As well as theoretical considerations, there are issues specific to the application area in which the model is to be developed.

Progressive rendering is, as explained earlier, a process of gradually refining a coarse rendering of the scene to a predefined quality limit. Multiresolution methods require an image that has at least been defined to the level of a pixel, to accommodate the need for the range of spatial resolutions within the model [185]. However, a region-based model can still cope with an unrefined image. A region-based importance approach can progressively guide the refinement of a synthetic scene, due to the low computational overhead of determining the importance of regions. Furthermore, it can update its region segmentation and importance values in a computationally efficient manner, to match the refinement level of the image. Due to these two points, it has been considered that a region-based visual importance approach has more utility. Therefore, an analysis of previous region-based approaches to visual attention is required in order to aid the design of a new visual importance model.

3.2 REGION-BASED VISUAL IMPORTANCE MODELS

There are a number of applications of region-based visual attention models to the area of image processing. In region-based approaches, the image is segmented into regions and a visual importance value assigned to the segmented regions, based upon the presence of visual features within the region. Importance mapping of images can improve the efficiency of image compression systems, as a higher degree of quantisation can be assigned to the areas of little visual interest [98, 127, 189]. Other application areas include progressive transmission of still [161] and video [102] images over low bandwidth links, where the most visually important areas of an image are assigned a higher priority in the transmission scheme. These applications improve the efficiency of their respective methodologies, while still maximising the perceptual image quality.

Maeder and Pham [99] have developed a colour importance system, which incorporates both global and local factors into its calculations. The system calculates a colour importance value *IMP* at every position (*i, j*). The global factors calculated include: the probability of a colour occurring in the image $Pr_g(C)$, the probability that a colour belongs to a particular colour group K_c $Pr(C|K_c)$, the variability of the colours in the image V_g, the variability amongst the colour groups V_{GK} and the variability within a particular group $V_G|K_c$. The local factors are calculated from an $m \times n$ block or segmented region, including the probability of colour value C being in an $m \times n$ block at pixel location (i, j) $Pr_{l(mxn)}(C)(i,j)$, or segmented region $Pr_{LR}(C)(i,j)$. Local cluster and variability factors are defined in similar ways for the regions and blocks.

Maeder [97] has also developed an approach using 3×3 pixel blocks to detect local edge strength, contrast and variance of pixel intensities, for grey scale intensities. These are incorporated into an importance map by choosing the minimum of the 3 factors at each pixel. The saliency map is divided into 8×8 blocks, which contain the average combined importance value within that block. This saliency map is then quantised into 4 values for the DCT matrix in JPEG compression, allowing high rates of image compression to be achieved while still preserving perceptual quality. The integration method is arbitrary in nature, with no psychophysical evidence for the use of the minimum of the three factors. However, Maeder notes that even with minimal tuning, the method is able to compress to quality levels of 40%, without the same level of perceptual distortion caused by the normal JPEG algorithm.

The success of previous models using both global and local importance calculations supports the incorporation of similar local and global effects into the visual model developed in this thesis. The global variability measures will be used in a somewhat modified form in this project. However, the colour importance system uses the colour probability distribution in the image for the modification of image processing methods, such as edge detection or colour quantisation. The importance model

developed in Chapter 4 is based upon local differences in hue, and not on absolute hue distributions throughout the scene.

Another system by Osberger [126-128] extends the importance map concept further by segmenting the image into regions based upon a variance analysis of the pixels, using a *recursive split and merge* technique [140]. These regions are then allocated an importance value, based upon the following criteria drawn from visual features in the image:

- contrast importance-the luminance difference between the region and its surrounds;
- contrast importance-the hue difference between the regions and its surrounds in CIE L*u*v* colour space[4];
- size importance-the ratio of region area to 1% of image area, with a fall off in the importance at a certain threshold;
- shape importance-the ratio of border pixels to total pixels, this value being highest for long thin edge-like regions;
- location importance-the ratio of image centre region (25%) pixels to region area;
- background importance–the ratio of border pixels to half the image border pixels, to account for foreground/background separation.

The importance values for the above features are squared and then summed into a saliency map, which is then normalised to a maximum value of 1.0. The size feature is processed in an absolute manner as a ratio of the region to 1% of the image area. However, psychophysical experimentation suggests that the differences in size around a region enhance visual importance, and so should be taken into account [177].

The model developed by Osberger, in a similar manner to Itti and Koch [70], uses normalisation techniques to account for activation within the hue and luminance

[4] CIE L*u*v* is a spatially deformed colour space where equivalent spatial displacements are as close as possible to perceptually equivalent colour differences [66].

feature dimensions. However, the exact technique is not mentioned in the references. A number of exponents have been added as parameters to the calculation of luminance and hue differences to enable more control of the effects of each feature dimension. Furthermore, the model has been extended to include temporal importance, with the addition of camera movement adaptation and the use of adaptive thresholds for motion [128].

Zhao et al. [58, 59, 68, 189] have also developed a region-based model for visual importance calculation based on the following features:

- size-the number of pixels in the region dictates its importance due to absolute size;
- position-regions closer to the centre of the screen are considered to be more visually important;
- compactness (boundary length to area ratio)-this ratio is greatest for a circle, considered most attention grabbing;
- border connection-high level of border connection suggests a background region;
- hue (CIE Lab and HSV)-the region will pop-out if its hue is different to surrounding regions;
- luminance-similar to hue pop-out, is based upon luminance differences to surrounding regions;
- saturation, the importance of this feature is taken from the average saturation of the pixels in the region in question - the more saturated the region, the more important it is to the human observer.

The visual importance of these regions is calculated from the above criteria and then passed through a fuzzy rule set. The fuzzy rule parameters are tuned using a neural net module [189], with training data gained from experiments with university students. The students compared the segmentation of an image with a normal image and allocated 3 levels: very important, somewhat important and not important to the regions segmented.

These results are used as training data for a neural network, which controls the weights used within the rule-base. The weight for each importance rule is constructed from the weights for each feature. This weighting scheme allows for interactions between the features, as well as an absolute weight for each rule in the system. It is assumed that, along with other fuzzy rule parameters, the neural network dictates the values of the weights.

The fuzzy feature importance system is then applied to image compression, where the JPEG quantisation matrix is biased so that the least visually interesting areas are highly compressed. Their evaluation of the correlation of the areas segmented by the system, with regards to human attention, is questionable due to its imprecise nature. The evaluation began with subjects choosing the most important areas of the images. They then measured the correlation of the image segmentation with the areas chosen by the viewers. It would be more appropriate to use either an eye movement evaluation method, or a more controlled experimental methodology to evaluate the correlation between areas of attention fixation and importance values predicted by the system.

The previous system could be improved with regards to the treatment of size and global effects. As with the system designed by Osberger [128], size is an absolute stimulus, whereas psychophysical literature suggests that the attractiveness of areas can be attributed to local differences in size around the object in question [177]. Also, the system only considers local calculations and does not consider global effects induced by the feature dimensions.

Tsumara et al. [161] describe a progressive transmission technique that uses gaze areas to order components of progressive images. Similar to work by Zabrodsky [188], the system sends a general low spatial frequency impression of the image first, followed by the higher frequency details of the image. However, this system does not use feature processing to determine the next region of the image to be sent. Instead, the order of transmission is deduced from eye tracking experiments performed with the image. The authors suggest that feature processing will be a

component of later stages of their research, due to the need to predict the gaze positions for an arbitrary image, obviating the need to perform time consuming eye movement experimentation.

The review has provided information about the utility of using particular sets of features for importance calculations. Certain features such as hue, luminance, size, contours and image location are ubiquitous among the sets of features processed. Furthermore, the use of texture measures, in particular contour concentrations, has been shown to be useful for region-based importance models [152] [79, 144]. Therefore, a similar subset of features has been chosen for this project: hue, luminance, size, contours, location, background/foreground and contour concentrations.

The newly developed system will differ from other region-based models primarily in the complete treatment of feature differences. Some of the features used in other models have been absolute in nature. In the system to be developed in this chapter, the conspicuousness of regions will be completely based upon feature differences, due to the experimental evidence shown in the psychophysical literature reviewed in Chapter 2 [122, 123, 177]. This will incorporate effects defined by feature differences that absolute processes cannot simulate. In addition, the importance values will be based upon a threshold concept, whereby the regions become prominent due to the local difference in a feature being above the level of background distractors. No gradient will be applied-the region will stand out due to suprathreshold differences and will be labelled as being of high importance in that feature dimension [122, 123].

The new importance module for the image synthesis approach will use fuzzy logic, for the reasons detailed at the start of this chapter. The following section provides a primer to the concepts used in applying rule-based fuzzy logic theory to control systems.

3.3 FUZZY CONTROL SYSTEM BACKGROUND

Fuzzy systems handle imprecise data by *fuzzifying* the truth of a logical statement[5]. Instead of the crisp values of True or False being discrete values of 1 and 0 respectively, the values are fuzzified to be any value between zero and one inclusive. Fuzzification of the truth defines the term *Degree Of Fulfillment* (DOF), which represents the truth level of a fuzzy term. The membership function for a fuzzy variable represents this degree of fulfilment value for the *universe of discourse* of the fuzzified term. An example membership function is shown in Figure 3.3 for the fuzzy variables *TempC* (temperature in degrees Celsius) and *Comfort*, with regards to water temperature.

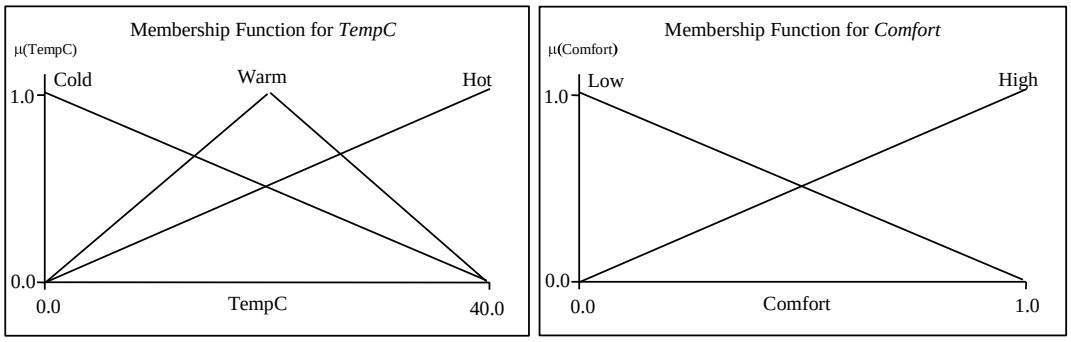


Figure 3.3 Example antecedent membership function variable *TempC*, and consequent function *Comfort*.

The DOF of the function at any temperature $\mu(TempC)$, is the degree of truth attributable to the relevant fuzzified term. The shapes of the functions can be gained from a number of sources, such as: expert knowledge, histogram data or mathematical formulae. The power of these membership functions is drawn from their ability to encode, in a manner similar to human reasoning, a relationship between a linguistic term and the numerical input value. It can be seen in Figure 3.3, as the temperature approaches zero the term *Cold* becomes more fulfilled, while the *Hot* term becomes less fulfilled. These fuzzified terms are used in inference statements, constituting the rule-base of the fuzzy control system. Some examples follow for the above membership functions:

[5] Unless otherwise noted, the principles in this section are taken from Berkan and Trubatch [7], or Yager and Filev [182].

IF *TempC* is *Cold* THEN *Comfort* is *Low*;
IF *TempC* is *Hot* THEN *Comfort* is *Low*;
IF *TempC* is *Warm* THEN *Comfort* is *High*.

The rule bases enable a fuzzy system engineer to form inferences about the state of the system for control purposes. The left part of the rule is titled the *antecedent*, whilst the right side of the term is known as the *consequent*. What inferentially links the two sides together are the processes of *implication*, *aggregation* and *defuzzification*, considered in turn below.

Implication is the process of mapping the antecedent fulfilment values to the consequent variable space, illustrated by Figure 3.4.

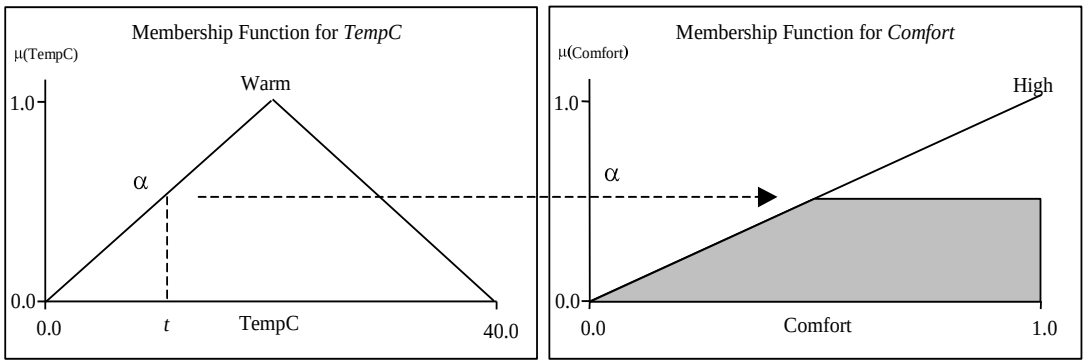


Figure 3.4 Illustration of the implication process that maps the antecedent value on the left to the consequent value on the right.

This implication process effects the logical connection between the antecedent and consequent. The temperature *t* is fuzzified into a DOF of α for the antecedent function *Warm*. The DOF value α is then mapped, via an implication operator (commonly the *minimum* activation value), to a fuzzy set in the consequent universe of discourse (represented by the grey region on the right hand side of Figure 3.4).

The process of aggregation takes place when the consequent fuzzy set is determined for the output functions. Here, the consequent values of the output fuzzy terms are combined for each of the rules in the system. The combination of these fulfillment

levels forms a fuzzy set representing the state of the system, as shown in Figure 3.5. From this fuzzified representation of the truth levels of the rules within the system, a defuzzified crisp value *D* must be derived for application to the control scenario in question. The method chosen is usually based upon the desired behaviour of a system. The defuzzified value is a crisp (single) value characterising the output fuzzy set generated by the rule-base of the system. A commonly used defuzzification method involves a weighted average of the values within the consequent variables, such as the *centre of area*. *Continuity* and *efficiency* issues are often the major criteria regarding the choice of defuzzification methodologies. Continuity prevents the system from jumping too quickly from one value to another, which may cause instabilities in the process the fuzzy model is seeking to control. Efficiency issues are paramount for real-time systems, in order to respond in a timely manner to changes in the process being controlled.

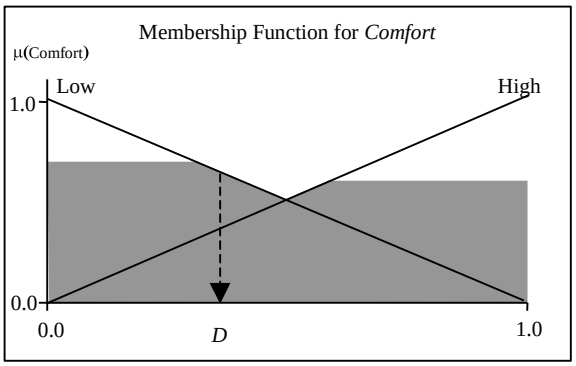


Figure 3.5 Example of aggregated fuzzy system (darkened regions) and defuzzification of system to produce a crisp value *D*.

All these aspects have been considered in detail for the design of the fuzzy logic importance model shown in Chapter 4.

3.4 DISCUSSION

This chapter has sought to investigate the state of the art in computational visual attention modelling. Multiresolution and region-based models of visual attention were presented and analysed to discern any deficiencies in their approach. From the evidence presented, it was concluded that a region-based approach using a fuzzy logic control system was the best technique for modelling the visual importance of

objects within a scene. The chapter then concluded with a primer devoted to relevant principles of rule-based fuzzy logic control, to provide a theoretical background to the design of the visual importance model presented in Chapter 4.

Chapter 4

A New Model of Visual Importance for Efficient Image Synthesis

In this chapter, a bottom-up model of visual importance is developed from principles discovered in the relevant literature. Chapter 2 has shown that there is a strong correlation between differences in spatial and temporal features and the application of visual attention. From this analysis arises the opportunity of computationally modelling this process and deriving useful applications. Principles derived from the literature review and encapsulated within the model to be presented in this chapter include: visual features used, feature difference effects, importance map concepts and inter-feature relationships.

The newly developed model seeks to utilise the rules gleaned from the psychophysical literature by using fuzzy logic to ascertain the relative visual importance of segmented visual regions. This model improves existing fuzzy models of visual importance by more completely modelling feature differences, by allowing for global effects of features within the visual field and by the inclusion of textural effects from contour attributes within the regions.

The visual importance model is eventually applied, in later chapters of this thesis, to the task of adaptive pixel sampling in ray tracing. The model contains two modules that have been developed to facilitate this process. One module guides the progressive refinement process, directing the rendering system to refine the most important contours first. The other module controls the pixel supersampling process, to refine areas regarded by the viewer.

Chapter 4 details the development of this model in the following manner. Sections 4.1 and 4.2 explain how the information from Chapter 2 and Chapter 3 has been incorporated into the overall design, with regards to: the selection of features to be modelled, the general approach to model development and the fuzzy logic

implication process. Section 4.3 then concludes with a summary of the achievements of this chapter.

The intended goal of the development of this new visual attention model is its application to the problem of image synthesis efficiency. The main conjecture of this thesis is that the application of a visual importance model to the control of ray tracing sampling rates will reap large computational cost savings, while minimising the perceptual distortion of the image.

The progressive ray tracing approach can be divided up into two main stages. The early stage is the refinement of low spatial frequency details, down to the size of a pixel. The contour importance module developed here orders the refinement process for contours, refining the most visually important contours first.

The second stage is the supersampling of pixels, for antialiasing purposes. At this stage the pixel is subdivided into quadrants for further sampling. This supersampling process may be fixed or adaptive. Fixed supersampling performs a spatially even pixel subdivision, whereas adaptive supersampling performs an asymmetric pixel subdivision that is sensitive to contours. The second region-based importance module controls the sampling at the subpixel level in order to reap savings in the sampling rate. In the following sections the two modules are explicated in detail, along with the reasoning in their development strategy.

4.1 CONTOUR IMPORTANCE MODULE

Much work has been carried out into image components that facilitate and enhance object recognition within a viewed scene. Evidence has been uncovered for the visual importance of contours with deep concavity, especially those that coterminate, forming junctions [4, 103]. The removal of these junctions from a contour image severely inhibits the ability of a viewer to recognise the image [10]. These terminations have therefore been postulated as being important to the processes of object recognition. It has also been postulated that these coterminations are strong attractors of visual attention [11], which coincides with the informative regions regarded by viewers [21, 184]. Other authors have noted the attention attracting

capability of the concentrations of contours in an image, and the attraction ability of the strength on the contrast forming the contour [144]. Finally, contours have been used in multiresolution visual attention systems, by looking for orientation differences that may attract the attention of the viewer [74, 107]. This model differs by using contours within a region-based framework.

This model has been designed to order and accelerate the refinement of contours in a progressive rendering system by the visual importance calculated according to fuzzy logic rules. In essence, a contour is important if it is strong (high contrasting) and highly curved.

The model is local in nature, drawing its information from an 8×8 pixel regular subdivision of the scene, using a quadtree. A contour analysis algorithm, called the Discontinuity Coherence Map (DCM), is used to ascertain contour information within the subdivisions [57]. The base DCM processing algorithm does not order or evaluate the importance of the detected contours. The newly developed model therefore introduces a more complex form of contour assessment, which both orders the refinement of the contours and accelerates the refinement of those contours that are considered important to image understanding.

The use of a region-based refinement model is rejected, due to the priming effect of the scene changes being expected to attract the attention of the viewer by default [101]. Therefore, the problem of progressive rendering is considered to be more of an image understanding or recognition problem than a bottom-up region-based visual importance issue[6]. Recognition is based more on the importance of certain contours in the scene rather than on the bottom-up visual importance of the regions within the scene [10]. As the module is based upon contour importance, three contour variables are obtained from the DCM analysis of the subdivision to facilitate this importance assessment: contrast, curvature and density. These variables and their contribution to the importance model are now described in detail.

[6] In hindsight this is not the case, as the regions considered most useful to image recognition and quality correlate to regions most salient to the HVS. This issue is discussed further in Chapter 9 under the heading of improvements to the present system.

Contrast is measured from the difference in the internal maximum and minimum subdivision luminance value, drawn from the Weber effect [166]. From this principle a simple perceptual model of the contrast in an image subdivision is derived. The contrast is evaluated using the following equation [110]:

$$C_{sub} = \Delta lum / (maxLum + minLum) \quad (4.1)$$

where:

C_{sub} is the luminance contrast value for the subdivision;
Δlum is the luminance contrast in the subdivision;
maxLum is the maximum luminance value within the subdivision;
minLum is the minimum luminance value within the subdivision.

If the luminance contrast exceeds an empirically derived values of 0.05 [57], then a contour is considered to exist in the subdivision and so further processing is performed.

Density is measured as a contour count within the subdivision, derived from the number of contour crossings contained within the boundary of the subdivision. This is calculated by thresholding the luminance values on the subdivision boundary into binary values 0 and 1. The threshold value is derived by the following equation from Guo [57]:

$$t_b = 0.5 \times (b_1 + d_1) \quad (4.2)$$

where:

t_b is the threshold to be used to binarise the subdivision values;
b_1 is the highest sample luminance value for the subdivision;
d_1 is the lowest sample luminance value for the subdivision.

The boundary is then analysed to find crossing points, which are the value transition locations. These represent an approximation of the number of contours within the subdivision (refer to Figure 4.1).

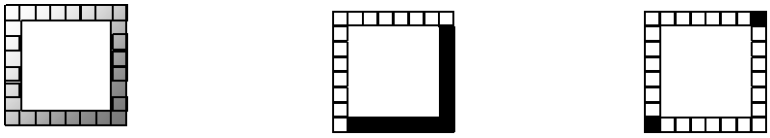


Figure 4.1 Illustration of the DCM method for ascertaining the number of contour crossing points within the boundary of a subdivision [57]. The left square shows the samples taken along the boundary of a subdivision. The middle square has been thresholded to show the transition points that are highlighted in the right square.

Curvature is calculated as the difference in radians between the tangents at the transition points in the subdivision (refer to Figure 4.2).

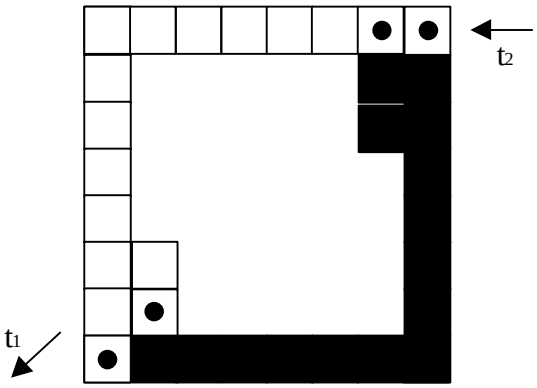


Figure 4.2 Illustration of the DCM method for ascertaining contour curvature within a subdivision [57]. Tangents at each transition point are calculated by making more samples inside, near the transition points. The tangents t1 and t2 are then found by matching the values at the transition points marked by the pixels marked with a black circle. The difference in the tangent angles is used as a curvature estimate.

The location of the subdivision is also included into the model, due to results indicating that viewers regard the centre 24% of the screen more than any other region of the image [41, 174, 181].

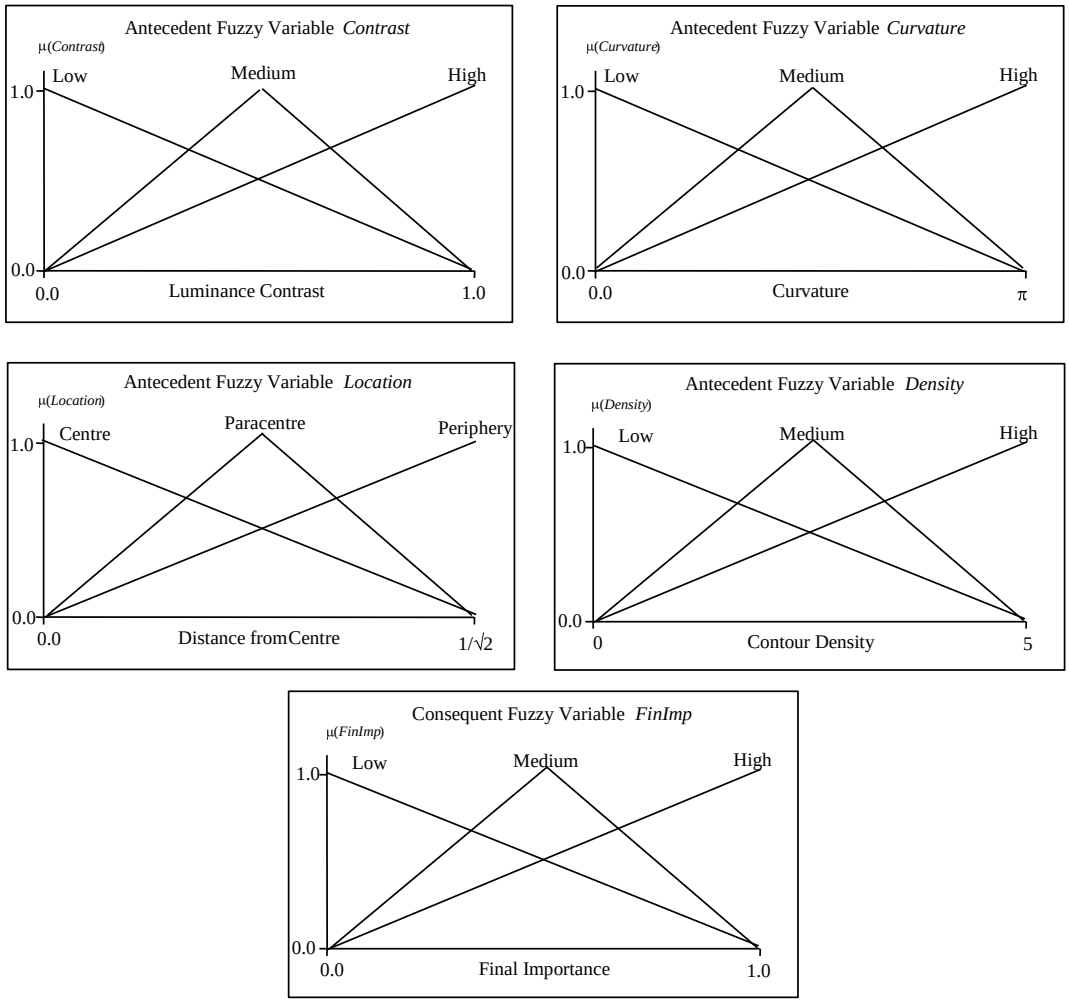


Figure 4.3 Illustration of the membership functions for the contour importance model, with four antecedent variables: Contrast, Curvature, Location, Density and the consequent variable FinImp.

Membership Function Development

As a part of the membership function design process, terms must be chosen to fuzzify the crisp truth of the variables being used. The contrast, density and curvature variables are fuzzified into three terms: Low, Medium and High. On the other hand, the location variable is fuzzified into the terms: Centre, Paracentre and Periphery, in order to characterise the spatial nature of the feature. Each of the variables is defined over a differing *universe of discourse* (function domain) according to the range of values available. The universes of discourse for each variable are:

- contrast–over the range of possible luminance values [0.0, 1.0];

- curvature–the tangent angle difference varies over [0.0, π] (to remove the sign from the calculations the value is derived from the absolute value of 2π - *curvature*);
- density–an appropriate ad hoc number of contrasts was chosen to be [0.0, 5.0], as having 5 crossings in the subdivision is a high density of contours within an 8×8 pixel subdivision;
- location–is the normalised window coordinate distance from the centre to the corner, for an assumed square image, ranging over [0.0, $1/\sqrt{2}$];
- finImp–is an arbitrarily defined discourse to indicate contour importance, ranging over [0.0, 1.0].

The membership functions defined on the above universes of discourse are illustrated in Figure 4.3. Triangle function shapes were chosen due to the lack of quantifiable data on contour importance. However, literature indicates that function shapes are not as influential as the actual implication process involved [7]. Therefore, the shape will still allow the module to ascertain the importance of subdivision contours. This leaves room for future work, to characterise further the relationships between aspects of this contour importance function.

Implication Methodology

The contour importance module uses a bounded sum implication methodology, incorporating weights that are applied to each rule–according to the feature it is processing [7]. Due to the bounded sum method of aggregation, the weights sum to 1.0. From experimentation with the test scenes, the following ad hoc weights in Table 4.1 work best:

Feature	Weight
Contrast	0.4
Density	0.3
Curvature	0.2
Location	0.1

Table 4.1 Table of weights for each of the contour model rules.

The following lists the high importance rule base:

IF *Contrast*	IS High	THEN *FinImp* IS High
IF *Curvature*	IS High	THEN *FinImp* IS High
IF *Density*	IS High	THEN *FinImp* IS High
IF *Loc*	IS Centre	THEN *FinImp* IS High

With medium and low rules being in a similar vein:

IF *Contrast*	IS Medium	THEN *FinImp* IS Medium
IF *Curvature*	IS Medium	THEN *FinImp* IS Medium
IF *Density*	IS Medium	THEN *FinImp* IS Medium
IF *Loc*	IS Paracentre	THEN *FinImp* IS Medium

IF *Contrast*	IS Low	THEN *FinImp* IS Low
IF *Curvature*	IS Low	THEN *FinImp* IS Low
IF *Density*	IS Low	THEN *FinImp* IS Low
IF *Loc*	IS Periphery	THEN *FinImp* IS Low

The *weighted fuzzy mean* defuzzification method is used to obtain a crisp importance value [86]. This method offers efficiency gains by making assumptions of symmetry with regards to the consequent function shapes. In the case of this system, the consequent shapes are triangular and symmetric. The shapes of the functions are then interpreted as a rectangle of height α (DOF value of antecedent) and width w (the power or width of the consequent function). The defuzzified value is then a ratio of the sum of the areas of the consequent functions, multiplied by the middle universe of discourse value a_i, and the sum of the total area of the consequent

functions. The formal equation to calculate the weighted fuzzy mean is shown below:

$$WFM(A) = \sum_{i=1}^{N_A} w_i \alpha_i a_i \Big/ \sum_{i=1}^{N_A} w_i \alpha_i \tag{4.3}$$

where:

WFM(A) is the defuzzified value for the fuzzy system;
N_A is the number of consequent membership functions;
w_i is the power (width) of the consequent function;
α_i is the activation value of the consequent function;
a_i is the numerical value of the consequent function (middle discourse value).

The weighted fuzzy mean defuzzification/implication methodology is one of the most efficient available, while still maintaining continuity of response. The method also provides a wide spread of values, in comparison to other area and maxima methods [7, 182]. The spread of values particularly suits the intended application in this project, due to the quantisation of the importance values into sampling rates for regions within the ray-traced image.

The final contour importance value is normalised to [0.0, 1.0] and stored in the *Contour Importance Map*–a subdivision map that identifies a spatial location with a contour importance value (refer to Section 5.3). The importance normalisation is performed using the following equation:

$$R_i = (R_i - MaxImp)/(MaxImp - MinImp) \tag{4.4}$$

where:

R_i is the importance value of region *i*;
MinImp is the minimum importance value of all the regions;
MaxImp is the maximum importance value of all the regions.

A fully worked example for the High consequent rules is now shown. The DCM from the progressively sampled image extracts the following values: *Contrast* 0.75, *Curvature* 0.3, *Density* 4 and *Location* 0.2. Each of these values is converted into a DOF value by the membership functions shown in Figure 4.4.

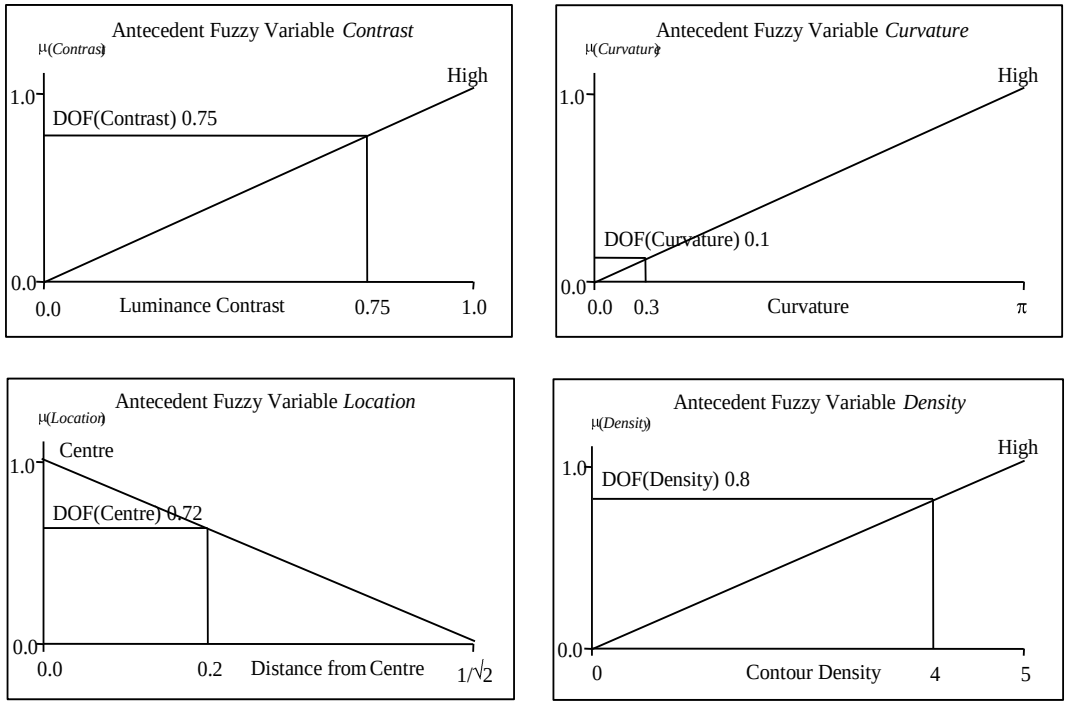


Figure 4.4 Illustration of DOF values drawn from the modified DCM algorithm for the High membership functions. Each value on the domain is convertedinto a DOF values for each of the High membership functions for each fuzzy variable.

The aggregation and defuzzification process uses the WFM technique. A multiple additive aggregation method is used, so the DOF values for contrast, curvature, location and density are added together and clipped to 1.0. The DOF value for the High consequent function is thus shown in Figure 4.5.

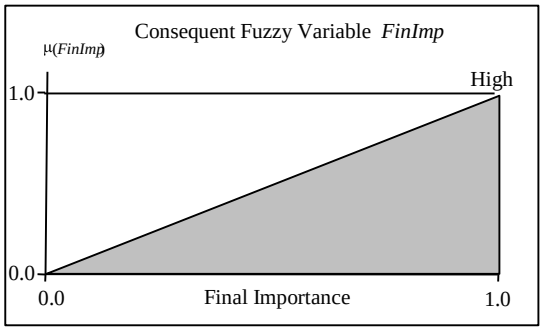


Figure 4.5 Illustration of the multiply additive DOF value for the High FinImp membership function example.

In Figure 4.6 the diagram shows the overlaid aggregated values for the Low, Medium and High membership functions and the final defuzzified value derived using the WFM technique. For the sake of brevity, only the High rule values have been calculated. The other consequent function values for Low and Medium have been assumed to be calculated previously using the same method as the High membership function.

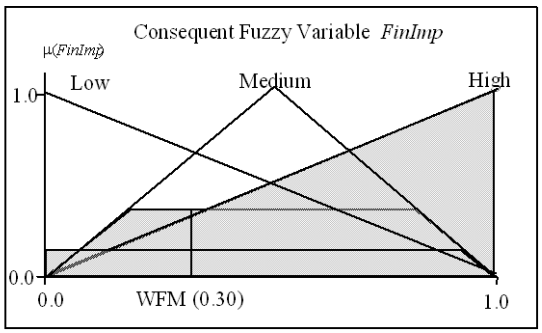


Figure 4.6 Illustration of the aggregated DOF values for the *FinImp* variable.

For this example, the three membership functions have fulfilment values of 0.1 (Low), 0.3 (Medium) and 1.0 (High). The calculation of the final defuzzified WFM value is carried out in the following fashion using Equation 4.2. Each membership function (Low, Medium and High) has its *width* multiplied by its *activation* value and the *middle discourse* value. These are then summed and divided by the sum of the widths multiplied by the activation values–translating for the examples to the following expression:

$$WFM(A) = \sum_{i=1}^{N_A} w_i \alpha_i a_i \Big/ \sum_{i=1}^{N_A} w_i \alpha_i$$

$= (0.1 \times 1.0 \times 0.5 + 0.3 \times 1.0 \times 0.5 + 1.0 \times 1.0 \times 0.5) / (0.1 \times 1.0 + 0.3 \times 1.0 + 1.0 \times 1.0)$

$= 0.30.$

The defuzzified value of 0.30 calculated above is thus entered as the importance value for the subdivision, and is used to control the progressive sampling of the scene within that subdivision. The above calculations are repeated for each subdivision in the contour importance map. An example contour importance map is illustrated in Figure 4.7 for an image of a head. Note that the contour map importance values are highest in subdivisions that contain a number of high contrasting, curved contours.

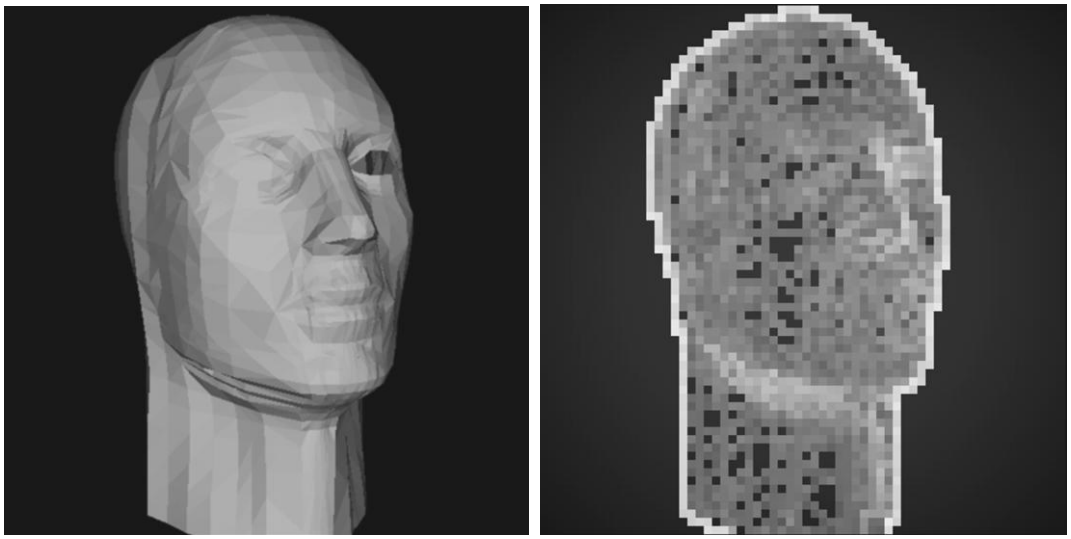


Figure 4.7 Illustration of the output of the normalised contour importance map for a head image, with the original image on the left and the generated contour importance map on the right.

The normalisation process facilitates the mapping of low and high importance values to the relevant variables in the rendering system. It ensures that the lowest value is mapped to 0.0 and the highest to 1.0, as the rendering system utilises a relative and not absolute importance value. That is, one subdivision is only important in comparison to other subdivisions in the image. The normalised contour importance value is used by the rendering system to progressively refine subdivisions in their visual importance order. Further details of this process are examined in Chapter 5.

4.2 REGION IMPORTANCE MODULE

From the reading performed in previous chapters, a number of basic rules about bottom-up forms of visual importance have been established. This has influenced the design of the visual importance model developed here. In particular, this has influenced the choice of visual features that are processed by the model, in order to obtain a visual importance value. A number of authors list some of the low-level image features so far discovered to influence the eye movements of a viewer:

- motion is considered the strongest attractor of attention [177];
- luminance contrast at the boundary of a region is a strong attractor of attention [144];
- hue contrast at the boundaries of different hued areas is one of the most obvious causes of pop-out and effortless texture segmentation [23, 39];
- contour concentration differences can lead to preattentive texture segmentation [80] and may attract the attention of the viewer [144];
- size differences [43];
- depth cues, from pictorial to stereo effects, are able to attract attention–especially foreground/background effects [177].

Some work has been carried out on the relationships of some of these features [124], but as yet, there is no mention of a model that completely describes these features, their ordering and their relative weights. However, some recent empirical work has begun to indicate general orderings and weights [128]. One aim of this project is to construct a more complete fuzzy preattentive feature relationship model by allowing for global effects and the use of contour concentrations to include texture factors. The ultimate intention is to use this model to devise appropriate techniques for efficient image synthesis.

The above features have been incorporated into a region-based fuzzy logic model of visual importance. The importance model fuzzifies the mean feature differences around segmented regions into three functions: Low, Medium and High.

One major improvement of this fuzzy logic importance model over others in the literature [35, 58] has been the introduction of membership functions that are adaptive in nature, in order to model the global effects mentioned later in this section. This has been implemented by passing the mean of the absolute value of background differences *m* to the system as a function shape parameter, so that the fuzzy threshold is dependent on the background feature variation in the image.

This concept can be illustrated by two extreme cases in the case of luminance differences. Scenes with low values of background activity within the image have the threshold as being the Just Noticeable Difference (JND) value, which is around one percent contrast for a grey level region on a constant level background [166]. The other extreme is a highly variant background, for example, a checkerboard. In this case, even the highest possible local contrast will not allow the region to become conspicuous, and so the mean difference value *m* forces the threshold to the far right of the domain.

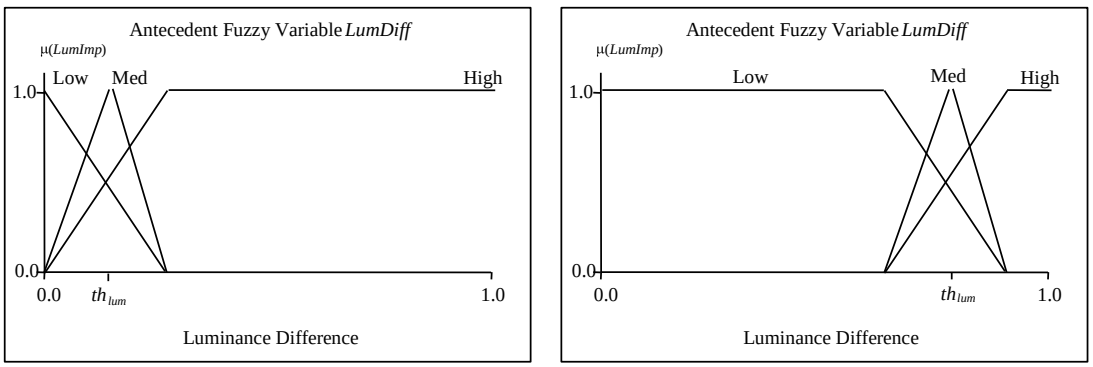


Figure 4.8 Example of the adaptive membership function shape approach. In the left diagram are the three membership functions centred around the Just Noticeable Difference (*JND*) threshold for luminance (around 1%), when the mean background differences are zero. On the right, the shapes are centred around the mean luminance differences (*m*), up to the extreme of 1.0. This moving threshold models the conspicuousness suppression caused by a highly variant background in the image, for example, a checkerboard.

This membership function models the sigmoidal effects analysed by Nothdurft with regards to global saliency effects. The results produced by Nothdurft show that the position on the feature domain varies according to the differences present in the background distractors. However, the basic nature of such conspicuousness functions does not change [122, 123]. In particular, the target will still seem prominent in a varying background, as long as there is a large enough local difference in visual features. This is consistent with the stimulus similarity model of Duncan and Humphrey [38], and other computational models [70, 128]. However, the newly presented model differs by considering pop-out to be sigmoidal in nature, as per Nothdurft, unlike other systems [128]. This is consistent with intuition, as the pop-out caused by a number of features, especially hue, are essentially threshold in nature, with a saturation point where the prominence reaches an upper limit. These results suggest that the mechanism which allows the pop-out to occur is designed to locate potential targets for further analysis, and not perform a measurement of the amount of feature difference between the objects in the scene. This is derived from the influence of top-down effects making an influence on the weightings of the features for search tasks. Therefore, the purpose of the membership functions presented here is to fuzzify the threshold where this pop-out occurs, to model the uncertainty of when a region is visually prominent or not.

Hence, the main thrust of this design is that once a region has popped-out, the addition of other feature differences will not make much difference to its attracting ability [124]. Simply put, the region once salient engages attention, but the fixation time depends upon top-down task oriented factors, which are beyond the scope of a bottom-up model as designed in this chapter. This approach also adds robustness to the system, due to the ability to sieve out the strong peaks in the image from the noisy background present in more natural scenes. This concurs with the evidence of experimenters finding viewers repeatedly regarding only a few regions within a natural image during unrestricted viewing conditions [21, 120, 151, 184].

Furthermore, anecdotal evidence is indicated from preliminary experiments performed for the development of this model indicating that the subjects had difficulty in assigning any sort of quantifiable value to the amount of pop-out

occurring in test stimuli [19]. As a result they were only able to easily assign a high, low or medium value to the level of conspicuousness. This in turn led to the development of the Low, Medium and High terms within the fuzzy functions, and the general sigmoidal shape.

Application area issues also arise for this model. Image synthesis applications will require that the images generated by the method do not contain artefacts that attract the attention of the viewer. It is unlikely that a bottom-up model of visual attention can ascertain the location of the viewers gaze in anything but a coarse manner. So it makes sense that a bottom-up attention system simply makes selections of regions that are likely to be regarded, without making subtle assessments of how long the person will regard the region. This is the methodology taken by Milanese, whereby the model uses a relaxation process to highlight the most salient objects and then thresholds the resultant saliency map to choose appropriate regions for object recognition purposes [107]. Similarly, in the image synthesis application area, the level of saliency can be divided into essentially regions that are regarded and regions that are ignored. The rendering resources can be concentrated on the former regions, without causing loss of quality in the latter. So the importance values calculated are heavily quantised due to the coarse nature of the model (refer to Chapter 5).

The fuzzy membership functions that model this threshold concept are shown in Figure 4.9. The threshold parameter t_{lum} is set according to the following formula:

$$t_{lum} = max(JND_{lum}, m) \tag{4.5}$$

where:

t_{lum} is the threshold to be fuzzified for the luminance membership function;
JND_{lum} is the just noticeable difference threshold for a luminance contrast to be detected;
m is the mean of luminance differences in the rest of the image.

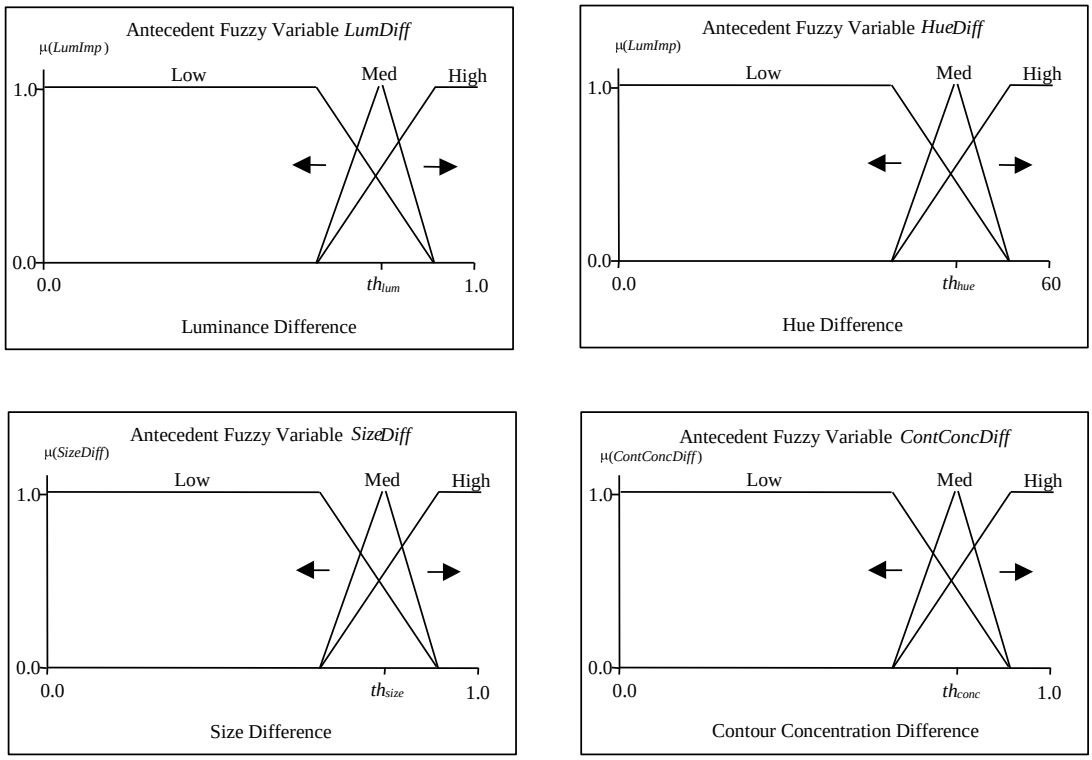


Figure 4.9 Illustration of the fuzzy, threshold-based membership functions for the features: luminance, hue, size and contour concentration.

A three-term membership function has been designed to best represent the perceptual phenomenon of visual conspicuousness. This is due to the phenomenon of visual pop-out being sigmoidal in nature [122], with a threshold and steep gradient. Piecewise linear trapezoidal membership function shapes have been used in this implementation for the sake of efficiency. Piece-wise linear approximations to the function shapes should not be a problem, as the relevant literature states that systems are more sensitive to the number of shape functions and their locations on the universe of discourse [7]. Other experiments previously performed indicate a small area of uncertainty around the threshold [19], so a third medium term has been added to the membership function.

These general threshold design principles have been implemented in several of the membership functions for features that rely upon differences to attract the attention of the viewer: luminance, hue, size, and contour concentration. Each of the relevant characteristics of these features is now considered in detail.

Hue difference is modelled on the hue angle difference of the region to its surrounds in HLS colour space. This is based on the assumption of the region being visually salient due to it being a different hue. That is, the region is salient when the hue angle difference indicates a different colour category, for example, red against green. Again, the principle is that once a region is significantly different then it will stand out, unless the differences surrounding it are greater. The universe of discourse for the function ranges over [0.0, 60.0]–60.0 degrees being the absolute limit of the hue difference needed to traverse from one hue category to another.

Size difference is modelled by the difference between the ratio of the image size taken by the region, and the average size differences in the surrounding image. Therefore, the difference has a universe of discourse ranging over [0.0, 1.0], allowing any image size to be processed. As it is inconclusive whether size exhibits the same pop-out effects as other features used in this model[7]. Yet, for two reasons the size feature will be treated in the same way as the other visual features in this visual attention model.

Firstly, it can be deduced that while an absolute model holds for a small number of objects in the scene, it does not capture the importance effects for relative differences that have been reported [177], and does not allow for a large number of objects in the scene. If there are a large number of objects in the scene of only small size then an absolute model cannot capture the effect of the size differences at this scale. An absolute model would consider the small size of the objects to remove any effect. However, even small objects, with a local spatial difference in size may stand out. Thus a relative size model, similar in manner to the other visual features, will capture these effects and the effects caused by few large objects in the scene.

Secondly, it is reasonable to keep the model consistent in its approach, for the sake of simplicity, and due to an expectation of the effects being similar for size when the appropriate research is carried out.

[7] Refer to Section 2.4.2 for a discussion of this issue.

Contour concentration differences are modelled using the information from the DCM that processes the progressive stage of the rendering (refer to Section 4.1). The region is analysed for the proportion of subdivisions that contain a contour. This proportion is a measure of the concentration of contours in the region, as a contribution to the saliency of the region. The universe of discourse ranges over [0.0, 1.0].

Other features that have been found attractive to viewer attention do not rely upon differences in features, namely; location and background/foreground differentiation. These features do not require a threshold parameter to allow for global effects, as they are inherently absolute in nature.

The location feature is modelled by taking the difference between the (x, y) position of the centroid of the segmented region and the centre of the image. This is performed in a similar manner to the previously described contour model (refer to Section 4.1).

Foreground/background segregation is an early and important feature of the human visual system [92], with experimentation showing its importance in deployment of visual attention [177]. Like other work [128, 189], this model uses the proportion of subdivisions in a region that contain half the border subdivisions. If this proportion is large, then the region will be considered to be a part of the background of the image, and so will be less attractive to the viewer. These values have again been fuzzified into three terms: Low, Medium and High.

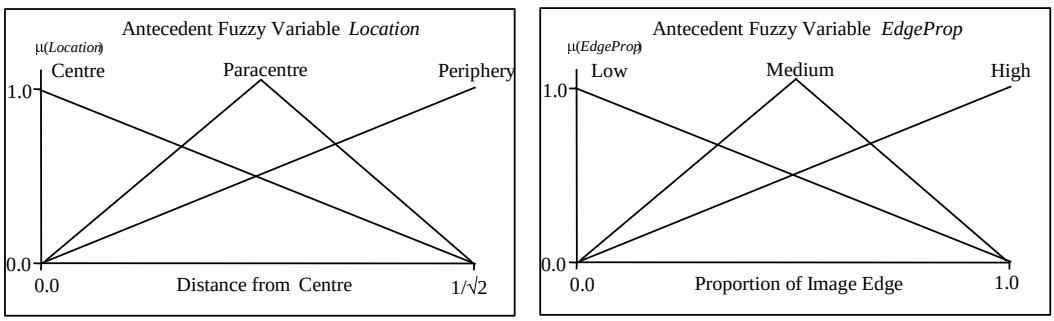


Figure 4.10 Diagram illustrating the non-threshold antecedent importance functions.

These feature-based fuzzy membership functions are then used in the implication process to ascertain the final visual importance of the segmented regions. The membership function for this final importance value FinImp is shown in Figure 4.11.

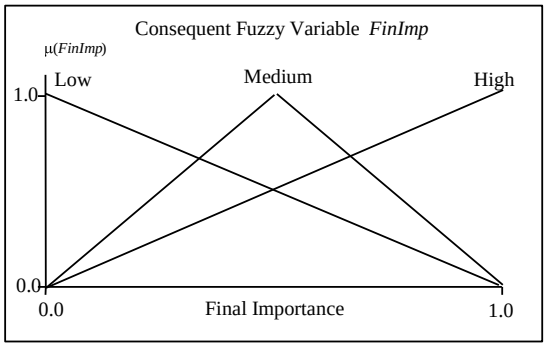


Figure 4.11 Diagram illustrating the consequent final importance function.

Region Module Implication Methodology

In the design of the fuzzy logic components of the system, consideration has been given to the three components of fuzzy deduction: *aggregation*, *defuzzification* and *implication*. The following sections describe the reasoning behind the choice of algorithms used to implement the fuzzy logic reasoning for the visual importance model.

It has been observed that the interaction of preattentive features is competitive in nature [70]. Furthermore, objects that are unique in appearance due to a single feature are more conspicuous than objects that are unique in appearance due to a conjunction of features [159, 177]. For example, a red circle among identical blue circles is more conspicuous than a red circle among red squares and blue circles. This component of feature interactions is handled by the use of the fuzzified thresholds in the membership functions mentioned earlier in this section. Regions that differ enough within a single dimension will be considered important, whereas the regions that differ due to combinations of features will not be considered salient in this model due to the lack of single feature dimension differences.

The overall interaction of features, however, is problematic. Evidence exists that the feature dimensions interact in an additive manner [124, 176], that is, the pop-out caused by differences in more than one feature dimension add together to cause a greater level of perceived pop-out. Successful models of the visual attention system have also been developed which are additive in nature [70, 127]. This additive nature is likely to be more complex than a simple addition, due to possible feature interactions [22, 23], differing feature weights [9, 128] and non-linearities [124]. The fuzzy implication method chosen is multiple additive, in order to model the overall additive nature of the combined importance of a region.

A bounded sum is used in the implementation, in a similar manner to the contour module described in Section 4.1. This form of aggregation of evidence simulates the additive nature of the importance of regions. Even though it is evident that a region stands out due to one feature dimension difference [155], the evidence above indicates that the importance of regions is enhance by multiple salient differences in feature dimensions. Therefore, the consequent membership functions have activation levels added together, instead of taking the minimum or maximum as is often the case [7, 182]. Multiple additive aggregation gives all feature dimensions a chance to contribute to the importance value of the region. This aggregation technique will still model a region becoming salient due to one feature dimension difference, as the other dimensions will contribute less due to the absence of feature differences.

The rules used are straightforward in nature. The process is best modelled using the general rule that if the local feature difference is high, and the local difference is above the global mean of feature differences, then the saliency of the object is high. From these concepts, the following list of rules constitute the fuzzy inference component of the region-based visual importance model:

IF *LumDiff*	IS High	THEN *FinImp* IS High
IF *HueDiff*	IS High	THEN *FinImp* IS High
IF *SizeDiff*	IS High	THEN *FinImp* IS High
IF *ContConc*	IS High	THEN *FinImp* IS High
IF *RegLoc*	IS Centre	THEN *FinImp* IS High

IF *EdgeProp*	IS Low	THEN *FinImp* IS High

The medium and low rules are similar to the High rules:

IF *LumDiff*	IS Medium	THEN *FinImp* IS Medium
IF *HueDiff*	IS Medium	THEN *FinImp* IS Medium
IF *SizeDiff*	IS Medium	THEN *FinImp* IS Medium
IF *ContConc*	IS Medium	THEN *FinImp* IS Medium
IF *RegLoc*	IS ParaCentre	THEN *FinImp* IS Medium
IF *EdgeProp*	IS Medium	THEN *FinImp* IS Medium

IF *LumDiff*	IS Low	THEN *FinImp* IS Low
IF *HueDiff*	IS Low	THEN *FinImp* IS Low
IF *SizeDiff*	IS Low	THEN *FinImp* IS Low
IF *ContConc*	IS Low	THEN *FinImp* IS Low
IF *RegLoc*	IS Periphery	THEN *FinImp* IS Low
IF *EdgeProp*	IS High	THEN *FinImp* IS Low

A multiply additive aggregation method is used [7], implemented as a bounded sum. This aggregation method enables the modelling of the importance as being a contribution of activation in a number of feature dimensions. Along with the multiple aggregation method, a set of weights has been implemented upon each of the rules. For now these weights are equal, except for the foreground/background feature, which has been made 2.5 times the others to enhance foreground/background differentiation.

The defuzzification and implication method is the weighted fuzzy mean, as used in the contour importance module (refer to Section 4.1).

The region importance, like the contour importance (refer to Section 4.1), is normalised to [0.0, 1.0] and stored in the *Region Importance Map*–a spatial map of regions within the image containing their associated visual importance (refer to Section 5.4). This is again due to the visual importance value being a relative

measure, with the lowest value being assumed to be 0.0 and the highest value being assumed to be 1.0 for pixel supersampling purposes. An example region importance map is shown in Figure 4.12.

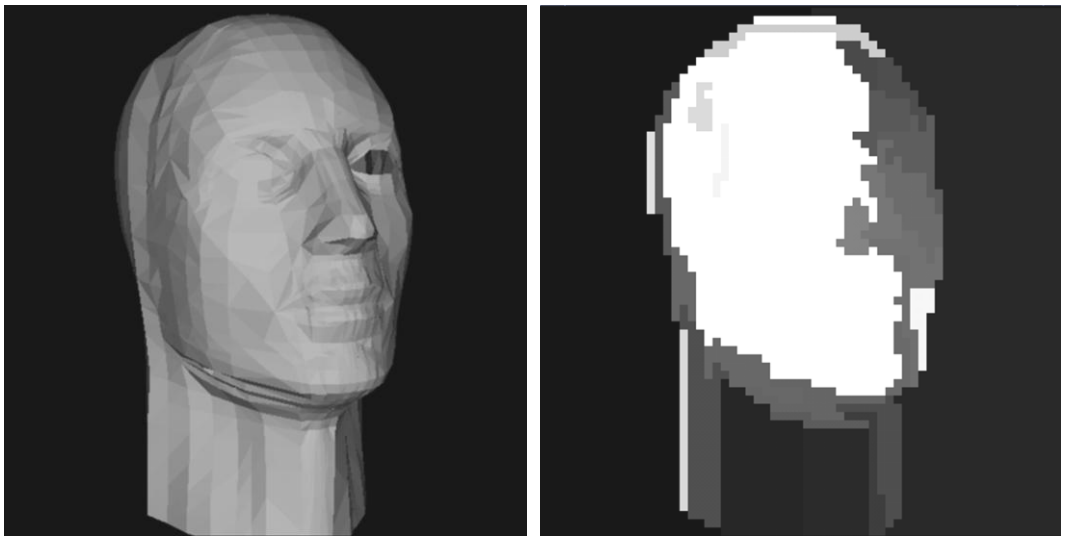


Figure 4.12 Illustration of normalised region importance map generated for the head image.

4.3 DISCUSSION

This chapter has described the development of two visual attention modules to be applied to the task of progressive and adaptive rendering; these being respectively the contour and region-based importance modules.

The contour fuzzy logic model highlights segmentation blocks that contain strong, highly curved contours and possible junctions. In Chapter 5 this model is applied to the area of progressive rendering, to order and accelerate those subdivisions that are considered important to perception of image quality. This seeks to improve the perceptual quality of the scene being rendered at an early stage of the progressive rendering process.

This chapter has also described a newly developed region-based fuzzy visual importance model. The model includes novel developments incorporating adaptive membership functions and contour concentration differences as an extra feature

dimension with which to evaluate visual importance. This model is also applied in Chapter 5 to supersampling algorithms used in ray tracing.

Chapter 5

Adaptive Image Synthesis Using a Visual Importance Model

Ray tracing, a commonly used hidden surface removal algorithm, uses complex lighting and surface shading techniques requiring large amounts of computing resources [46]. *Progressive ray-tracing* addresses the large time scales for image synthesis by presenting an initial low fidelity image to the viewer [100]. This low fidelity image is often stored as a regular subdivision of the image in a *quadtree* data structure [141]. The quadtree is then progressively refined until the final image is at full resolution. *Adaptive sampling* is a further modification to this method, where the sampling is concentrated around contours in a scene, continuing until a stop condition is met [129]. This stop condition may be from objective or perceptual measures of image refinement. Adaptive approaches reap significant savings in the number of samples made. However, while these methods are effective in producing scenes of high quality, they still expend effort on refining regions not important to the perception of the scene.

Both these approaches to ray tracing may benefit from the application of concepts drawn from psychological research into visual attention (refer to Chapter 2 and Chapter 3). The visual importance of regions in the scene can be used to control the progressive rendering approach, to reap further savings in sampling overhead. As a consequence, this chapter presents an approach incorporating region-based visual importance into progressive ray tracing techniques.

Efficient rendering is treated in this approach as a two-stage process. The first stage involves a new approach to progressive rendering techniques. Here, the issue is the choice of which spatial region in the scene to refine first, and by what magnitude. In order to facilitate this process, a contour importance model developed in Section 4.1 using fuzzy logic and results from psychological experimentation has been implemented. The main goal is to guide the progressive rendering algorithm to refine contours that are visually important to the viewer. This goal is achieved by

performing a regular quadtree subdivision of the scene, followed by application of the DCM to ascertain contour information within the subdivisions. These contours are evaluated for their saliency to the human visual system. The most important subdivisions are processed first and accelerated through the refinement stages, thus improving the visual quality of an early image. This technique is expected to be especially effective in the case of complex scene databases, where the cost of firing a single ray is large. In these circumstances the samples must be made in visually important regions, in order to efficiently present an effective early image.

The second stage involves a modified form of adaptive ray tracing, to render more efficiently the final high-quality scene. The region importance model developed in Section 3.2 analyses a coarse segmentation of the scene to calculate the visual importance of a region, and thus modulate the stop condition of the adaptive rendering. The main goal here is to sample heavily in the visually important regions of the scene. This differs from the edge problem discussed earlier, due to the contextual value placed on the regions. The edges are only processed locally, whereas global inter-region comparisons are performed in order to give an indication of contextual effects upon the visual importance of the region.

Two main rewards are reaped from this approach: an improvement in the visual quality of early progressive images, and savings in sampling rates for final high quality images, with minimal perceptual degradation.

As stated previously, progressive ray-tracing is the process of refining an image from a coarse representation until the final high fidelity image is produced. Progressive ray tracing techniques begin by coarsely sampling the scene to gain a first impression of the contents. This is usually represented as a segmentation of the scene, by either an adaptive quadtree representation [15, 100, 129], or a Delaunay triangulation [125, 130]. An adaptive regular subdivision was chosen due to a number of reasons. Firstly, it has proved to produce visually superior results, at low sampling rates [57]. Secondly, the quadtree gives a simple and efficient way of evaluating the importance of contours in the scene and the segmentation of scene regions, which is harder to

obtain from a Delaunay triangulation. Finally, it is the segmentation method used in the DCM, which has been utilised earlier in the approach.

These adaptive segmentations can be based upon object-space features in the scene geometry or by using image-space features from the early stages of the actual rendering. Using object-space features brings with it the advantages of pre-processing the scene geometry, in order to ascertain image-space features before they are rendered. However, this approach is prohibitive on two fronts. The object-space methods are inherently inaccurate when it comes to image-space features, due to the lack of information about their final appearance. Secondly, they are restricted in what image-space features they can analyse. Reflections, non-polygonal geometry and shadows are just some of the problems with this method.

Another approach is to use graphics hardware to pre-render the scene and obtain information from the image generated, which is then used to guide the adaptive process as an Oracle [130, 185]. Even though the strength of this method is the early approximate scene rendering, there are still problems with simplified lighting models. Moreover, there are limitations in scan-line algorithms with regards to the correct representation of reflections, shadows, displacement/bump maps and other advanced rendering techniques [112].

Scan-line methods render a polygon by directly scan converting polygons into pixels by a process called *rasterisation*, exploiting the *edge-coherence* of pixels along a line in image-space. Due to the rasterisation of polygons without reference to other polygons in the scene, the information available for the correct representation of reflections is not available. Therefore, a simple mapping of an image on an imaginary sphere surrounding the object being rendered performs the scan-line rendering of reflections [13]. This is efficient and appropriate for real-time systems, but does not create a true rendering of the reflections off objects within the scene, especially object to object reflections. Ray-tracing presents a simple solution by spawning secondary rays from intersection points on surfaces to render true

reflections. Similar problems occur within scan-line techniques used to represent shadows and bump maps[8].

These problems indicate that at present the technique of using image-space features generated by the ray-tracer itself is the best way to deal with progressive rendering, notwithstanding any advances in the area of hardware driven scan-line algorithms [67].

The approach taken here recursively subdivides the scene into 8×8 pixel elementary subdivisions. These subdivisions are then analysed for contours, and ordered according to their visual importance-based upon contrast levels, number of edges and curvature estimates. These subdivisions are then further sampled using the evaluated importance order. This should improve the perceptual quality of the image being presented, due to further sampling of the image in regions with a high concentration of high contrast, high curvature contours.

In addition to the progressive approach, a super-sampling technique has been implemented which uses a region-based visual attention model to regulate the stop condition of the image spatial subdivision, to prevent any further unnecessary sampling. Existing non-perceptual methods use statistical techniques to establish a measure of the homogeneity of a subdivision [57, 100, 110, 129]. While these methods are able to improve the efficiency of present ray-tracing algorithms by concentrating samples in regions of contrast, they do not account effectively for the visibility of the contrast value being used as a decision metric for further subdivision.

An improvement to this approach is the modification of the depth of the tree representing the refinement of the scene by allowing for the inferior colour sensitivity of the HVS [106]. The achromatic channel contains the highest levels of acuity within the HVS, followed by the red-green and blue yellow channels respectively. An opponent colour system is used within the ray-tracer to emulate the three HVS opponent colour channels. The spatial segmentation of the ray traced

[8] Bump mapping uses a projected image to modify the appearance of geometric surfaces within a scene (refer to Section 6.1)

image is modulated by the sensitivity of the three colour channels–that is, the achromatic channel had the highest level of subdivision, followed in order by the red-green and blue-yellow channels. This reaps modest efficiency savings, without perceptual loss of image acuity.

More sophisticated perceptual measures have been used to indicate to the renderer whether further subdivision and sampling of the scene is actually visible to the viewer [15, 111]. In this case, sophisticated models of early human vision, incorporating: opponent colour spaces, contrast sensitivity functions and *visual masking*[9] control the sampling rates of the ray tracers. Due to their sensitivity to only those spatial signals perceived by humans, savings in samples made by the respective ray-tracing systems reap further efficiency gains, in some cases of the order of $1/10^{th}$ of the original required without much perceptual loss of quality [14].

Many global illumination operators have been devised to deal with object-space measures of varying sophistication; however, only a few have dealt with image-space perceptual measures [47, 111, 134]. Among them, Yee [186], has implemented a multi -resolution visual attention model proposed by Koch and Ullman [83]. Yee has also incorporated the *Visible Difference Predictor* developed by Daly [30] into the approach.

The Visible Difference Predictor is a system developed to produce a probability map showing the visibility of differences between a degraded and original form of an image. The system passes an image through three main stages to achieve this aim: *amplitude nonlinearity*, *contrast sensitivity function* and *detection mechanisms*.

The amplitude nonlinearity stage models the nonlinear responses of the HVS to differing luminance levels. The contrast sensitivity function models the HVS sensitivity to spatial frequencies in an image. This stage accounts for optical effects from the eye, sampling effects induced by the cone photoreceptor and both passive

[9] Visual masking is the suppression of HVS signal detection by the superimposing of another similar signal.

and active neural connection effects. The final detection mechanisms stage contains the following subsections:

- spatial frequency hierarchy–models the frequency selectivity of the HVS (not its sensitivity) and creates a framework for the multiple detection mechanisms;
- masking functions–dealing with masking effects of superimposed spatial frequencies;
- psychometric function–which defines appropriate thresholds;
- probability summation–combines the responses of all the detectors into a unified perceptual response, that is, a visibility map.

The implementation of the VDP is used to control the ray-caching and sampling stages in a *Monte Carlo* ray tracing system [167]. These Monte Carlo integration algorithms are used to calculate the diffuse interreflectance (global illumination) contribution from other surfaces within the scene being rendered. Yee reports efficiency gains of 6 to 8 times the base-rate, again with minimally perceived error [185]. However, the approach requires a hardware-assisted, directly lit pre-rendering of the scene in order to act as an oracle to the global illumination process. This use of a hardware-accelerated prerendering has the already discussed weakness of lacking appropriate image-space information to make decisions on what will be visually important to the viewer.

Except for the multiresolution attention model used by Yee [185], the other adaptive methodologies do not incorporate any form of visual importance into their methods. The new approach presented here differs from the previous methodologies by using a region-based visual attention model, tightly integrated into the progressive rendering process, to facilitate efficient rendering by making further savings in sampling rates in areas not regarded by the viewer. Furthermore, this approach does not require hardware support to prerender the scene, but instead uses the progressive rendering performed by the ray-tracing system as input in its decision making processes.

5.1 A NEW IMAGE SYNTHESIS APPROACH BASED ON VISUAL ATTENTION

The newly developed fuzzy logic model of human visual attention has been integrated into a progressive and adaptive ray-tracer architecture. An overview of the architecture is shown in Figure 5.1.

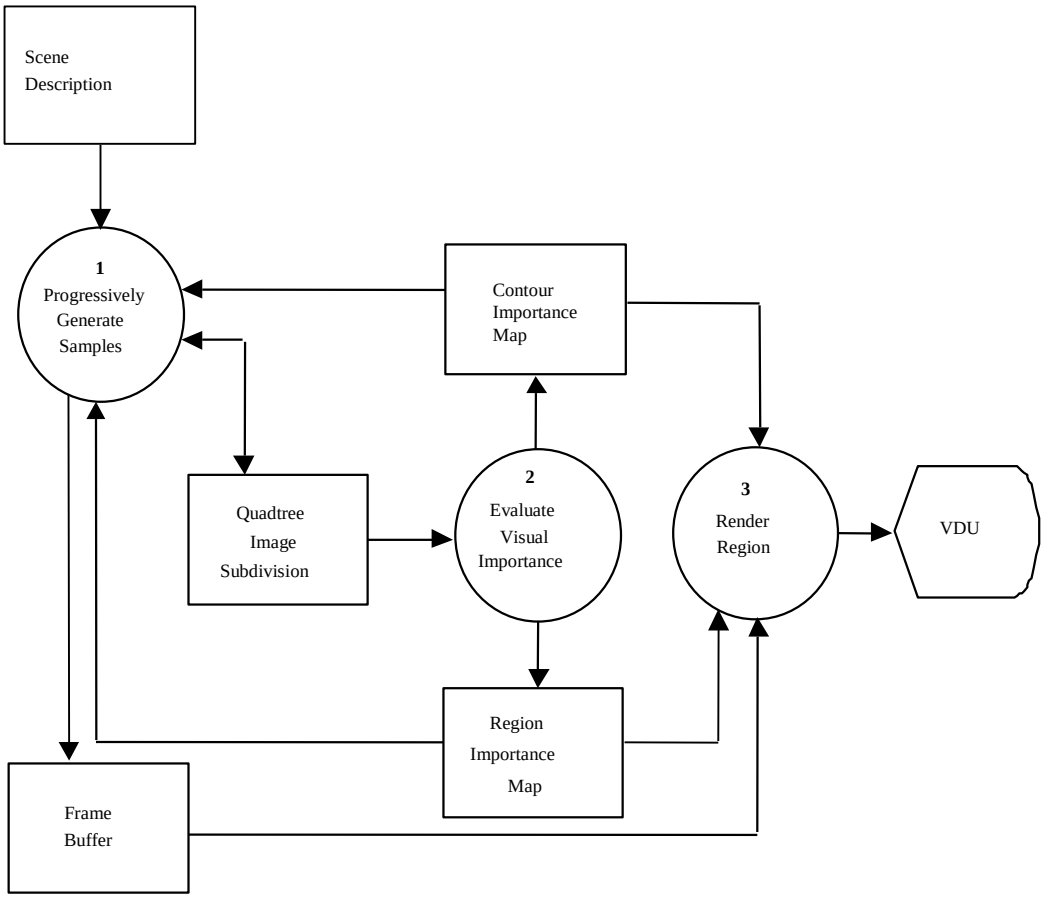


Figure 5.1 Overview flow diagram of the attention-based ray tracing system.

The implementation reads a Renderman® [162] scene description file, into a scene database. The *Visual Importance Model* evaluates both the contours in the image and the regions segmented from the image so far sampled, assigning them appropriate importance values. The *Contour Importance Map* is used to store contour importance information, whilst the *Region Importance Map* stores importance information for supersampling purposes. The *Progressive Sample Generator* uses information gained from the Contour Importance Map and the

Region Importance Map, to adaptively sample the image. A *Finite Element Renderer* can then produce an image, on demand, from the DCM and Frame Buffer [57].

Some elementary principles of ray tracing are now outlined in the next section, in order to form a theoretical basis for the development of adaptive sampling strategies[10].

5.2 RAY-TRACING PRINCIPLES

Ray tracing is a hidden surface removal algorithm that uses a vector-oriented model of the propagation of light within a scene. First, a centre of projection is defined from which rays are fired into the scene through pixels in the image plane. The ray may or may not strike an object in the scene. If an object is struck then an intersection point is calculated. A shading model is then applied to determine the colour of the object at the intersection. The shading model incorporates a model of the lights and the reflectance properties of the surface with which the ray has intersected. Calculations are performed to ascertain the colour of the light reflected from the surface into the synthetic camera view. This value is then placed in the frame buffer at the image-space pixel location passed through by the ray. All image-space pixels are treated in such a manner to create the final image (refer to Figure 5.2).

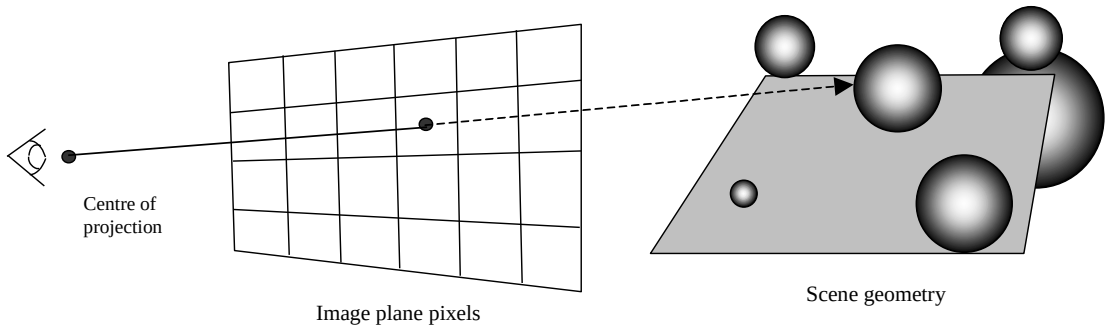
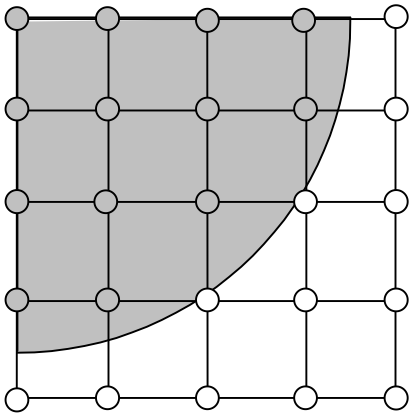


Figure 5.2 Diagram illustrating the ray being fired through the pixel into the scene geometry [46].

[10] Unless otherwise cited, the concepts are drawn from Foley and van Dam [46].

As with other sampling applications [175], there is the problem of dealing with aliasing effects caused by discretely sampling a continously defined function below the *Nyquist limit*[11]. This is handled by applying antialiasing techniques to cause more than one sample to be made at each pixel, otherwise known as *supersampling*. In supersampling, a similar process is carried out by rendering the image at the same resolution as required, and then subdividing the pixel and sampling it to smooth jagged edges [175]. The value of the pixel used in the final antialiased image is an average of the samples made within the area covered by the pixel in image-space–the image-space pixel being a discrete interval within the real valued dimensions of the scene. With regards to the latter technique, a number of methods have been developed to deal with supersampling, these being respectively: *regular*, *adaptive* and *stochastic*.

Regular supersampling involves the subdivision of the pixel into a regular grid, with sampling being carried out at each of the vertices, as per Figure 5.3. This reduces the aliasing of the image by performing more samples per pixel. For practical purposes, a limit of four samples in both the x and y dimensions is considered enough to reduce the aliasing in most edges within a typical computer generated scene, due to the sharp fall off in spatial frequencies [172]. However, as the sampling rate is constant across the pixel, there are a large number of unnecessary samples in quadrants with low to no contrast.



[11] The Nyquist limit dictates that a signal must be sampled at double its highest frequency component in order to replicate the original signal [175]. In the domain of image processing, this translates to specifying the lowest spatial sampling rate required to remove unsightly aliasing effects.

Figure 5.3 Regular sampling grid overlaid on a single pixel. Each of the circles represents a supersample of the pixel space.

Adaptive sampling improves regular sampling by making objective decisions about the variation of the pixel samples within a quadrant. This adaptive approach exploits the fact that dense sampling is only required around edges within a scene, due to the presence of higher spatial frequencies. Constant regions only require sparse samples to represent the low frequency signals present [175]. The quadrant is subdivided if the pixel samples it contains are above a contrast threshold, as per Figure 5.4. This process continues until the maximum limit of samples per pixel is reached. This limit may be defined by an arbitrary constant, an objective measure of contrast [110] or by a measure of perceptual contrast visibility [15].

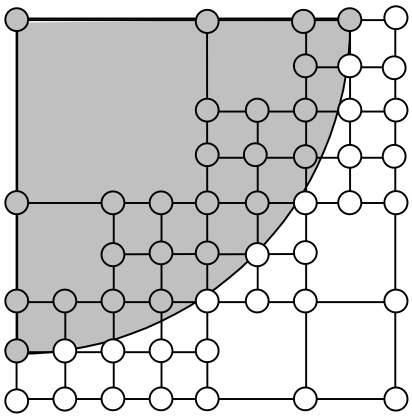


Figure 5.4 Adaptive sampling grid overlaid on a single pixel near the edge of geometry, illustrating the sensitivity of the method to contrast values.

Stochastic sampling jitters the regular grid in order to introduce noise into the sample regime (refer to Figure 5.5) [29]. The addition of the noise to the signal is more acceptable to the HVS than uniform sampling at the same rate, due to the breaking up of the regular aliasing frequency.

It should be noted that these supersampling techniques also work at levels above the size of a pixel. Progressive forms of image synthesis often use these same forms of sampling for a quadrant larger than a pixel, to guide the refinement of the image [57,

100, 129]. A modified form of adaptive sampling is used in Section 5.3 for the progressive sampling module.

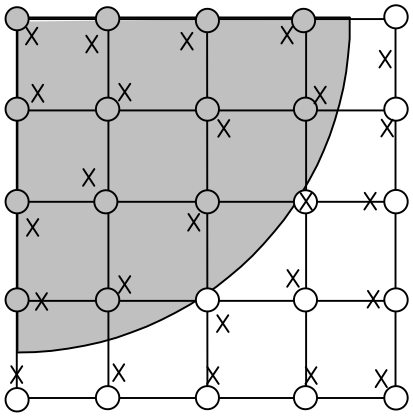


Figure 5.5 Illustration of a jittered regular sampling grid used in stochastic sampling strategies. Xs mark the jittered sampling locations.

Each of these antialiasing methods assumes that the viewer equally regards each region in an image. According to psychophysical research this is not the case. Human viewers regarding images tend to fixate on a limited number of regions that correlate with visual feature differences (refer to Chapter 2). The following sections detail the modulation of adaptive sampling approaches by the visual importance of the region being supersampled, in order to gain further efficiency improvements.

5.3 PROGRESSIVE SAMPLE GENERATION AND THE CONTOUR IMPORTANCE MAP

The image is regularly subdivided until a grid of 8×8 pixel subdivisions is generated. This grid forms the basis for the Contour Importance Map and the Region Importance Map. The subdivisions are analysed by an extended form of the DCM for contrast information, as a guide to further subdivision and sampling. This is performed by interpolating and thresholding the samples on the outline of the subdivision in order to obtain the points where contours cross (refer to Section 4.1). The process provides for each subdivision a measure of its contrast, number of contours and the contour curvature. Blocks are categorised as *smooth* (no contour present), *simple* (one relatively straight contour present) or *complex* (more than one contour, or one contour that is curved).

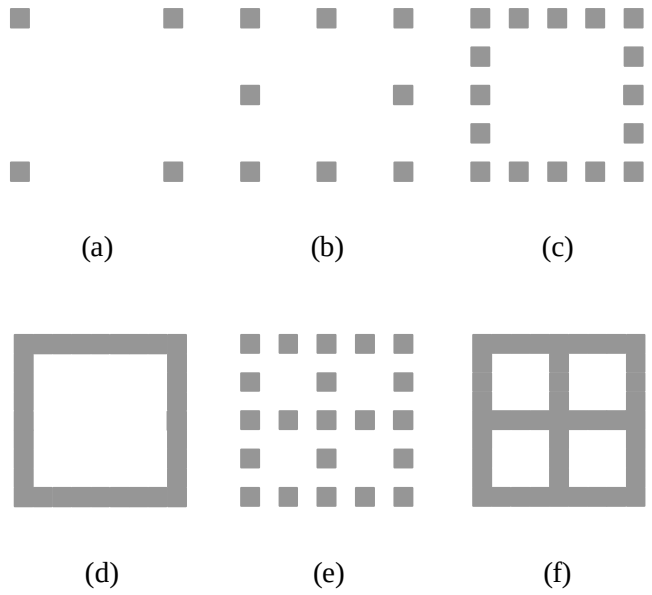


Figure 5.6 Diagram of subdivision sampling sequence for both simple and complex contour subdivisions. The grey squares indicate sampled pixels. Simple contour subdivisions follow the sequence (a)→(d), while complex contour subdivisions follow the sequence (a)→(c), (e)→(f). This is a modified form of the sequence used in the base DCM[57].

We extend the DCM algorithm by incorporating a fuzzy logic visual attention model that uses the contour information to calculate the visual importance of a subdivision. The contour categorising capabilities of the DCM have been extended by including measures of texture and bump map information–for more details on this extension refer to Chapter 6. The sample generator uses the subdivision contour information to guide further sampling. The general outline of the progressive sampling of the subdivisions is displayed in Figure 5.6, including a description of the differences in sampling steps for simple and complex nodes. The major difference being that complex subdivisions are further subdivided during the sampling pass, in order to better approximate interior details.

After each sampling run and importance evaluation, an array of pointers to the subdivisions is sorted according to the results from the visual model. This means that the subdivisions are then further processed in order of visual importance. After the first sampling run, a sort is only triggered if the average values within the elements have changed by a predetermined threshold.

In addition to the previous importance ordering, the approach has been modified so that subdivisions deemed more important are accelerated through the sampling hierarchy. For example, a complex subdivision with a high level of importance may jump from step *(a)* to step *(c)* (refer to Figure 5.6), thereby accelerating the refinement process, and so enhancing the quality of the image in visually important regions. However, this acceleration process must be applied in a judicious fashion, due to the inherent trade offs in the approach. Any acceleration of refinement in a region means that at any point other regions will be at a lesser state of refinement. Accelerate the important regions too much, and the rest of the regions are left in too low a state of refinement to improve the appearance of the image as a whole.

After appropriate experimentation with the images shown in this chapter, a maximum acceleration rate of two steps per pass has been settled upon as an appropriate magnitude. This acceleration rate is represented as an increment derived from the following formula:

$$Inc_{samp} = round(min(2.0, I_{seg} + 1.0)) \tag{5.1}$$

where:

Inc_{samp} is the sample increment value of 1 or 2;
I_{seg} is the importance of the subdivision being sampled, calculated by the visual importance model [0.0, 1.0].

After each sampling run each subdivision in the grid is further subdivided in the normal fashion of a quadtree, and the previous sampling regime is performed on the new subdivisions. Any contrast information gained from the DCM is passed to the Contour Importance Map at the 8×8 pixel level. The progressive refinement process is carried out until the subdivisions are 2×2 pixels in size. At this stage the sampling of the four corners of the subdivision is a sampling of each pixel in the subdivision, and so the supersampling stage begins.

5.4 REGION SEGMENTATION AND ADAPTIVE SAMPLING

The final stage in the progressive rendering of the scene is the supersampling of each pixel for antialiasing purposes [29]. The new approach uses a region-based visual attention model to ascertain the visual importance of regions in the image. The image is segmented at the 8×8 pixel subdivision level, into regions of similar hue and luminance values. This creates a list of regions able to be processed for importance information, as per other models [128, 189]. These importance values are stored in the Region Importance Map (refer to Figure 5.7).

A number of issues arise with the need to segment the scene into importance regions. Computational efficiency needs to be weighed against the need for discernment accuracy. The region segmentation is performed at the 8×8 pixel level, and not at the single pixel level, due to the computational savings from segmenting the scene at a relatively coarse resolution. The region segmentation is performed progressively during the refinement of the image from the elemental subdivision stage, as detailed in Algorithm 5.1. As stated before, the image is refined up until the size of a pixel, where the supersampling stage begins. Supersampling is therefore performed as a later pass, incorporating the first samples taken in the pixel to maximise efficiency.

Viewing factors also influence the decision to segment the scene at an 8x8 pixel level. Some methods use multiresolution methods to assign an importance on a pixel by pixel basis [71]. However, evidence suggests that viewers do not fixate on single pixels, but instead fixate on regions in the image being viewed–derived from research showing collections of preattentive features being perceived in a region-based manner [178] (refer to Section 3.2). Further support is drawn from previous region-based visual attention models that have used 16×16 pixel blocks [127] and 8×8 blocks [97] as bases for region segmentation. The approach developed here uses an 8×8 pixel block for the importance maps, due in part to this sized subdivision being a basis for the DCM, and the efficiency issues previously mentioned.

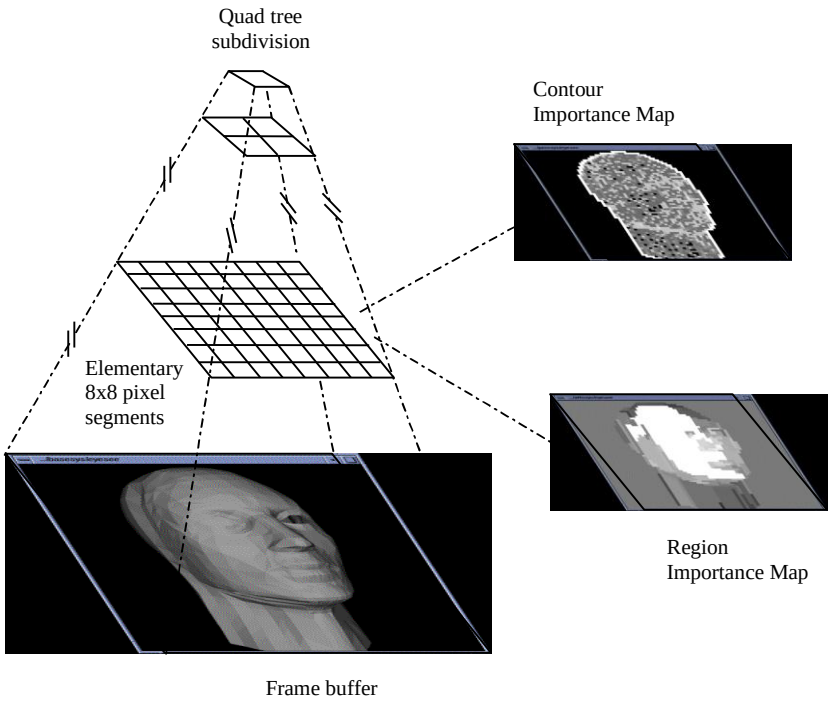


Figure 5.7 Relationship between importance map data structures in adaptive rendering approach.

The regions are assembled from the elemental subdivisions using a merge algorithm [140], with hue and luminance values for a subdivision compared by a fuzzy logic system. Luminance and hue thresholds are used, both representing the proportional similarity required between subdivisions before they are merged–that is 0.0 means totally dissimilar and 1.0 being identical. For this implementation the ad hoc value of 0.98 for both hue and luminance thresholds has worked well. An example segmentation of a head scene is shown below in Figure 5.8.

An issue for this coarse segmentation is the effect of aliasing introduced by the change of sampling rates across the boundaries of regions with differing importance. It is possible that the importance values could be linearly interpolated across the region boundaries in order to prevent sudden changes in image quality introducing *blockiness* into the image being produced. However, any image quality discontinuities present would be induced by the variation in sampling rate for the region, and not by the importance change gradient. In the implementation presented in this thesis, the subdivision rate for each pixel ranges from 1 to 4. No matter what the spatial change in region importance, the subdivision rate is quantised to four

values, which may bring about sudden changes in image quality. Therefore, the aliasing introduced is inherent to the quantisation of subdivision values, so the region importance values are not interpolated across the region boundaries.

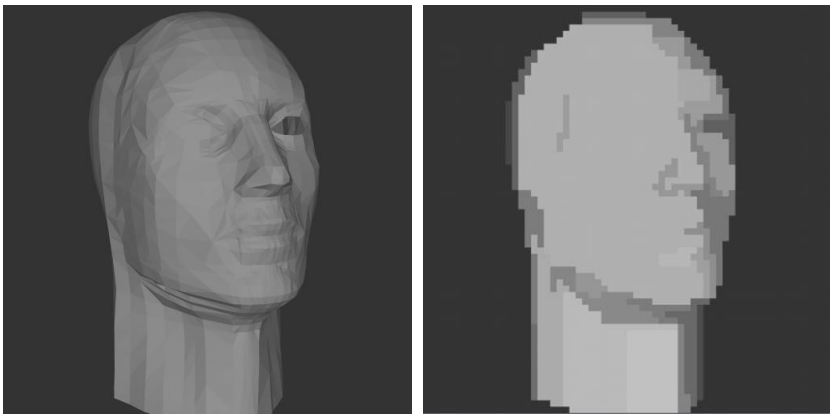


Figure 5.8 Illustration of the output of the segmentation algorithm (right) from an example head image (left).

The single samples taken previously during the progressive rendering process are incorporated into the supersampling by assuming they have been made at the bottom-left corner of the pixel. This differs from the standard practice of using the centre of the pixel, however, for this implementation the use of the bottom-left hand corner facilitated easier adaptive rendering strategies. This was due to the addition of further samples into the pixel at increments away from the bottom-left hand corner. In general, the supersampling function may be represented as the equation:

$$S_{seg} = I_{seg} \times SampFunc(p) \tag{5.2}$$

where:

S_{seg} is the super-sampling rate for the region;

I_{seg} is the visual importance value for the region being processed [0, 1];

SampFunc is a function determining the supersampling rate for the pixel *p* within the region by means other than visual importance, such as mentioned in Section 5.2

Regular super-sampling has been used in this implementation. It is expected that the approach should easily accommodate other methods such as *adaptive* [129] and *distributed* [29] sampling. Adaptive and stochastic algorithms would be in the general formula above, with the *SampFunc* being the algorithm used in the adaptive or distributed form of super-sampling. For the purposes of this project, as a proof of concept, the SampFunc has been implemented in two ways: *flat-rate* and *perceptually-based* supersampling. This will enable efficiency comparisons between two broad supersampling methodology areas, and should give a clearer picture of the performance of the new importance-based rendering approach.

5.4.1 Flat-rate Supersampling

The flat-rate supersampling method is the simplest to implement, as it consists of a constant magnitude of pixel subdivision across the scene–that is, the *SampFunc* is a constant. This can be implemented by multiplying the supersampling rate by the value from the importance maps.

In this new implementation the importance value calculated from the Region Importance Map I_{reg} has been used:

$$I_{seg} = I_{reg} \tag{5.3}$$

where:

I_{reg} is the importance of the segmented region;
I_{seg} is the final importance of the region to be refined.

The flat-rate supersampling method has a maximum subdivision rate applied to the pixels called *MaxSamples*, for this implementation varying over [1, 4]. Thus, a final high quality image has a maximum samples per pixel ranging over [4, 16]. Therefore, any pixel within an elementary subdivision has a subdivision rate S_{seg} of:

$$S_{seg} = I_{seg} \times MaxSamples \tag{5.4}$$

where:

S_{seg} is the number of samples for the pixel;

I_{seg} is the importance of the region;

MaxSamples is an implementation dependent maximum number of subdivisions for the pixel (in this implementation set to 4).

5.4.2 Perceptual Supersampling

In this second method, a perceptual image quality metric is used to control the stop condition on the refinement of the image. Myszkowski has explored this perceptual approach by experimenting with applications of the Daly Visual Difference Predictor [30] to global illumination problems [111]. It was found to be an effective difference metric with regards to global illumination issues, such as: progressive indirect lighting solutions, lighting stopping conditions and mesh discontinuity processing. Results indicated that it facilitated efficiency gains in the rendering of the scene. Bolin and Meyer [14, 15, 105] have taken a different approach, by using a modified Sarnoff *Visible Difference Metric* (VDM) [95], and applying it to Monte Carlo ray tracing techniques.

The Sarnoff VDM works in a similar manner to the Visible Difference Predictor by Daly [30]. However, the Sarnoff VDM seeks to model more closely the physiological nature of the HVS, whereas the Daly VDP is more psychophysically based. The major stages in the system are:

- *cone fundamentals*–splits the signals into small, medium and large frequency cone responses;
- cortex filtering–sets up a multiresolution model of the orientation filters contained within the visual cortex;
- *local contrast*–models the non-linear response to light;
- *chromatic aberration*–models the effects induces by the optics of the human eye;
- *opponents contrast space*–models the opponent colour space of one achromatic and two colour channels contained in the HVS;

- *csf filtering*–models the chromatic and achromatic components of the contrast sensitivity functions;
- *masking transducer*–mimics the masking effects of similar superimposed spatial frequencies;
- *spatial pooling*–implemented as a filter to mimic the peak sensitivity of the HVS to five cycle sinusoidal signals.

The two images to be compared are processed by these stages and then passed through a final distance summation stage to create a visual difference map–showing the difference in image quality in all spatial locations.

Again, the system was able to detect masking and contrast effects, and delivered efficiency increases for the rendering algorithm. Yee has also implemented a variant of the Daly Visual Difference Predictor [186], and has used it to improve the efficiency of Monte Carlo lighting solutions, in particular, the ray caching and sampling rates. The approach was able to modulate the number of samples made for the global lighting integral by the visibility of the refinement made to the pixel, and its visual importance within the image. This perceptual visibility component was incorporated into a modified version of a multiresolution visual attention system, as described in Section 3.1.

Along with other approaches [134], these methods of image comparison are limited in their application, due to the overhead of transforming the image into frequency or wavelet spaces.

Neumann et al. have developed an image-space metric for image quality comparisons, incorporating the analysis of random rectangles projected over the two images to be compared [114]. The use of the contrast sensitivity function allows the method to be sensitive to the visibility of certain spatial frequencies. The metric is simple in nature and does not account for effects such as chromatic aberration and masking effects of spatial frequencies. However, it is highly efficient, as it does not require a frequency space transformation of the images being compared. For now, it

will suffice as a proof of concept with regards to incorporating perceptual image synthesis into the region-based visual attention model.

The perceptual image difference method uses efficient techniques for calculating the difference between the average hue of the rectangles in L*u*v* colour space, and is weighted by the size of the rectangle and human visual contrast sensitivity. A threshold is applied to each rectangle error in turn, such that if the difference between two rectangles is above the threshold, then the error is added to the combined error for the entire image.

The stop condition is used in this approach in the following manner:

1. A scene is generated at an arbitrary level of resolution in the progressive renderer.
2. The values for the rectangles are calculated and stored.
3. The next iteration of the renderer generates an image at a higher super-sampling rate.
4. The values for the rectangles in the new image are calculated and stored.
5. The differences between the rectangles in each image are thresholded.
6. The image synthesis process continues until the cumulative error for all the rectangles between two refinement levels is zero.

To incorporate this metric into the adaptive rendering algorithm, the rendered image is convolved with the importance values derived from the region importance model. While calculating the rectangle error, each value from the image $P_{x,y}$ is multiplied by the region importance value I_{reg}, in the following manner:

$$R_{x,y} = I_{reg} \times P_{x,y} \tag{5.5}$$

where:

$R_{x,y}$ the pixel value contributing to the rectangle in the perceptual difference metric;

I_{reg} is the region importance value of the pixel [0, 1];

$P_{x,y}$ is the pixel value in the image at position (x, y).

As a consequence, the importance of each pixel weights the calculation of the difference between the rectangles. An intended corollary of this process is the modulation of the sampling performed within each pixel of the image by the visual importance of the segmented region. The next section will detail the incorporation of these methodologies into an integrated visual importance-biased rendering algorithm.

5.5 ALGORITHM DESCRIPTION

The progressive rendering algorithm detailed in this section refines the scene to the level of a pixel, for the final supersampling phase to take place. The pseudocode is listed in Algorithm 5.1. The algorithm has three main stages for each step in the refinement process. The first is the decision of how many steps to refine the subdivision–based upon previous contour importance calculations. The second is the generation of a list of subdivisions from the subdivision being processed. These subdivisions, if any, are then sampled, subdivided and evaluated for contour information. Finally, if the average statistics have changed for the subdivision, then the subdivisions are remerged into regions and importance values calculated for both the contours and regions.

The following pseudocode in Algorithm 5.2 details the *EvalRegImp* function called in Algorithm 5.1. Its general structure contains two major sections to process the segmented regions. The *EvalRegGlobalDiff* function is called to calculate the global average of feature differences within the whole image (refer to Algorithm 5.3). The loop then processes each region again, to ascertain the local differences in features surrounding the region being processed. These feature difference values are then passed to the fuzzy logic module to assign a visual importance value to the region. The visual importance values are then normalized to [0, 1].

```
Inputs:     Nil
Outputs:    Nil

Set stepMax ← 6                                  { Set Maximum number of steps. }
Set impSeg ← {e1, e2, … , en}                    { Subdivide image to elementary 8×8 pixel
subdivisions. }
Set impSeg[seg].step ← 0 ∀seg                    { Reset refinement step for each subdivision. }

for subLevel = 1 to 3 do                         { For every subdivision step to pixel size. }

    if subLevel ≥ 2 then                         { Adjust maximum number of sampling steps }
        Set stepMax ← 1                          { to refinement level. }
    else
        Decrement stepMax by 1
    end if

    for step = 0 to stepMax do                   { For every subdivision sampling step. }
        for seg = 0 to numElem do                { For every 8 × 8 subdivision }

            if impSeg[seg].step < stepMax - 1    { If subdivision already fully sampled }
              or impSeg[seg].step = 0 then
                if step > 0 then                 { Pass, on first time through. }
                    if impSeg[seg].imp < 0.5 then { Set step according to subdivision importance }
                        Increment impSeg[seg].step by 1
                    else
                        Increment impSeg[seg].step by 2
                    end if
                end if
            end if

            { Subdivide subdivision (if needed) and sample }

            Set numSubSeg ← 0
            GenSegList(impSeg[seg], ElemTreeDepth + subLevel)

            for j = 0 to numSubSeg -1 do
                if subSegList[j].edgeType ≠ SMOOTH then
                    SampBound(segList[j], impSeg[seg].step)
                end if
            end for

            if segChanged = True then            { Sort and subdivide if changes in features }
                Normalise(impSeg)
                Sort impSeg by impSeg[seg].imp ∀seg
                RegionSeg(impSeg, regions)
                EvalRegImp(impSeg, regions)
            end if

        end for
    end for
end for
```

Algorithm 5.1 Progressive rendering algorithm pseudocode.

Procedure: EvalRegImp

Inputs: List of *regions* containing segmented region information for image.
List of importance subdivisions *elemSeg* containing 8 × 8 pixel level information

Outputs: Nil

EvalRegGlobalDiff(*regions, numReg, lumDiffMean, hueDiffMean, sizeDiffMean, contDiffMean*)

for *regNum* = 0 to *numReg* **do** { For each region do }
    Set surrLum, surrHue, surrContDens, surrSize ← 0;

    for *surr* = 0 to *regions*[*regNum*].*bordCount* **do** { For each surrounding region do }
        Set surrReg ← elemSeg[regions[regNum].border[surr]].regNum
        Add the surrounding local feature values of region *surrReg* to *surrLum, surrContDens, surrSize, surrHue*
    end for

    { Obtain local feature difference values, with respect to surrounding average feature values }

    Set lumDiff ← |regions[regNum].lumAvg - surrLum / regions[regNum].bordCount|
    Set hueDiff ← |regions[regNum].hueAvg - surrHue / regions[regNum].bordCount|
    Set contDensDiff ← |(regions[regNum].contCount / regions[regNum].segCount) - (surrContDens / regions[regNum].bordCount)|
    Set sizeDiff ← |regions[regNum].segCount / numElem) - surrSize / regions[regNum].bordCount|

    { Obtain absolute feature values }

    Set loc ← ((regions[regNum].centreY - dimElem / 2.0)^2 + (regions[regNum].centreX - dimElem / 2.0)^2)$^{-2}$ /
        ((*dimElem* / 2.0)^2 + (*dimElem* / 2.0)^2)$^{-2}$
    Set edgeProp ← regions[regNum].imEdgeCount / (dimElem × 2.0 - 2.0)

    { Obtain visual importance of region using fuzzy logic system designed in Section 4.2 }

    RegImp(imp, lumDiff, lumDiffMean, hueDiff, hueDiffMean, sizeDiff, sizeDiffMean, contDensDiff, contDiffMean, loc, edgeProp)
    Set regions[regNum].imp ← imp

    if regImpMax < imp **then**
        Set regImpMax ← imp
    end if

    if qTree->regImpMin > imp **then**
        Set regImpMin ← imp
    end if
end for

NormRegImp(*regions, regImpMax, regImpMin*) { Normalise values to [0, 1] }

Algorithm 5.2 Algorithm listing for EvalRegImp procedure.

```
Procedure: EvalRegGlobalDiff

Input:    List of segmented regions containing visual feature information
          Number of regions numReg
Output:   A number of global feature activation variables: lumDiffMean, hueDiffMean, sizeDiffMean, contDiffMean

Set diffCount ← 0

for regNum = 0 to numReg do                                   { For each region do }
    for surr = 0 to regions[regNum].bordCount do              { For each surrounding region do }

        Set surrReg ← elemSeg[regions[regNum].border[surr]].regnum   { Obtain surrounding region number
                                                                       from index }

        Set revSurrReg ← 0                                    { Find the link to the present region
                                                                regNum from the surrounding region
                                                                surrReg }}
        while regNum <> elemSeg[regions[surrReg].border[revSurrReg]].regNum do
            Increment revSurrReg by 1
        end while

        { If the present region has not been compared to the surrounding region then add feature values to global feature }
        { difference variables }

        if not regions[surrReg].checked[revSurrReg] then
            Add difference between surrReg and regNum regions to total variables: lumDiffMean, contDiffMean,
                sizeDiffMean, hueDiffMean.
            Set regions[regNum].checked[surr] ← True;         { These two regions have been
                                                                compared }
            Increment diffCount by 1
        end if
    end for
end for

if diffCount <> 0 then
    Divide lumDiffMean, contDiffMean, sizeDiffMean, hueDiffMean by diffCount   { Obtain averages for global feature
                                                                                 differences }
end if
```

Algorithm 5.3 Algorithm listing of EvalRegGlobalDiff procedure, which calculates global feature difference values.

The modified DCM method of progressive image refinement shown in Algorithm 5.1 remains O(n) with regards to the number of 8×8 pixel subdivisions in the image. The space requirements are unchanged from the original algorithm as well.

The contour importance approach requires only one multiply for each subdivision to set the acceleration of the refinement. The method uses the DCM information already available, therefore no overhead is incurred gaining the contour information from the subdivisions as they are being refined. The subdivisions are only processed for contours locally; no comparison is made with other subdivisions within the image. The fuzzy subsystem has the same computational overhead no matter what the contour information within the subdivision. Therefore, the time complexity of this method is O(n) with regards to the number of subdivisions. In practice, the

importance evaluation at the end of each refinement stage takes less than one second on an R5000 Silicon Graphics O2, for a 513×513 pixel image, including the sorting of the importance map. Memory requirements are a contour importance map, aggregated from the subdivisions at the 8×8 pixel level of the quadtree subdivision, and so are also O(n) with regards to the size of the image. Two more words are required for each subdivision, in addition to the information stored for the DCM algorithm. One word contains the subdivision importance value, the other is a pointer to the appropriate quadtree subdivision for the importance map.

The region-based importance method has larger space requirements due to the region segmentation component of the algorithm. The worst case is that every elemental subdivision in the importance map is segmented as a region by the merge algorithm. For this scenario the expression for the region importance calculations will be performed on four times the number of subdivisions within the image, as the segmentation uses a four-connected merge algorithm [140]. In addition, a second loop uses global values calculated to bias the final importance value for the regions (refer to Section 4.2). Therefore, the region-based importance calculations are O(n), with regards to the number of elemental subdivisions in the scene. Space requirements are the storing of 24 words of information (average hue, average luminance, number of contours, number of subdivisions, number of edge subdivisions and region location) within the region to efficiently calculate importance values–leaving the algorithm with an O(n) space requirement in machine words, with respect to the number of regions in the image.

Once this progressive algorithm has completed, the supersampling process can then subdivide each pixel, modulated by the visual importance of each pixel region.

5.6 OBJECTIVE IMPLEMENTATION EVALUATION

In order to evaluate the new progressive approach a number of scenes have been assembled as a representation of the broad categories of images used in rendering systems. These categories are: a single object within the centre of the screen, a scene

taken inside a building and an outdoor scene with a horizon. Examples of these scenes are shown below in Figure 5.9.

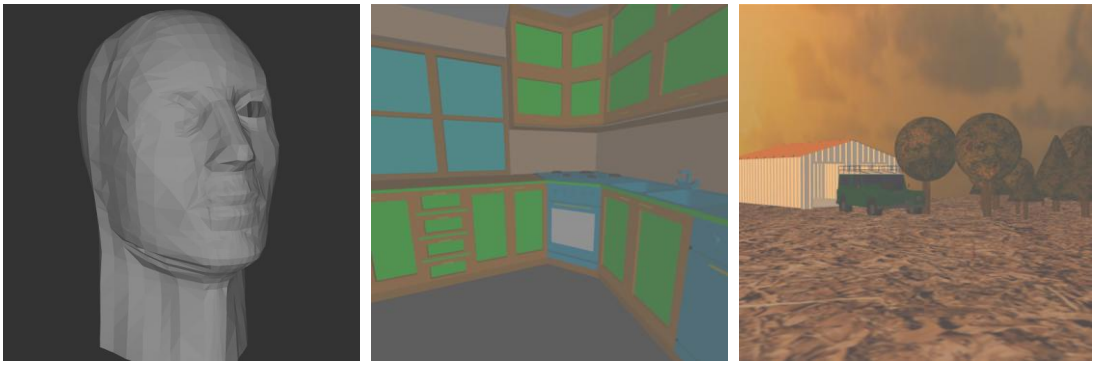


Figure 5.9 Example scenes used in the evaluation process. From left to right they are a single object, an indoor scene and an outdoor scene.

Each represents a general category of scenes, which can be developed for image synthesis. The first is a single object within the centre of the scene. This is often the way CAD images are constructed for viewing prototype renderings. The second scene is of a typical indoor image with floor and roof and multiple objects. The final and most complex scene involves a strong perspective and a horizon, with textured objects, making the scene noisy in nature. Each scene was rendered at a resolution of 513 × 513 pixels. More details about each scene are included in Table 5.1.

Scene	Number of Polygons	Reason for Inclusion
Head	~3,000	Single object in centre of image.
Kitchen	~2,000	Indoor scene with more complexity and colour.
Farm	~12,000	Outdoor scene with colour, texture and high levels of complexity.

Table 5.1 Details of each scene used in the evaluation of the rendering approach, both progressive and supersampling.

As progressive rendering is a temporal improvement in the quality of the image, the comparison of image quality will be performed over a series of images. An objective evaluation metric was applied to the progressive images generated by the system over the first 10% of the samples, to gain an indication of the improvement of the

images. Each image was rendered at 1% sampling intervals (1%, 2%, 3%...10%), giving 10 images in all.

The visual importance-based supersampling methodology is a form of degradation, which is more acceptable to the viewer due to the degradation being moved to areas that are less noticeable. To do this the evaluation involved the comparison of a normal approach using a level of supersampling, and the region-biased method using the same maximum super-sampling rate. As there were expectations of varying performance from both forms of supersampling at differing supersampling rates, the upper bound sampling rates are varied for both methods.

For the flat-rate method, the super-sampling rate maximum was varied from 4 to 1 subdivisions per pixel. For the perceptual method, the threshold used for comparisons between the two stages of refinement was varied from 10 to 50. This latter threshold controls the magnitude of error tolerated between the images. Previous experimentation had shown that the error thresholds below 10 become prohibitive due to non-termination of the algorithm within a reasonable amount of time. While error thresholds above 50 became superfluous due to the error value always being less than the threshold. Furthermore, the variation of error thresholds gives an indication of the optimum sampling rates that enhance the image quality of degraded images.

The objective methodology used to evaluate the approach is detailed in the next section. Subjective image quality comparisons have also been performed, and are reported in Chapter 8.

5.6.1 Objective Evaluation Metric

The objective measure used is the L1 / L2 norm error ratio method used in similar work in the progressive rendering research area [57, 130]. Using the L_1 and L_2 norms the error is quantified between a work image sampled at the highest level for each method and an importance-biased image. First, an error image E is calculated by subtracting the degraded image D from the work image W. The L_1 ratio L_2 ratio is calculated using the following equations:

$$L_{1rat} = ||E||_1 / ||W||_1 \quad (5.6)$$

$$L_{2rat} = ||E||_2 / ||W||_2 \quad (5.7)$$

where:

$||.||_1$ is the L1 norm;
$||.||_2$ is the L2 norm;
E is the difference image between the final work image and the importance-biased image;
W is the final high-quality work image generated.

The norm values are then used as a numeric quantification of the error between the two images for both the progressive and supersampling methods. The L1 error ratio value indicates the maximum column sum value of the error image E [49]. This is defined by the following equation:

$$\|M\|_1 = \max_k \sum_{i=1}^{n} |m_{ik}| \quad (5.8)$$

where:

M is the matrix being evaluated;
m_{ik} is the matrix element at the i^{th} row and k^{th} column.

As a consequence, the L1 value gives an indication of the largest error value between the two images.

The L2 error ratio value is derived from the maximum eigenvalue of the error image expression E^TE [142]. This is the closest matrix norm to the Euclidean norm for vectors and thus gives an overall value for the pixel by pixel *distance* between the degraded and work images–in the way an inner product of two functions works with vectors. Let λ_1, λ_2,..., λ_n be the real Eigenvectors of matrix M. The following equation details the matrix L2 norm:

$$\|M\|_2 = |\lambda_1| \quad |\lambda| = \max_i |\lambda_i| \tag{5.9}$$

The presentation of both the L1 and L2 values give a good characterisation of the differences between the images by presenting both the maximum and overall distance. Due to the degradation being spatially non-uniform and thus creating large differences in some areas and not in others, it is expected that the L1 error ratio values may be high in comparison to the L2 error ratio values. Furthermore, due to the same effect, the L1 and L2 values may give opposite results when comparing importance-biased and original methods.

An error image has also been presented to give a visual indication of the locations of the differences between the images rendered with and without a visual importance bias. The error values in the image have been negated and thresholded, in order to aid visualisation and reproduction. That is, the dark pixels indicate locations where image differences occur, but they do not illustrate the magnitude of the differences. Furthermore, image regions have been highlighted and magnified to help illustrate the differences between the biased and non-biased images in regions considered insignificant by the visual importance model.

5.6.2 Objective Progressive Rendering Evaluation

The results for evaluation of the progressive rendering approach for the three test scenes are listed in the following sections. Each section contains: images sampled at a number of points in the rendering process by both the base and importance-biased systems, tables of L1/L2 norm ratios, pixel sampling images, contour importance maps and a discussion of the results for the evaluated scene.

Head Scene

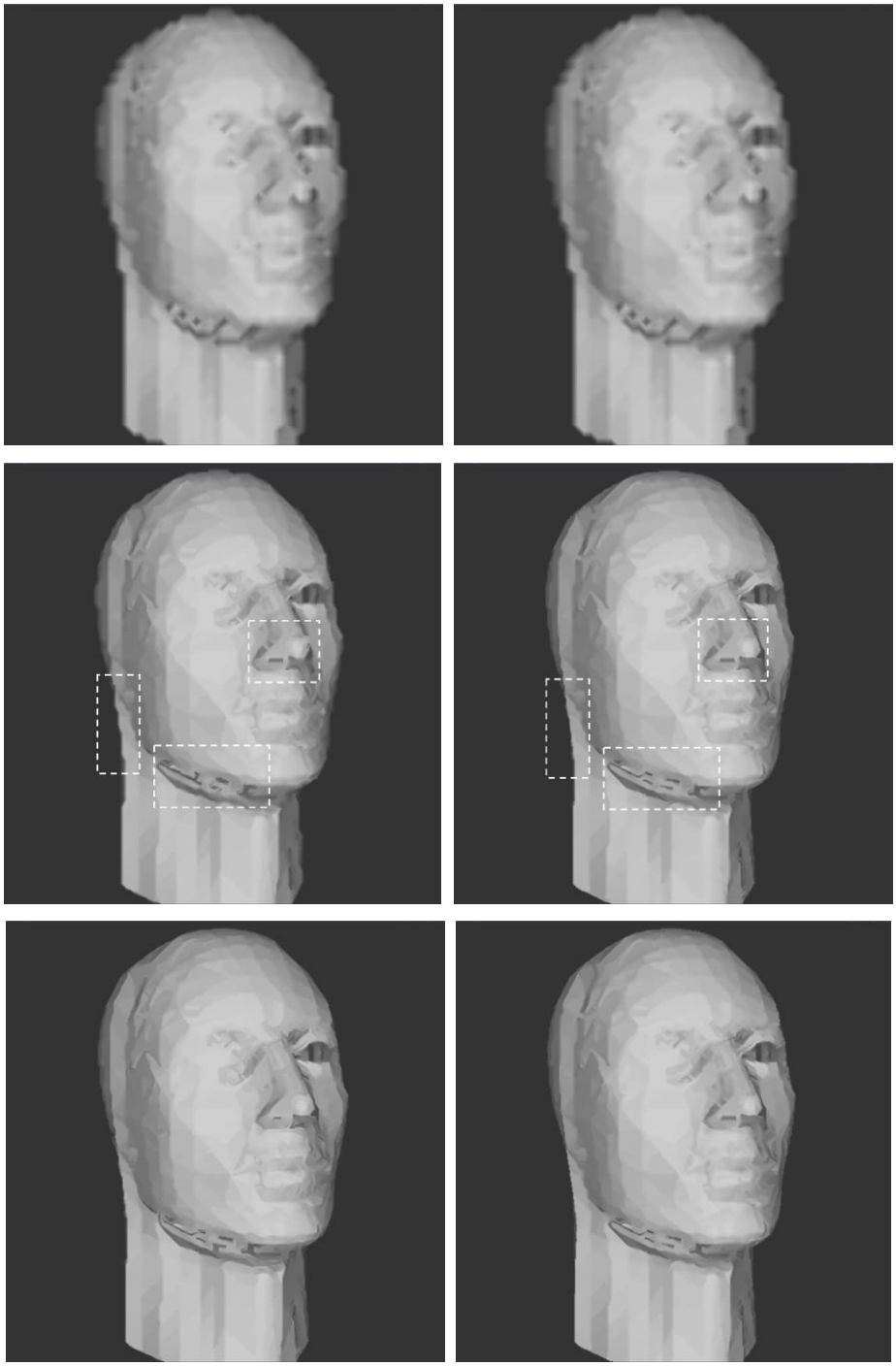


Figure 5.10 A series of images illustrating the improvement brought about by the use of importance acceleration. The images on the left are base images using the normal DCM method of sampling, while the images on the right are accelerated using the new method. The first image is 1.6% sampled, the second is 8% sampled-where the improvement is most discernable-and the final image is 10% sampled. The dashed rectangles highlight areas of greatest difference.

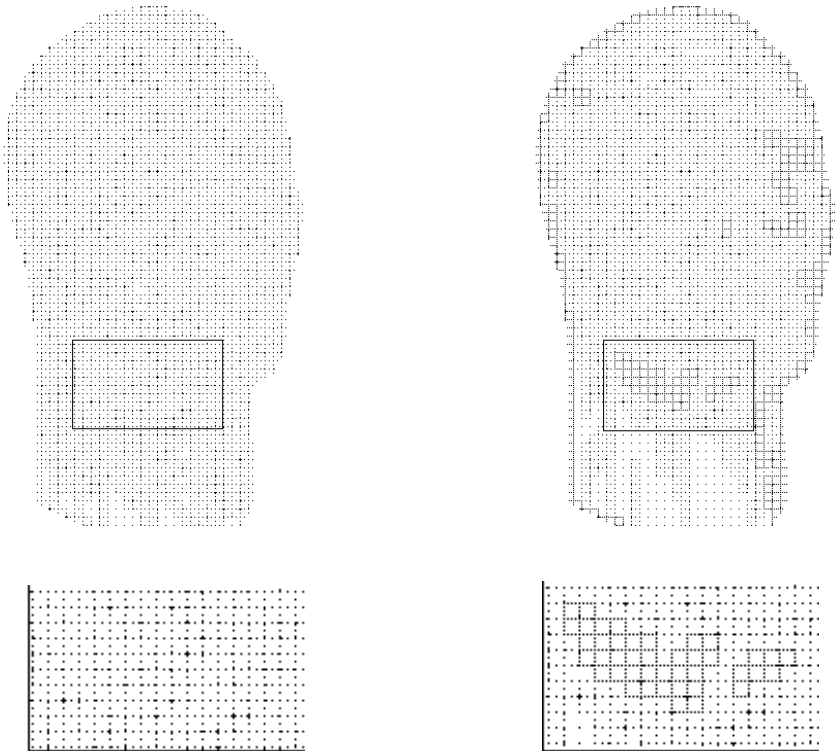


Figure 5.11 A comparison of the sampling performed for the 8% image, which shows the most improvement. The base method is shown on the left and the accelerated method on the right. The rectangle in each image has been magnified and placed underneath, highlighting some of the subdivisions that have been selected for accelerated refinement.

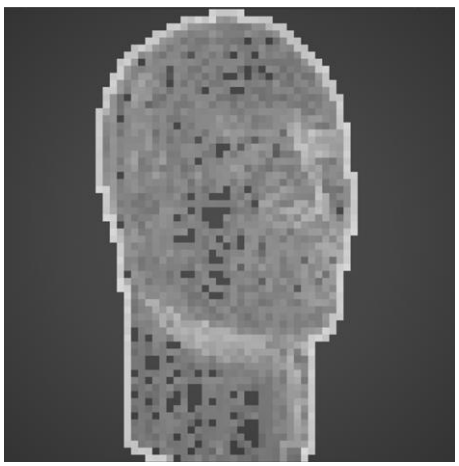


Figure 5.12 The contour importance map generated by the system. The bright subdivisions are the most visually important.

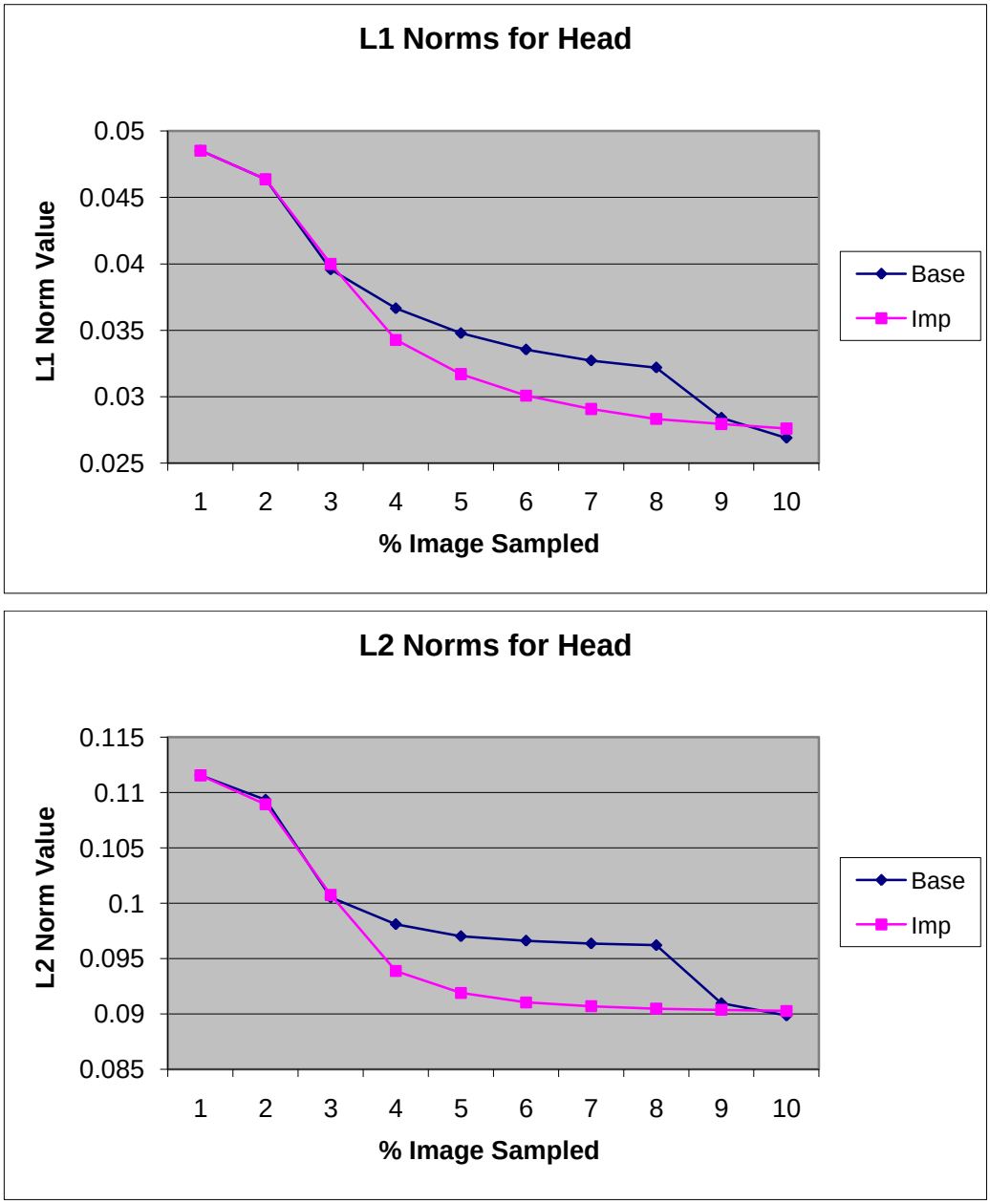


Figure 5.13 Graphs of relative L1 and L2 norm ratios for images at 1% sampling intervals, with the non-importance method marked as *Base* and the new visual importance method marked as *Imp*.

% Image Sampled	L1 Ratio Difference	L2 Ratio Difference
1	0.0000	0.0000
2	0.0000	0.0004
3	0.0004	0.0002
4	0.0024	0.0042
5	0.0031	0.0051
6	0.0035	0.0056
7	0.0037	0.0057
8	0.0039	0.0057
9	0.0005	0.0006
10	0.0007	0.0004

Table 5.2 Table of L1 and L2 differences shown in Figure 5.10. The table entries are calculated by taking the absolute value of the differences between the base and accelerated norm values, at the respective sample percentage.

For the Head image, there is a visible improvement in the visual quality of the image at the 8% sampling point. The white rectangles in Figure 5.10 highlight the areas which have improved in quality. This is also exhibited in the norm graph and the difference table, with the largest difference between the two images being at 8% of the image sampled. The images converge again in quality, due to the convergence of both the methods upon the final image at a later stage in the rendering process.

Kitchen Image

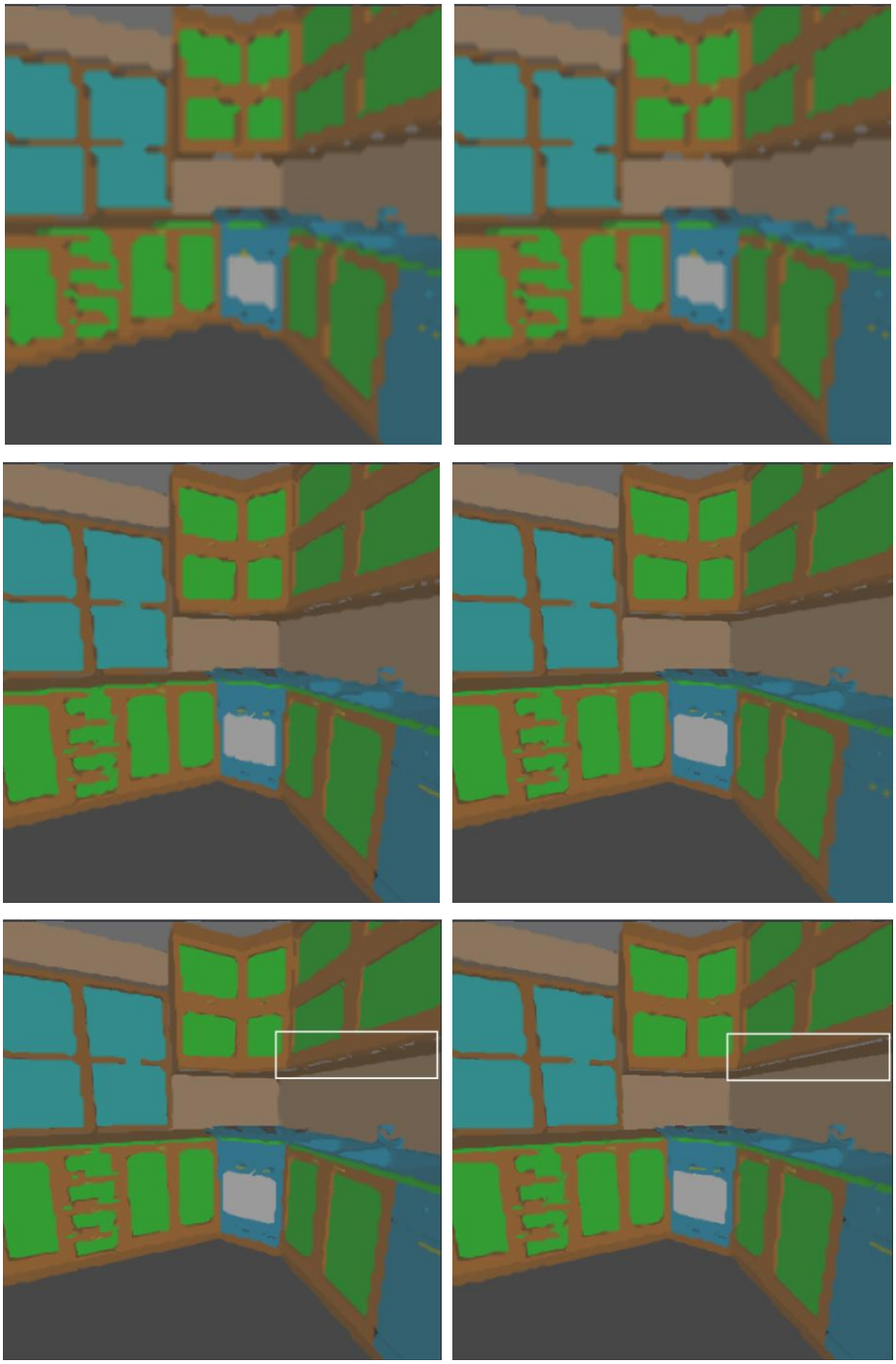


Figure 5.14 Progressively rendered images of the kitchen scene. The images in the left column are rendered using the base system, while the images on the right are rendered with the importance-based acceleration method. The top row of images is 1.6% sampled, the middle row is 8% sampled and the bottom is 10% sampled. The white rectangle highlights a refined area within the 10% sampled image.

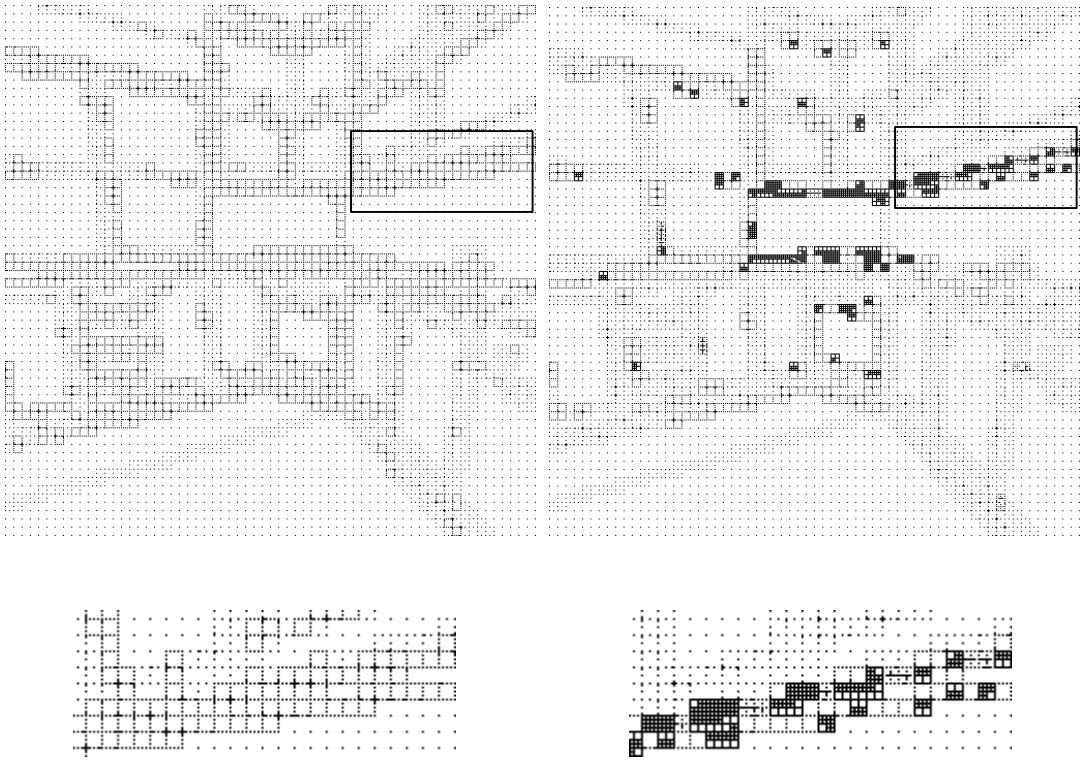


Figure 5.15 A comparison of the sampling performed for the 10% image. The base method is shown on the left and the accelerated method on the right. The rectangle in each image has been magnified and placed underneath, highlighting some of the subdivisions that have been selected for accelerated refinement.

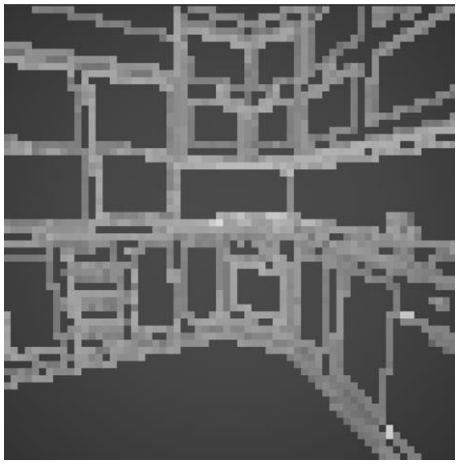
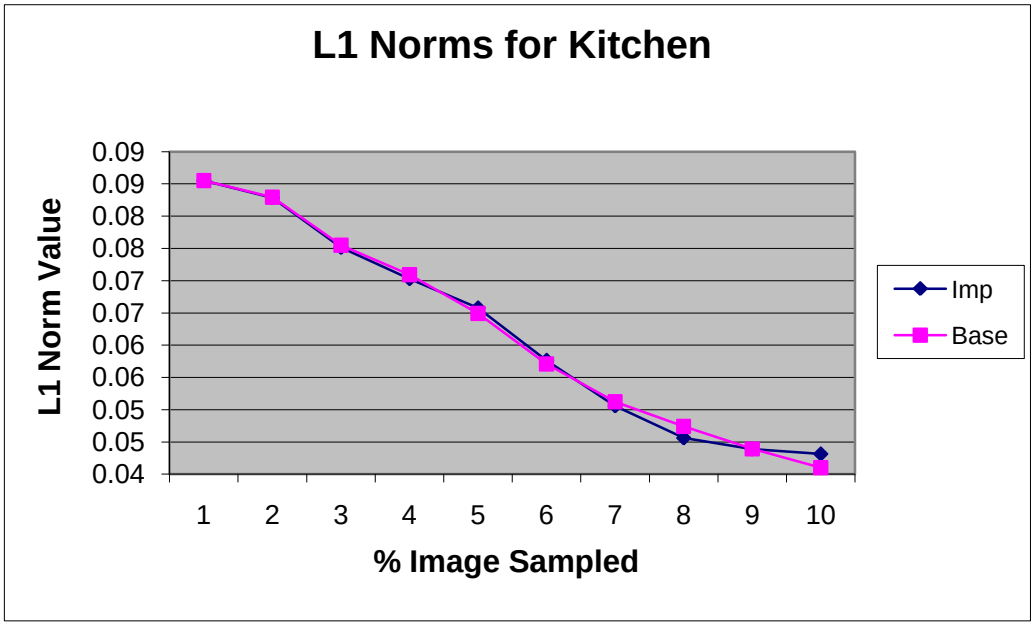


Figure 5.16 Contour importance map of the kitchen scene. Visually important subdivisions produce lighter coloured squares.



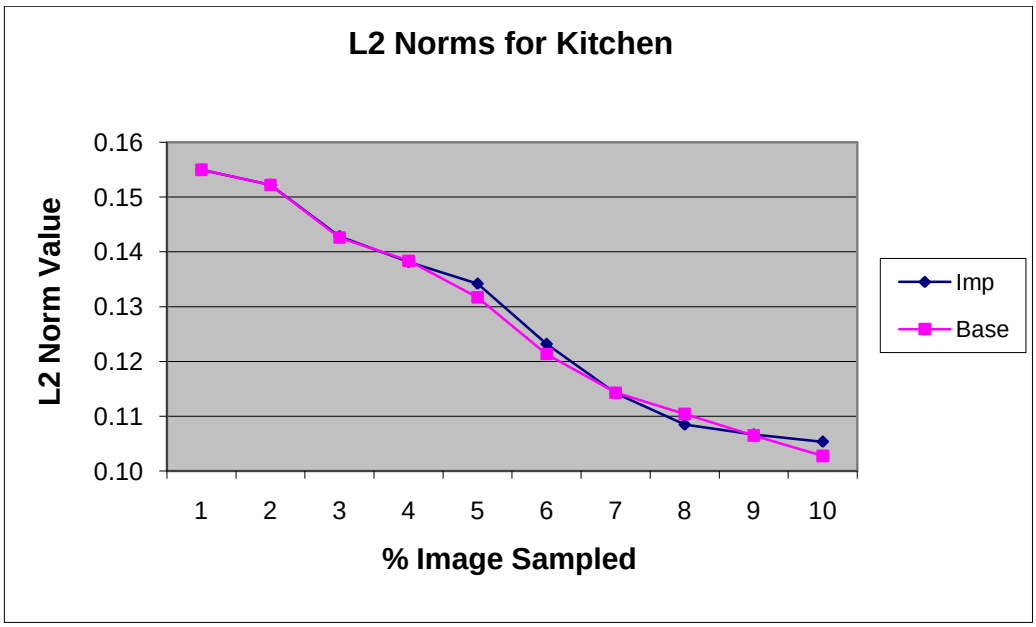


Figure 5.17 Graphs of relative L1 and L2 norm ratios for images at 1% sampling intervals, with the non-importance method marked as *Base* and the new visual importance method marked as *Imp*.

% Image Sampled	L1 Ratio Difference	L2 Ratio Difference
1	0.0000	0.0000
2	0.0001	0.0000
3	0.0004	0.0002
4	0.0007	0.0002
5	0.0008	0.0025
6	0.0006	0.0019
7	0.0007	0.0001
8	0.0018	0.0020
9	0.0000	0.0002
10	0.0021	0.0026

Table 5.3 Table of L1 and L2 differences shown in Figure 5.17 for the kitchen scene. The difference values are calculated by taken the absolute value of the differences between the base and accelerated images, at the respective number of samples.

Compared to the head scene the kitchen scene did not exhibit the same level of image quality improvement via importance acceleration. The subdivisions that have been improved in quality have not improved the overall impression of the quality of the scene. Significantly, the L1 and L2 objective measures have not shown any real discernible differences.

Farm Images

Figure 5.18 Progressively rendered images of the farm scene. The images in the left column are rendered using the base system, while the images on the right are rendered with the importance-based acceleration method. The top row of images is 1.6% sampled, the middle row is 8% sampled and the bottom is 10% sampled. The white rectangles highlight and compare refined regions from both methods.

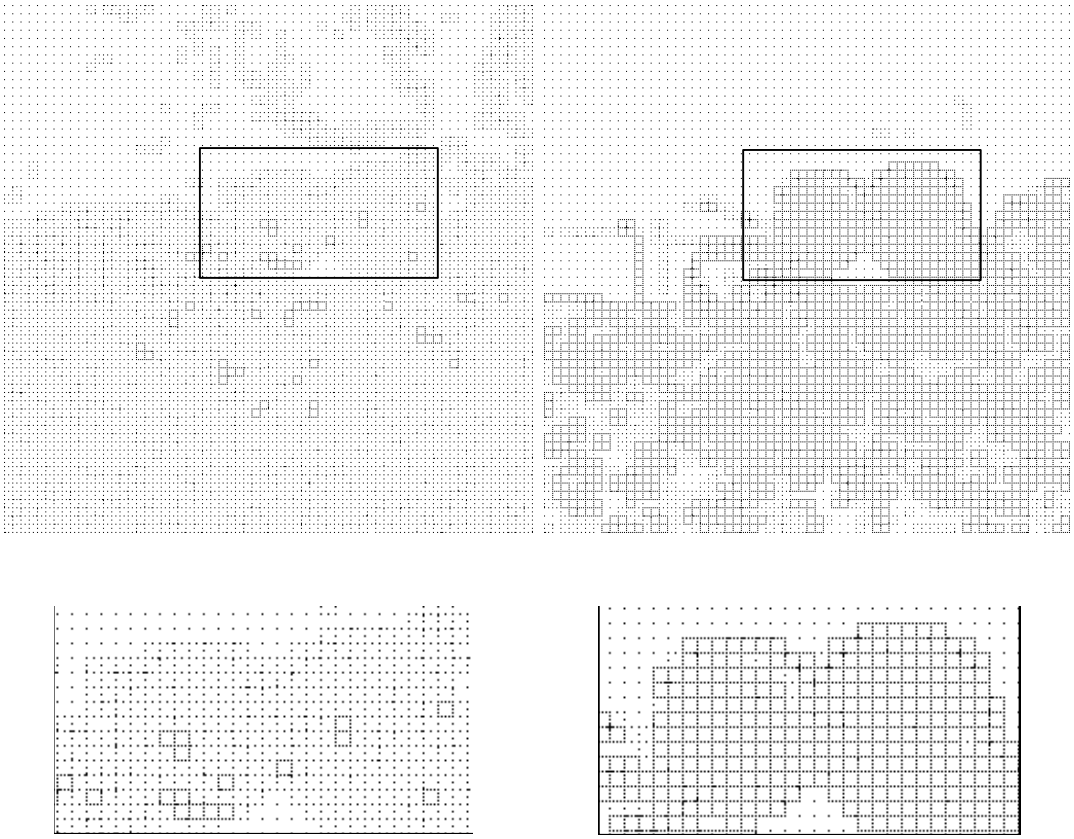


Figure 5.19 A comparison of the sampling performed for the 8% image. The base method is shown on the left and the accelerated method on the right. The rectangle in each image has been magnified and placed underneath, highlighting some of the subdivisions that have been selected for accelerated refinement.

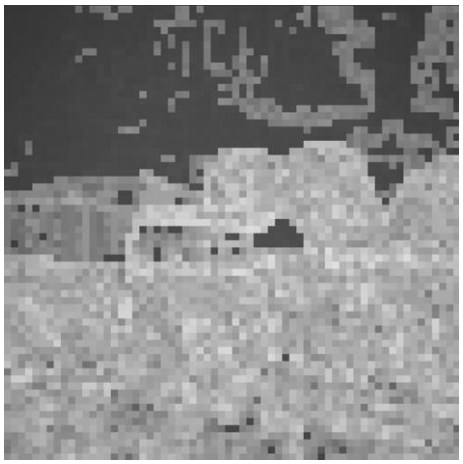


Figure 5.20 Contour importance map of the farm scene. Visually important subdivisions produce lighter coloured squares.

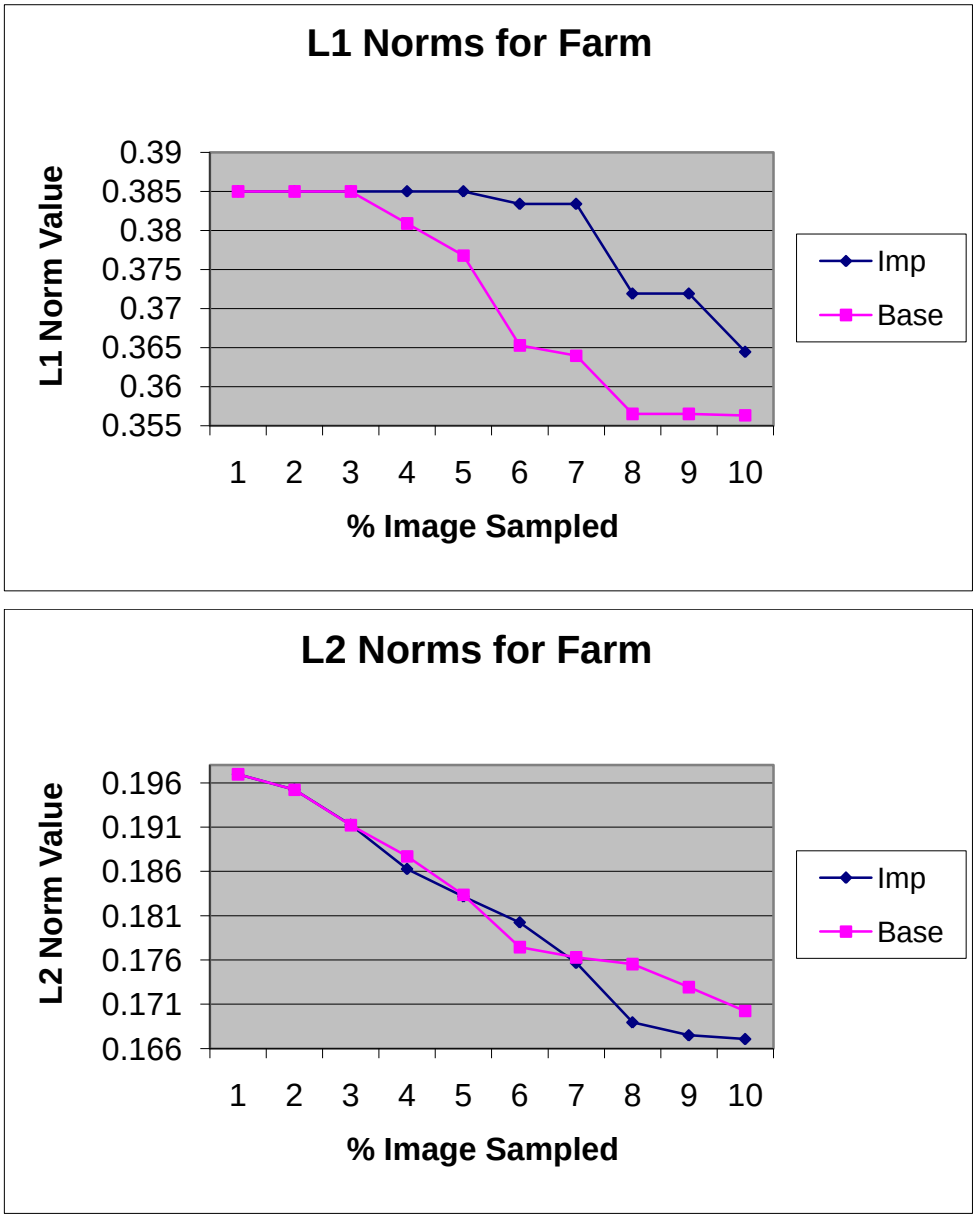


Figure 5.21 Graphs of relative L1 and L2 norm ratios for images at 1% sampling intervals, with the non-importance method marked as *Base* and the new visual importance method marked as *Imp*.

% Image Sampled	L1 Ratio Difference	L2 Ratio Difference
1	0.000000	0.000000
2	0.000000	0.000024
3	0.000000	0.000037
4	0.004102	0.001427
5	0.008218	0.000192
6	0.018119	0.002791
7	0.019440	0.000627
8	0.015411	0.006583
9	0.015411	0.005437
10	0.008145	0.003164

Table 5.4 Table of L1 and L2 differences shown in Figure 5.17 for the farm scene. The difference values are calculated by taken the absolute value of the differences between the base and accelerated images, at the respective number of samples.

Despite its complex nature and number of contours, the farm scene still exhibits visible levels of improvement within the region outlined by the white rectangle in Figure 5.18. This is also supported by the difference between the values in the L2 norm graph (refer to Figure 5.21) and the table of values (refer to Table 5.4). Due to the textured nature of the image, the boundaries between dissimilar textures become very important contours. Therefore, the scene can be considered to be similar in content to the head image, with only a few important contours contributing to the visual quality of the image. As a result, any improvement in these major contours may be quite noticeable. However, some obvious aliasing effects occurred on the barn, above the door. Subjective testing carried out in Chapter 9 addresses any perceptual impressions of the differences between the images. Furthermore, the large difference with the L1 norm values is expected, due to the L1 norm indicating the sum differences between the images. Most of these large differences occur within the heavily textured regions, and so are invisible due to the masking effects of the surrounding texture.

Discussion

The images show that as more sampling occurs, the two rendering methods converge on the same final image. What is evident, though, is the effectiveness of using a measure of contour importance and utilising this as a heuristic to guide the further

refinement of the scene. The early image quality increases are due to this acceleration.

The method tends to work for an image that contains high levels of order and small number of contours. If an image is like the head scene, containing only a small proportion of contours for the whole image, then the method has some leeway to apportion more detail to areas than others, without losing image quality in toto. With more complex scenes like the kitchen, the ability to apportion extra detail in selected areas is reduced markedly. This is indicated by the L1 and L2 norm graphs generated for the kitchen scene. The objective difference between the images is of an almost insignificant magnitude. Perusal of the images subjectively confirms this observation.

This effect is even more prevalent within spatially noisy scenes, like the farm. This scene, as indicated in the contour importance map, essentially contains a contour in every subdivision, thereby making it hard to improve the quality of the image by accelerating the refinement of certain contours. Secondly, the noise in the image introduces inherent masking effects, due to the superposition of one frequency upon another in the scene [166]. The refinements are therefore lost within the noise generated by the textures within the scene.

Another issue is the degradation introduced into the rest of the image, by the reassignment of the samples to those subdivisions deemed to have important contours. This means that while some contours are improved, others are degraded and thus cause further aliasing. The issue here is the effect on the subjective visual quality of the whole image. Potentially the quality of the image could be compromised by the degradation of unimportant contour subdivisions. On the contrary, the improvement in the important contours could enhance the overall quality of the image. This is dependent on the contents of the image. For example, the barn in the farm image (Figure 5.18) is left unrefined as its contours are relatively unimportant, compared to the edges of the trees. The subjective testing section of this thesis (Chapter 8), deals with this issue in more detail.

5.6.3 Objective Supersampling Evaluation

An objective evaluation methodology was also applied to the images generated by the region-based supersampling method. Objective evaluation of the images was carried out in the following manner. The supersampling method (flat or perceptual) is used to render a work image at a predefined level of quality, without importance biasing. A degraded image using the same parameters is then rendered using the newly developed visual importance model. A difference image is derived from these two images, along with L1 and L2 ratios as defined in Section 5.6.1. The difference image is used to indicate the spatial location of differences between the two images with darkened pixels, but does not indicate the magnitude of the difference. This method is used to give an indication of the quality of an image at various levels of degradation, with reference to the supersampling method being used.

The first set of images have been generated using the flat-rate supersampling method with subdivision rates being 2, 3 and 4 subdivisions per pixel. This gives sampling rates ranging from 4 to 16 samples per pixel, using a regular sampling distribution. The work image is sampled at a constant rate for each pixel, whereas the region-biased image is sampled at a rate depending on the importance of the region (refer to Section 5.4.1).

For the perceptual method of supersampling control (refer to Section 5.4.2) a similar method of testing was performed. The image difference predictor used in this method contains an arbitrary threshold parameter [114]. This threshold is the amount of error to be tolerated before requiring a refinement of the image. If the absolute error between the newly refined image and the old image is larger than the predefined threshold, then the image is further refined. In the experiments performed for this chapter, the threshold has been varied from 10 to 50. The upper limit of 50 is due to the work and biased images being essentially the same after the threshold value passes 50.

Tables of L1/L2 norm values and relative timings for each parameter value have also been generated for each scene. Relative and not absolute timings were used, as they

provide relevant information about the efficiency gains reaped from the importance methods. Absolute values are not so informative, due to the inefficient prototype nature of the rendering system developed. Small example sections of the images have also been cropped and magnified to provide examples of the image distortion caused by modification of the sampling rate across the image. The above objective methodology has been applied in the following sections to the head, kitchen and farm scenes respectively.

Head Scene Flat-Rate

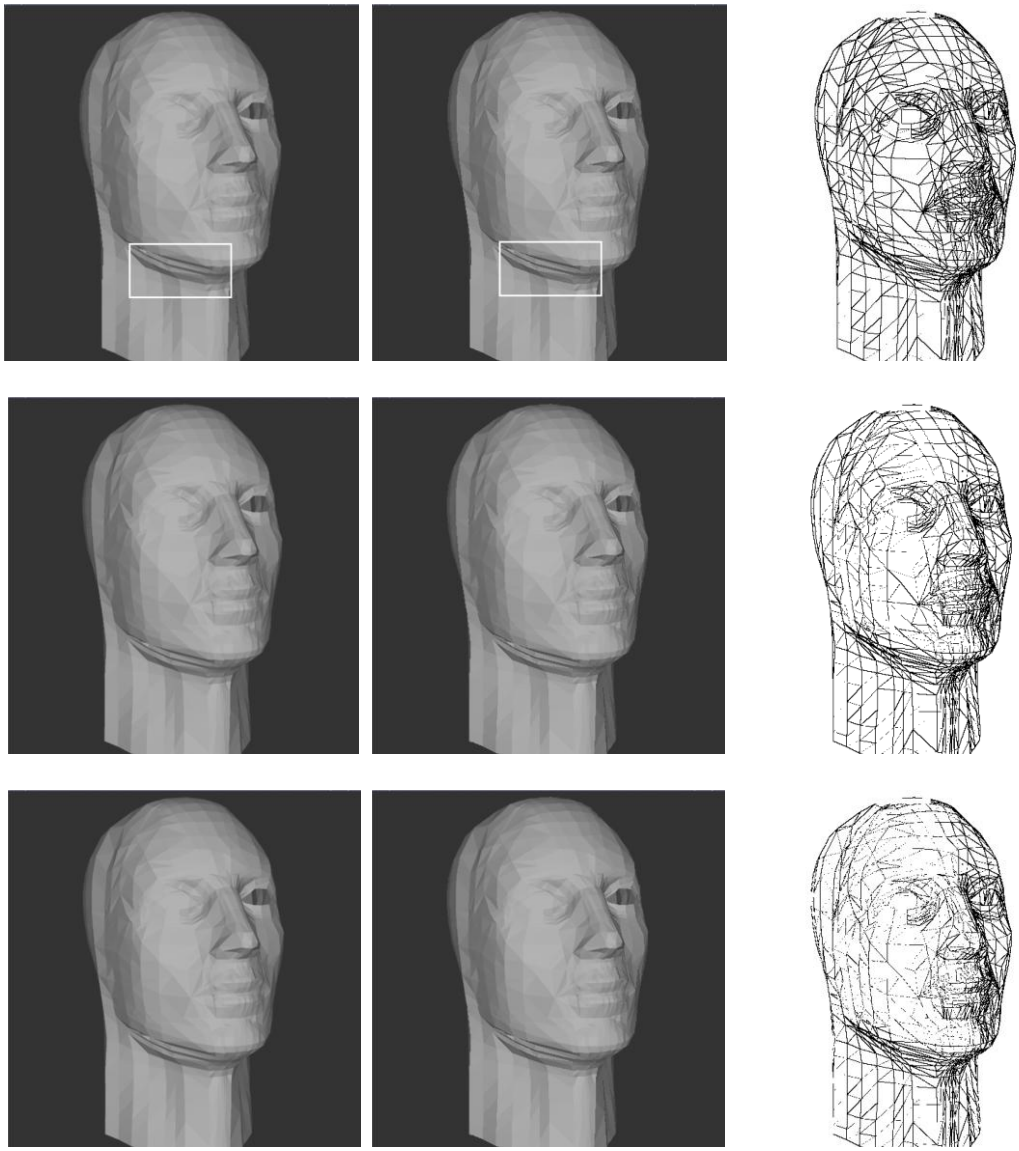


Figure 5.22 A series of images showing the output from the flat-rate method. The images on the left are the work images generated at a constant level of pixel supersampling. The middle images have been generated using a region-biased method. The difference between the images is shown on the right. The rows represent the maximum number of samples per pixel with the top row being 4 the middle 9 and the bottom 16 samples per pixel respectively.

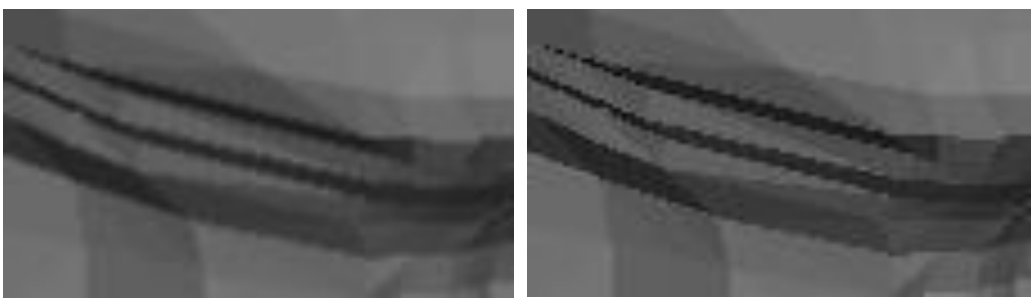
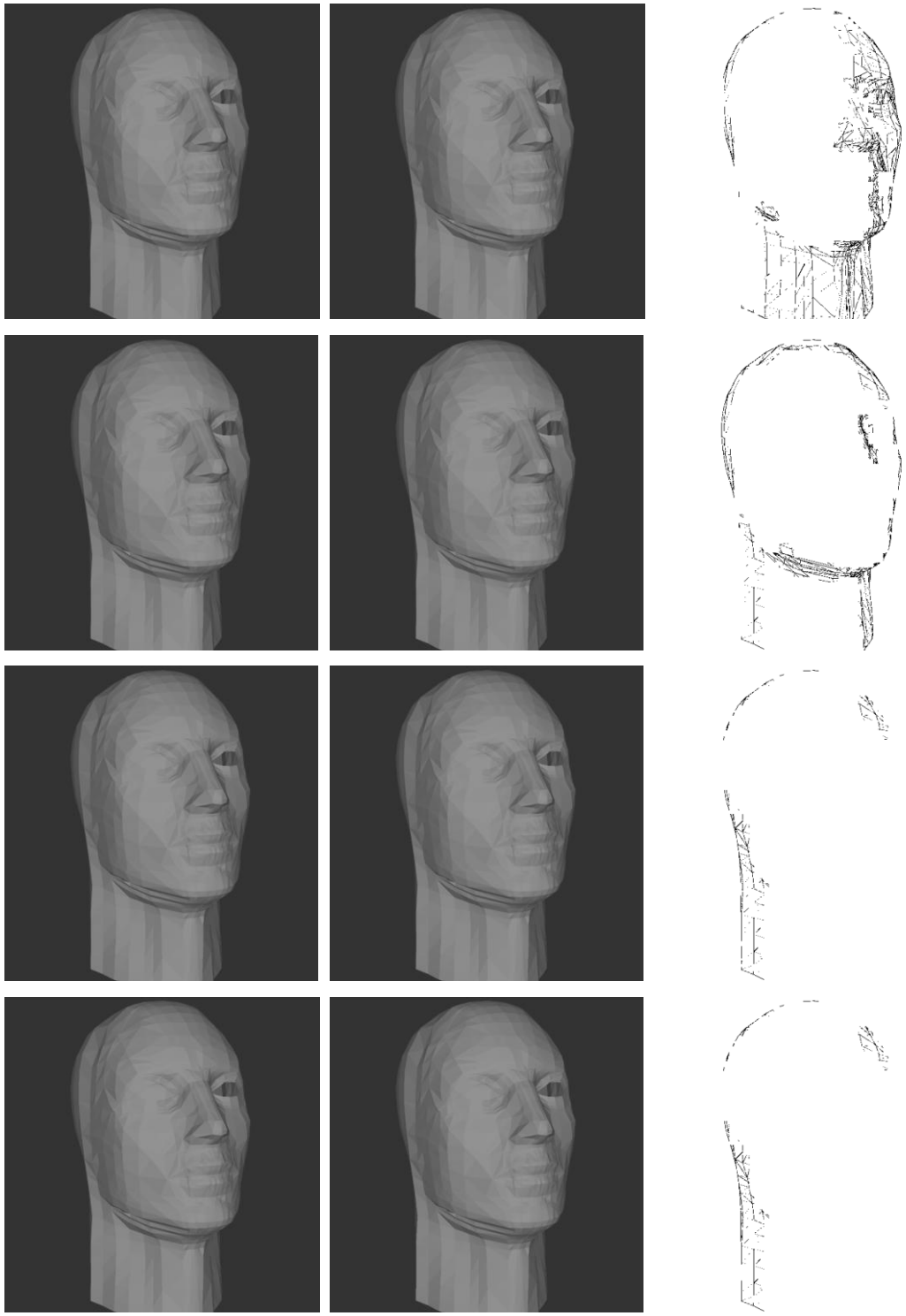


Figure 5.23 Illustration of quality differences caused by the reduction in pixel sampling within the white rectangles shown in Figure 5.22. The base image is on the left, while the biased image is on the right.

Maximum Sample Rate	Average Samples Per Pixel	Relative Time With Respect to Non-biased Image	L1 Ratio (L2 Ratio)
Non-biased 4	4.00	-	-
Region-biased 4	1.01	0.5	0.0142 (0.0462)
Non-biased 9	9.00	-	-
Region-biased 9	1.83	0.3	0.0094 (0.0386)
Non-biased 16	16.00	-	-
Non-biased 16	3.07	0.3	0.0086 (0.0388)

Table 5.5 Results of the flat-rate rendering methodology showing samples, relative times and norm error ratios for each image generated with or without attention-based biasing, at varying levels of fidelity.

Head Scene Perceptual



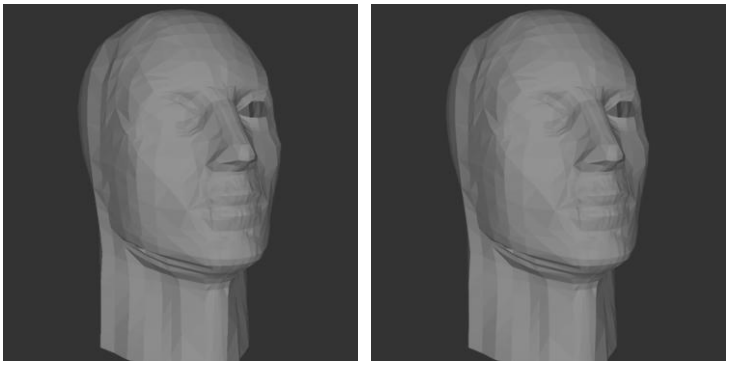


Figure 5.24 Images generated using the perceptual method, with a high quality work image on the left, the importance-biased image in the middle, and the difference between the two on the right. The rows represent the error threshold measure used to control the quality of the image; ranging from 10 in the top row to 50 in the bottom row.

Perc Method and Threshold	Average Samples Per Pixel	Relative Time With Repect to Non-biased Image	L1 Ratio (L2 Ratio)
Non-biased 10	79.31	-	-
Region-biased 10	7.75	0.1	0.0102 (0.0037)
Non-biased 20	14.51	-	-
Region-biased 20	7.04	0.5	0.0051 (0.0027)
Non-biased 30	10.51	-	-
Region-biased 30	6.72	0.7	0.0159 (0.0033)
Non-biased 40	7.59	-	-
Region-biased 40	4.00	0.6	0.0334 (0.0072)
Non-biased 50	4.00	-	-
Region-biased 50	4.00	1.0	0.0000 (0.0000)

Table 5.6 Results of the perceptual rendering methodology showing samples, relative times and norm error ratios for each image generated with or without attention-based biasing, at varying levels of fidlelity.

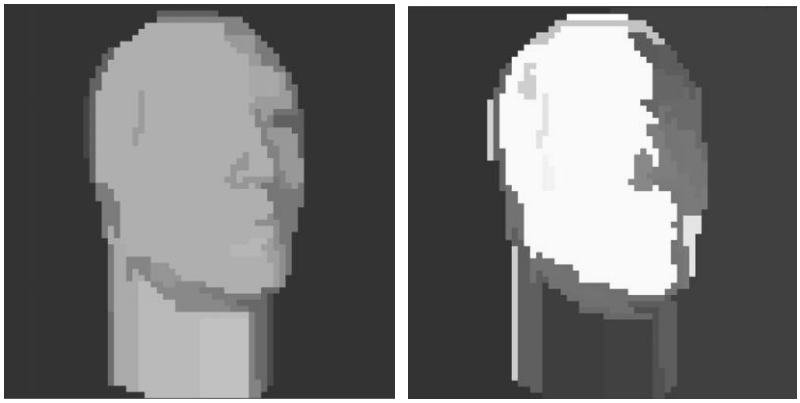


Figure 5.25 Region segmentation images, with the raw segmentation on the left, coloured with random grey shades to indicate the segmentation performed. On the right is the region importance map generated, with the lighter regions being assigned higher importance values, ranging over [0.0, 1.0].

With the head scene, a number of points can be made. Firstly, both flat and perceptual methods reap large savings in the rendering of images, of at least a half the time for a normal rendering. Secondly, the flat method of sampling modulation, while quick and efficient, is insensitive to contours within the scene. As can be seen from the difference images in Figure 5.22, the differences fall mainly on the contours. This degradation of quality within the edges causes unsightly aliasing to appear along certain edges (refer to Figure 5.23). It improves with the increase in maximal supersampling to sixteen samples per pixel, but still presents a problem as a method of supersampling when used with region biasing.

The perceptual method of supersampling is more sensitive to the presence of edges within the scene, due to the use of the contrast sensitivity function within its algorithm. This occurs almost evenly, no matter how large the error threshold. In the table of norm values for the perceptual method, it can be seen that the method again saves half the time for a perceptual rendering. It also can be noted that the method improves in comparison to the base perceptual method as it renders a more and more error filled image. Overall, the quality of the scenes using the perceptual method is much better than the flat supersampling method.

Kitchen Scene Flat-Rate

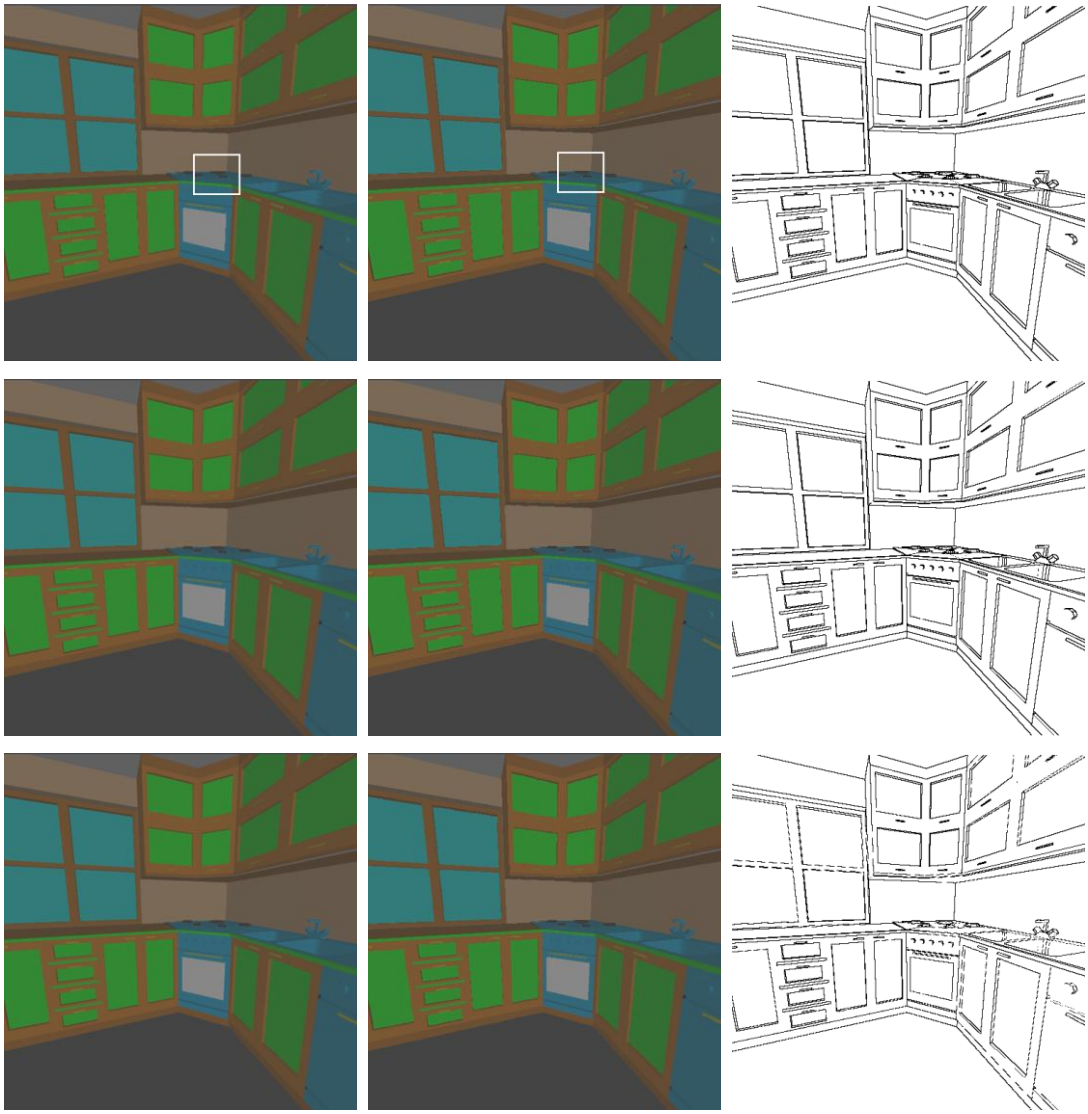


Figure 5.26 A series of images showing the output from the flat-rate method. The images on the left are the work images generated at a constant level of pixel supersampling. The middle images have been generated using a region-biased method. The difference between the images is shown on the right. The rows represent the maximum number of samples per pixel with the top row being 4 the middle 9 and the bottom 16 samples per pixel respectively.

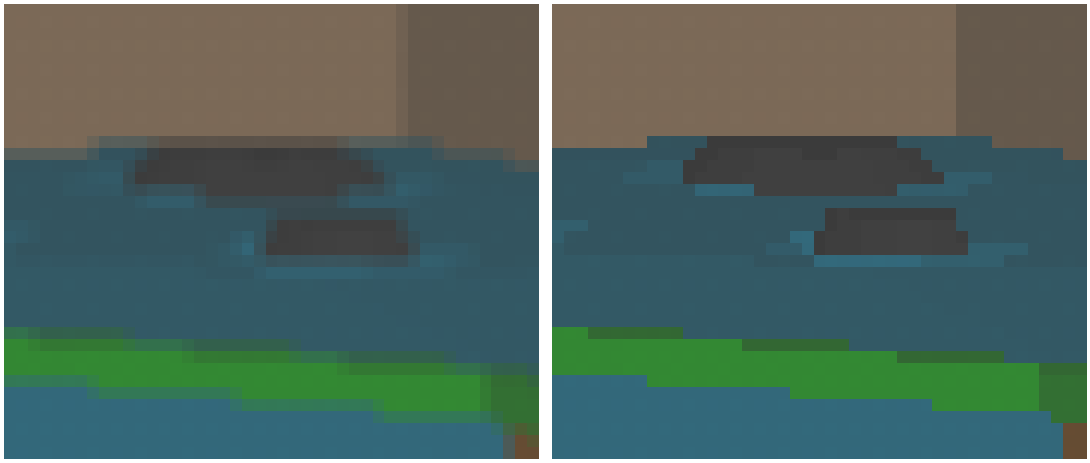
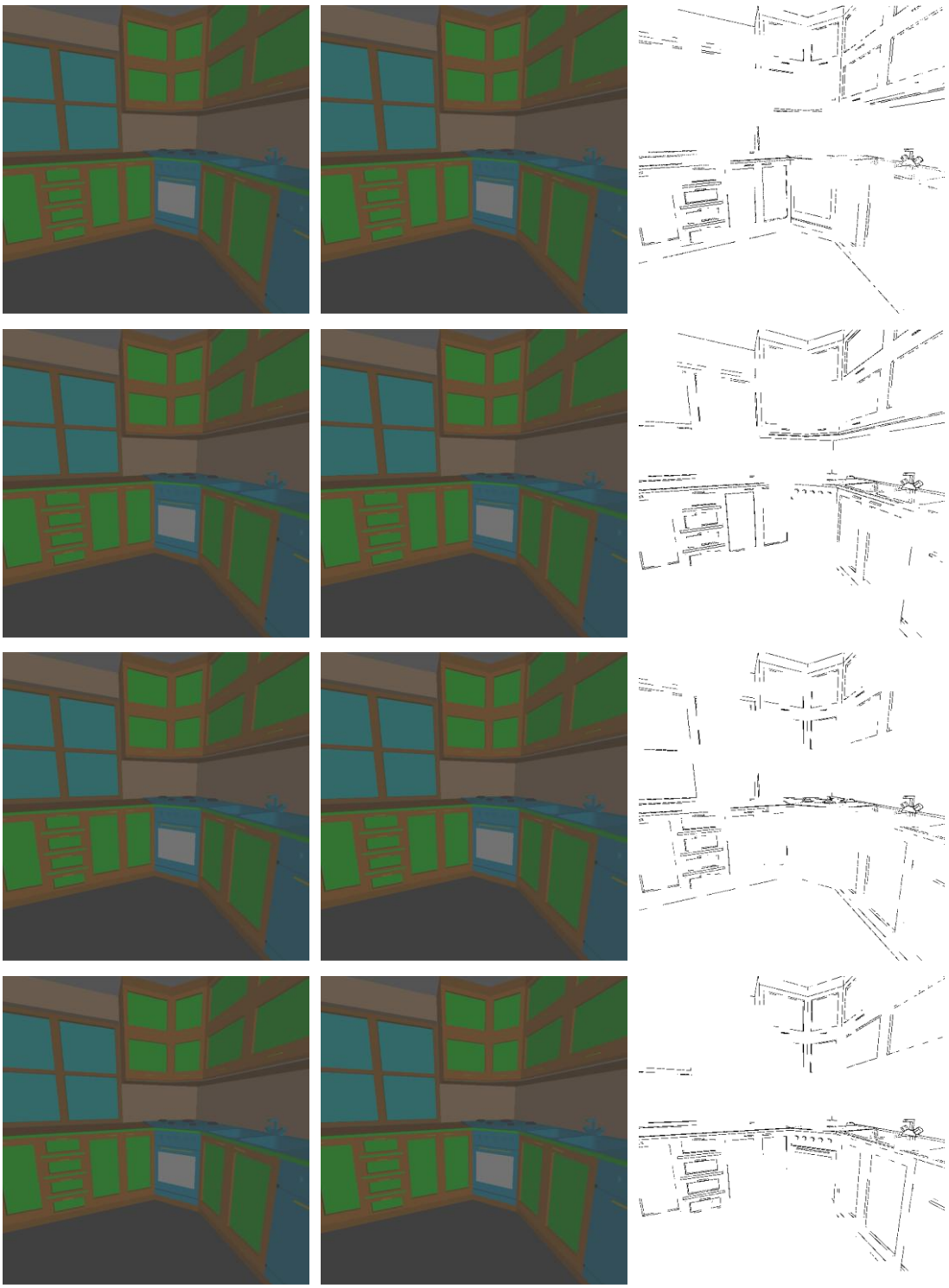


Figure 5.27 Blown up illustrations of the differences in image quality between kitchen images within the region highlighted by white rectangles in Figure 5.26.

Maximum Sample Rate	Average Samples Per Pixel	Relative Time With Respect to Non-biased Image	L1 Ratio (L2 Ratio)
Non-biased 4	4.00	-	-
Region-biased 4	1.02	0.4	0.0184 (0.0250)
Non-biased 9	9.00	-	-
Region-biased 9	1.92	0.3	0.0236 (0.0220)
Non-biased 16	16.00	-	-
Non-biased 16	3.31	0.2	0.0163 (0.0166)

Table 5.7 Results of the flat-rate rendering methodology showing samples, relative times and norm error ratios for each image generated, with or without region biasing, at varying levels of fidlelity.

Kitchen Scene Perceptual



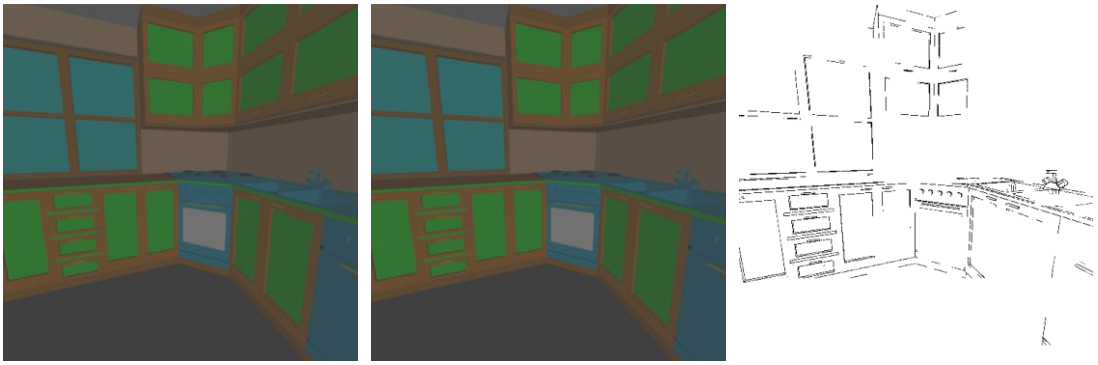


Figure 5.28 Images generated using the perceptual method, with a high quality work image on the left, the importance-biased image in the middle, and the difference between the two on the right. The rows represent the error threshold measure used to control the quality of the image; ranging from 10 in the top row to 50 in the bottom row.

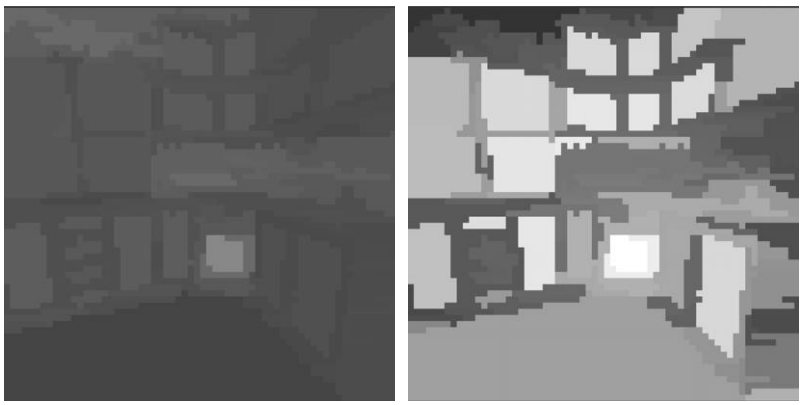


Figure 5.29 Region segmentation images, with the raw segmentation on the left, coloured with random grey shades to indicate the segmentation performed. On the right is the region importance map generated, with the lighter regions being assigned higher importance values.

Perceptual Method and Threshold	Average Samples Per Pixel	Relative Time With Respect to Non-biased Image	L1 Ratio (L2 Ratio)
Non-biased 10	55.42	-	-
Region-biased 10	31.22	0.6	0.0147 (0.0043)
Non-biased 20	23.40	-	-
Region-biased 20	15.58	0.7	0.0125 (0.0052)
Non-biased 30	14.38	-	-
Region-biased 30	7.54	0.6	0.0131 (0.0065)
Non-biased 40	9.27	-	-
Region-biased 40	5.89	0.6	0.0279 (0.0073)
Non-biased 50	7.54	-	-
Region-biased 50	4.07	0.6	0.0313 (0.0087)

Table 5.8 Results of the perceptual rendering methodology showing samples, relative times and norm error ratios for each image generated with or without attention-based biasing, at varying levels of fidelity.

Region-biased flat supersampling offers the same advantages and disadvantages for the kitchen scene as discovered for the head scene. Similar effects occur with the lack of sensitivity to edges in the scene. Subsequently, edges are aliased badly when the region-biasing is applied to flat supersampling. With the perceptual method of sampling, again, the sensitivity to contours in the image enables the method to produce less aliasing of the contours in the scene. Similar time savings are recorded with around half the time taken to render the scenes using a region-biased supersampling method.

Farm Scene Flat-Rate

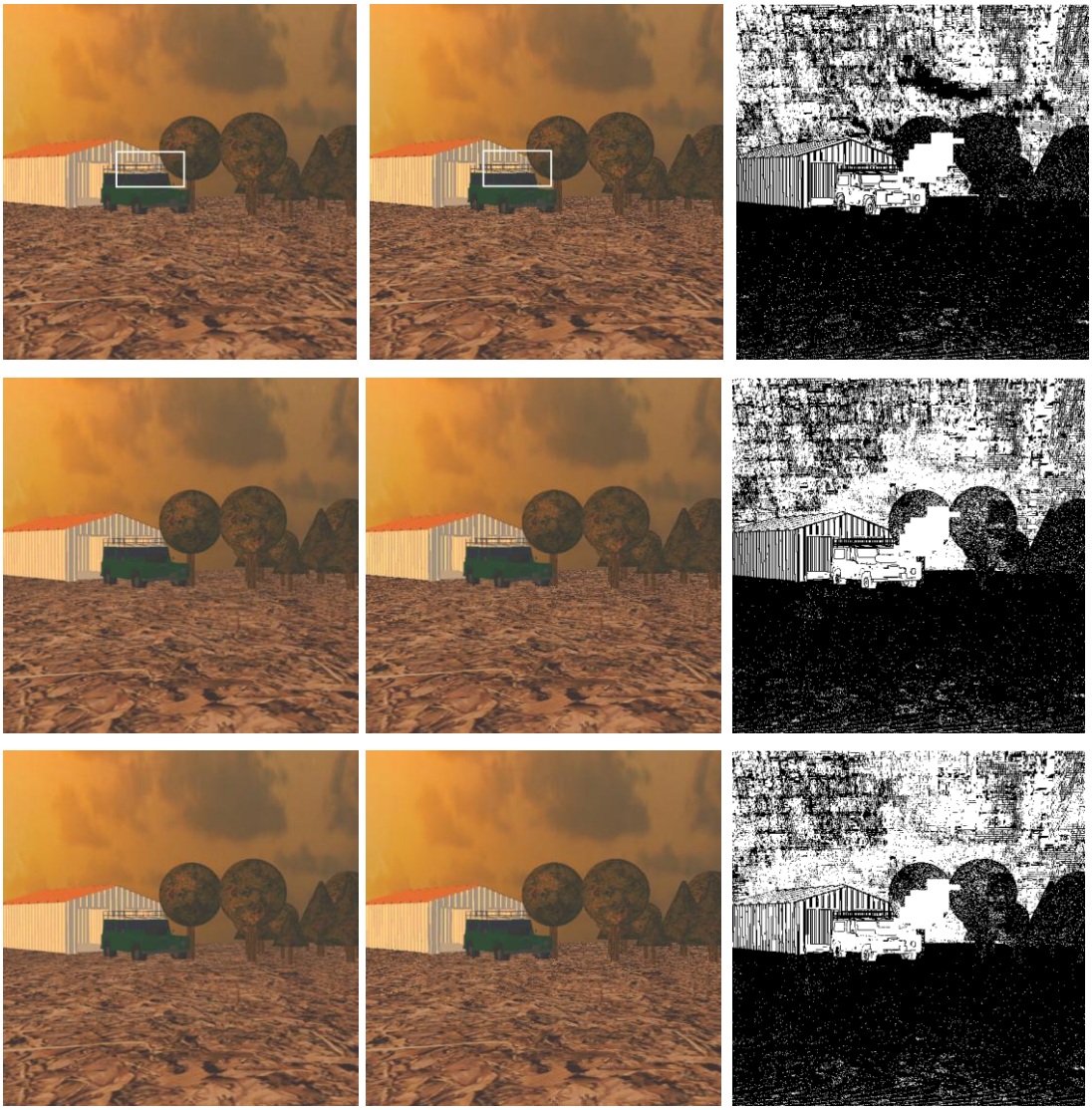


Figure 5.30 A series of images showing the output from the flat-rate method. The images on the left are the work images generated at a constant level of pixel supersampling. The middle images have been generated using a region-biased method. The difference between the images is shown on the right. The rows represent the maximum number of samples per pixel with the top row being 4 the middle 9 and the bottom 16 samples per pixel respectively.

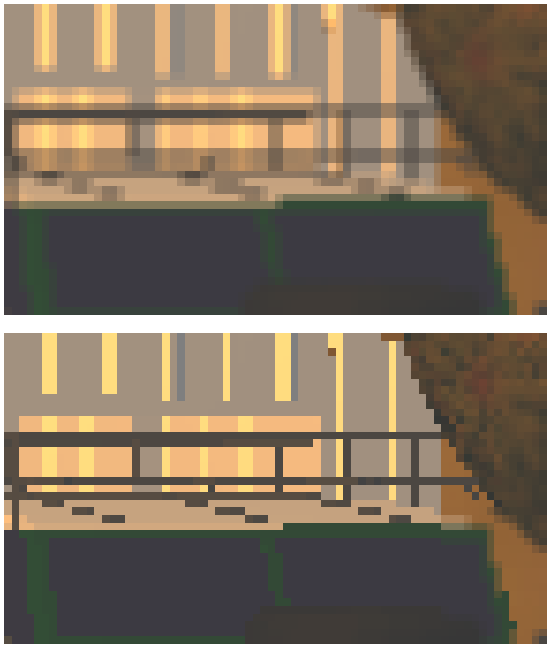
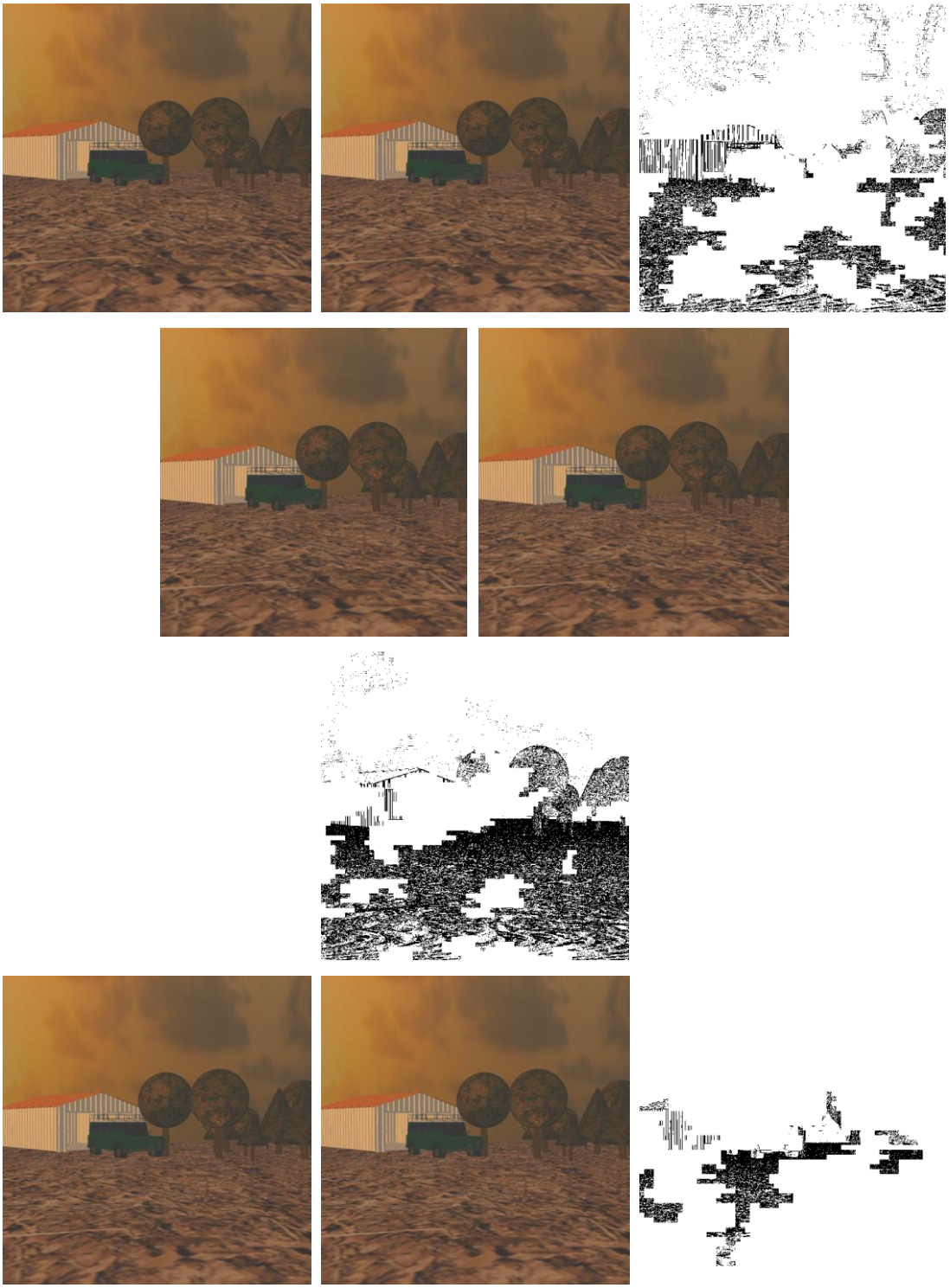


Figure 5.31 Blown up illustrations of the differences in image quality between farm images within the region highlighted by white rectangles in Figure 5.30.

Maximum Sample Rate	Average Samples Per Pixel	Relative Time With Respect to Non-biased Image	L1 Ratio (L2 Ratio)
Non-biased 4	4.00	-	-
Region-biased 4	1.05	0.3	0.1091 (0.1126)
Non-biased 9	9.00	-	-
Region-biased 9	1.62	0.2	0.1164 (0.1075)
Non-biased 16	16.00	-	-
Non-biased 16	2.37	0.2	0.0945 (0.1064)

Table 5.9 Results of the flat-rate rendering methodology showing samples, relative times and norm error ratios for each image generated, with or without region biasing, at varying levels of fidlelity.

Farm Scene Perceptual



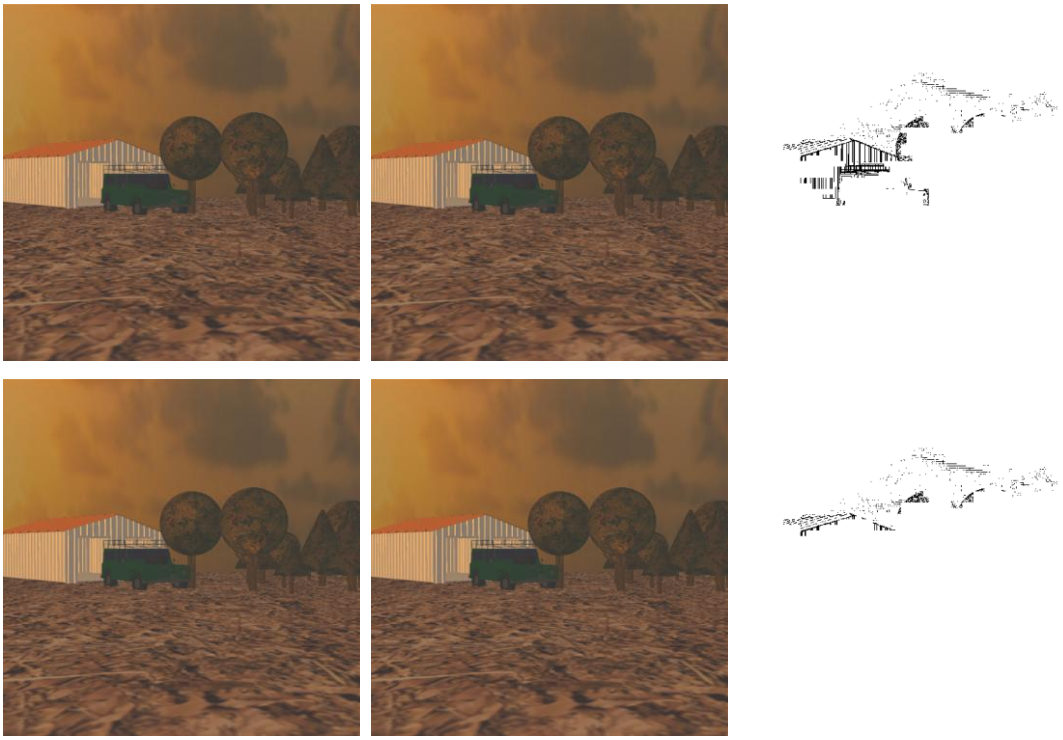


Figure 5.32 Images generated using the perceptual method, with a high quality work image on the left, the importance-biased image in the middle, and the difference between the two on the right. The rows represent the error threshold measure used to control the quality of the image; ranging from 10 in the top row to 50 in the bottom row.

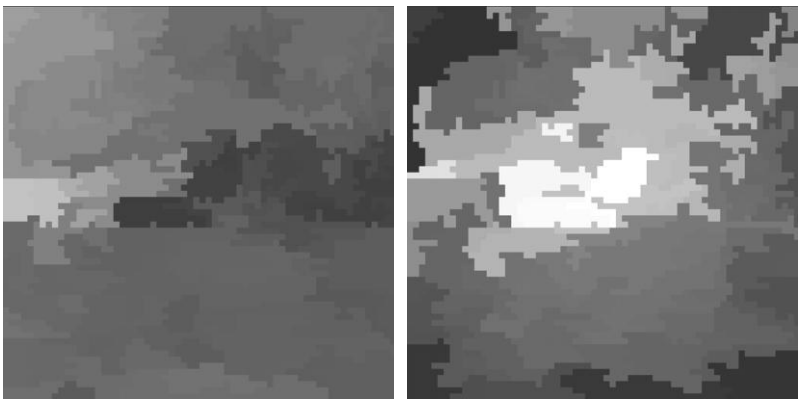


Figure 5.33 Region segmentation images, with the raw segmentation on the left, coloured with random grey shades to indicate the segmentation performed. On the right is the region importance map generated, with the lighter regions being assigned higher importance values.

Perc Method and Threshold	Average Samples Per Pixel	Relative Time With Respect to Non-biased Image	L1 Ratio (L2 Ratio)
Non-biased 10	32.83	-	-
Region-biased 10	19.16	0.6	0.0161 (0.0099)
Non-biased 20	18.11	-	-
Region-biased 20	6.48	0.4	0.0161 (0.0245)
Non-biased 30	6.50	-	-
Region-biased 30	5.24	0.8	0.0176 (0.0141)
Non-biased 40	5.14	-	-
Region-biased 40	4.14	0.8	0.0212 (0.0167)
Non-biased 50	4.73	-	-
Region-biased 50	4.00	0.8	0.0045 (0.0046)

Table 5.10 Results of the perceptual rendering methodology showing samples, relative times and norm error ratios for each image generated with or without attention-based biasing, at varying levels of fidelity.

In a similar manner to the head and kitchen scenes, the methodology is able to reduce the number of samples made in the areas that are visually less important. Similar effects are observed, including the most time savings occurring with the highest quality images.

5.7 DISCUSSION

Progressive rendering approaches in this chapter have been shown to benefit from judicious application of visual importance models. A large saving in time is gained from using the region-based supersampling method. The flat-rate method had the greatest gains in time efficiency. However, a significant amount of image difference resulted. The perceptual method generated better results according to the objective error values, and yet was still able to take less time to render the image. This quality difference was due, in part, to the method used to implement the perceptual image difference algorithm.

The perceptual supersampling method subdivided each pixel at least once, in order to detect any possible improvement from further supersampling. Therefore, this perceptual approach automatically provides an improved image quality. Related to this is a law of diminishing returns regarding the quality of the image in relation to subdivisions. The number of subdivisions is usually limited to four, as this antialiases most of the sampled frequencies in a typical synthesised image [172].

With the scenes presented in this chapter, continuing past four subdivisions with the perceptually controlled supersampling did not produce any visible differences. As a consequence, it seems that as long as the degradation does not create visible aliasing, then the loss of samples could occur anywhere in the image without being a viewing problem. The differences may only be visible with close comparison of images rendered with and without the visual importance method. Images viewed alone would probably not elicit a negative response from a viewer. This issue is discussed further in the subjective testing results discussion in Chapter 8.

However, the spatial frequency content of the image will alias at lower sampling rates, and so the rendering method used should be sensitive to this aliasing effect. This sensitivity is reliant on the perceptual sampling method used. To conclude, it would seem that any use of visual importance processing must incorporate a perceptual module that tests for the visibility of any spatial frequencies within the image. In the case of the method used in this chapter, the modification of the error tolerance level for each sampled rectangle would allow control over the differences between the work and the degraded images. Such parametric control would allow the user to degrade or enhance the image quality to a level acceptable for the intended application.

The contour importance approach has been shown to give a limited improvement with scenes of lower complexity. The error values and subjective inspection indicate a measure of improvement in the quality of the image with reference to the final fully rendered image. However, the method has been shown to struggle with scenes that contain a large number of complex contours, due to the problem of apportioning refinement to particular regions in the image over others, and thereby losing quality in those other regions. This is in essence a signal to noise problem. If the image is noisy, then any changes within the scene disappear due to the masking effects from the other numerous unrefined edges. It should be noted, however, that the importance-biased progressive method is at least able to maintain the quality of the image at the same sampling rate, with a different sampling strategy. This progressive methodology still has potential for simpler scenes, and extensions suggested in Section 9.2.2 may improve its performance with more complex images.

Incorporating Texture Importance into Adaptive Rendering

Image information is used in 3D rendering to remove the need for geometric modelling of all the details in a scene. Its effective handling can facilitate both the efficiency and fidelity of scene renderings. This chapter is an investigation into the further use of visual importance concepts, with regards to efficient texture mapping of 3D polyhedra.

The texture sampling approach has been incorporated into the sample generator outlined in Chapter 5. The technique utilises the concept of *texture coherence*, which is the similarity of the appearance of the texture in texture-space and the final image-space on the screen. *Coherence* is commonly used in computer graphics to enhance the efficiency of many rendering algorithms [46]. This coherence concept assumes a relative similarity between one component of a rendering and another. In the case of this application, a contour contained within the texture map, when transformed to the surface of a polygon, is likely to be similar in appearance. The contours, although scaled, translated and sheared, are still preserved on the final projected polygon surface.

This texture coherence is used to regulate the sampling of textures and to ascertain the presence of contours for further analysis. This brings gains in the efficiency of texture sampling scenes, by removing the need for unwanted samples. These techniques also improve the quality of images early in the progressive rendering process by discovering and highlighting contours present in texture information, without having to render the subdivision being analysed.

Section 6.1 details how image data is used in 3D image synthesis and provides a theoretical background to the work in this chapter. Section 6.2 describes new techniques developed to adaptively sample textures based upon the visual importance of the projected region. Section 6.3 outlines a new improvement to DCM bump map processing. The chapter concludes with a discussion of achievements in Section 6.4.

6.1 THE USE OF IMAGE INFORMATION IN IMAGE SYNTHESIS

Under the broad heading of *texture mapping*, methods have been developed that enable the high-detail modelling of objects within a scene, without having to extend the geometry of the underlying object representation [24]. Often the application of an image over a simple underlying geometric model can give the appearance of high-detail, without the heavy overhead of processing extra geometry. The source of this texture image may be a captured digital image, or it may be generated by a mathematical function as a *procedural texture* [40]. An example of a captured digital image used as a texture is shown in Figure 6.1. The simple square polygon has an image applied to it to give the appearance of a brick wall.

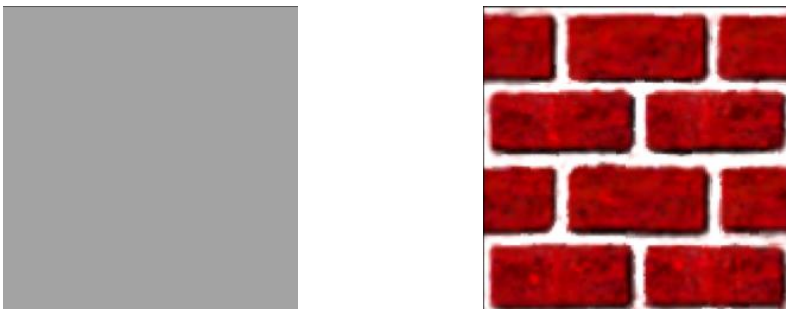


Figure 6.1 A texture mapping illustration. The image on the left is a blank polygon, while the image on the right is the same polygon with a texture map applied.

Texturing is achieved by mapping the image-space pixels, as they are being shaded, to a *texel* (TEXture pixEL). This is shown schematically in Figure 6.2, where the pixel is being shaded with a value sampled from the mapped texel in the image.

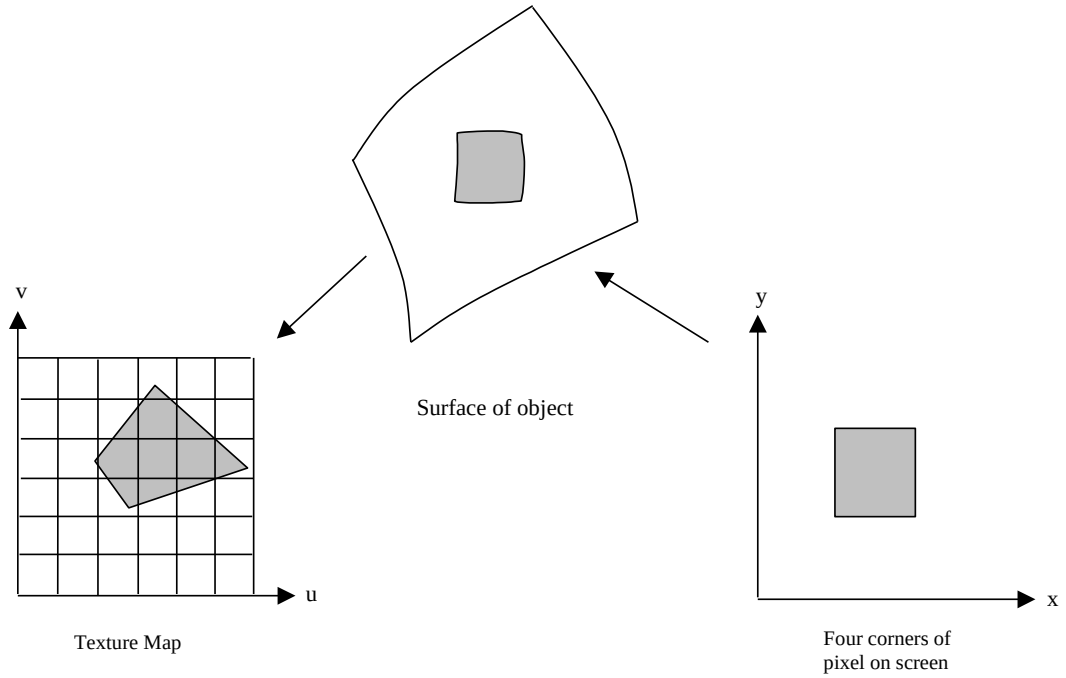


Figure 6.2 Illustration of the process of mapping a pixel on the surface of geometry being rendered to a texel [46].

A further development, called *bump mapping*, uses an image as an achromatic height field, to modify the surface of the geometry being rendered [13]. Bump mapping is performed by perturbing the surface normal of the pixel being rendered by the partial derivative obtained from the image at that point in texture-space. The resultant perturbed surface normal modulates the shading properties of the polygon when any lighting calculations are performed. Therefore, the intensity values in the bump map are converted into pseudo geometry. This facilitates the modelling of complex surface properties, such as gouges, scratches etc., without having to create a complex underlying geometric model. Instead, the intensity gradients in the image are used to implicitly model the desired geometry, and can be considered to be related to the ability of the HVS to infer 3D geometry directly from shading gradients [177]. An example of bump mapping is illustrated in Figure 6.3.

Figure 6.3 An example of bump mapping. A plain polygon is on the left, while a bump mapped polygon is on the right.

In a similar process to texture mapping, the intercept on the surface of the object is mapped to u, v coordinates in the texture map. The calculation of the partial derivatives of the surface at that point in the bump map is executed by sampling nearby texels in the bump map. For any point *u*, *v* within the bump map, the partial derivatives B_u and B_v are approximated by the following equation:

$$B_u = B(u + 1, v) - B(u, v) \quad (6.1)$$

$$B_v = B(u, v + 1) - B(u, v) \quad (6.2)$$

where:

u is the u coordinate location within the bump map;
v is the v coordinate location within the bump map;
B_u is the partial derivative of the bump map surface in the direction of u;
B_v is the partial derivative of the bump map surface in the direction of v;
B is the bump map image.

Adding the cross product of the two partial derivatives to the surface normal derives a simple planar polygon bump map method. This method is less complex than normally used [12], due to the planar nature of the polygons used in this implementation removing the need for partial derivatives to be calculated from the polygon surface. Perturbing the surface normal *N* requires the addition of the cross product of the partial derivatives. This is shown in the following equation:

$$N' = N + (B_u \times B_v) \quad (6.3)$$

where:

N' is the resultant perturbed surface normal;
N is the surface normal of the polygon;
B_u is the partial derivative of the bump map surface in the direction of u (represented as a vector);
B_v is the partial derivative of the bump map surface in the direction of v (represented as a vector).

A problem with texture mapping is the discrete sampling of the texture-space by another discrete sampling space, the image plane. This is indicated in Figure 6.2, with the projected pixel not fitting pixel boundaries exactly. This means that high frequencies in the texture map are not represented correctly upon being sampled for image-space rendering, causing unsightly aliasing. Much work has been carried out into the effective sampling of texture maps to reduce texture aliasing [42, 52, 55, 61, 84, 119, 173]. A common method is the use of a filter with a support of more than one pixel in size, in order to remove high frequencies from the texture. The filtering method works in a similar manner to the antialiasing methods used in pixel supersampling, whereby further samples are gathered next to the texel being sampled, and are then averaged together to form the final texel sample. Some of these filtering techniques have been made adaptive to the levels of contrast within a projected pixel [42, 52, 55, 61, 84], but have not sought to deal with visual importance as presented in this thesis.

This absence of regard for visual importance is reflected in the test scenes used to test the texturing methods. Most texturing methods are only tested with simple scenes, such as checkerboard patterns with strong perspective transformations [61]. While adequate for testing antialiasing effects, these scenes prevent the assessment of the effect of the texturing method within the context of a more complex scene. In a complex scene a large proportion of the image is ignored by the HVS, with only a

few interesting regions being regarded [184]. This indicates a large amount of redundant texture sampling effort, which could be saved if the high quality sampling was only applied to important regions.

The next section details a novel method of dealing with filter support sizes for texture filters. This extends the work developed in Chapter 3 and Chapter 5 by applying visual importance concepts to the resampling of texture maps. The premise is that the visual importance of a region should also influence the resampling of textures, not just sampling by primary rays.

6.2 TEXTURE IMPORTANCE MAPPING

Adaptive texture sampling methods have been developed which seek to sample the texture heavily in areas of high frequency information [42, 52, 55, 61, 84, 119]. However, in a manner similar to adaptive rendering, this adaptive rate is calculated from texture-space contour information, and not from the image-space saliency of the texture when it is finally projected. An adaptive texture sampling scheme, based upon the visual importance of the image-space region, has been incorporated into the previously described adaptive rendering approach (see Chapter 5). This method has been implemented by modifying the support of a texture filter in the texture resampling stage of the rendering system.

The calculation of the importance of the texture in image-space is accomplished by using the region-based importance maps to model the visual importance of image regions at an early stage. Thus, savings in texture resampling overhead can be gained from the judicious use of region importance values, without high computational overhead.

The second factor to be considered is the projected size of a pixel from image-space to texture-space texels [61]. This changes the conditions in which the visual importance approach may work. The methods used within this thesis rely on the ability to reap efficiency gains from the reduction of sampling within regions considered to be unimportant to the viewer. This becomes a problem for the texture

sampling application as the nature of the visual quality changes, depending on the size of the projected image-space pixel.

The projected-pixel size problem can be divided into three categories:

- m:1–many texture-space texels to one image-space pixel;
- 1:1–one texture-space texel to one image-space texel;
- 1:m–one texture-space texel to many image-space pixels.

For the first category, the sampling problem is similar conceptually to the supersampling of pixels in image-space. To improve the quality of the image, the approach should sample around the intersecting pixel to obtain a more accurate value for the image-space pixel, as one sample (texel) is not an accurate integral over the projected pixel area. For the other cases (1:1 and 1:m), further sampling is used to low pass filter (blur) the texture to improve its appearance by softening any edges in the texture [61]. For this thesis, as a proof of concept, the method will simulate the two sets of circumstances using scaled textures and modified sampling regimes. The goal of this work is to evaluate the application of a visual importance regime to modulation of this resampling.

Due to time constraints, adaptive texture mapping techniques have not been implemented. However, this filter support approach is expected to generalise to other adaptive methods due to the global principle of minimising the cost involved in sampling the said textures, in areas that are not visually important. This sampling principle could be used to choose between isotropic and anisotropic filters used in MIP-mapping [173] (refer to Figure 6.4). For example, *anisotropic*[12] filters could be used in visually important areas [42], while a cheaper *isotropic* box filtering could be used within visually unimportant regions that can tolerate more error in their integral calculation [173].

[12] An isotropic filter has directionally symmetric properties, while an anisotropic filter has different properties in different directions.

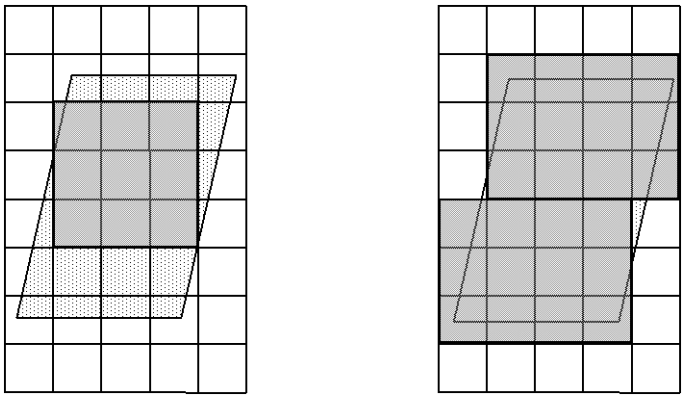


Figure 6.4 Illustration of the difference between isotropic filtering (left) and anisotropic filtering (right) in texture-space (grid). Both texture filters are represented by the grey areas in the diagram. The anisotropic filter better captures the shape of the projected pixel in texture coordinates (dotted quadrilateral), and thus produces more correct texturing in perspective distorted sections of an image. However, the adaptive nature of the filter introduces costs into the texture integral calculations.

The box filtering method in this chapter can be formulated as a linear equation relating the filter size S_{filt} to the texel to pixel size ratio t / p, and the degree of importance of the region in which the pixel is contained I_{reg}. The relationship is:

$$S_{filt} = \begin{cases} MaxSupp \times I_{reg} & \alpha \times r \geq MaxSupp \\ MaxSupp & r < MaxSupp \end{cases} \tag{6.4}$$

where,

S_{filt} is the support size of the texture filter in texels;

α is a user parameter controlling the trade off between efficiency and quality–set to 1.0 for this implementation;

r is the texel to pixel size ratio t / p;

MaxSupp is the maximum filter support size constant–set to 3 for this implementation;

I_{reg} is the visual importance of the image-space region containing the texture value.

The maximum size of the filter is set to allow the filter support to vary from one to three. In essence, the support of the filter can be arbitrary in size [175]. For the sake

of this implementation, a 3×3 filter maximum support is adequate as a proof of concept. The S_{filt} value is rounded to the nearest integer, thereby making the filter support one, two or three texture pixels in width. The filter chosen is determined by the implementation constraints. For the purposes of this project, a simple box filter was used [175]. The disadvantage of this form of texture refiltering is the blurring that occurs due to the removal of high frequencies from the signal. However, the approach could be applied to any filter with finite support, or with any of the more advanced methods involving more sophisticated convolution calculations [61].

This relationship therefore brings about less sampling in areas that are not important to the HVS. However, this method does not work for pixels that have a texel to pixel ratio of less than one, as the concept of removing samples from the less important regions tends to produce better visual quality in low importance areas. This is due to the previously mentioned blurring effect, which is enacted to antialias the texture [61]. Blurring is not related positively to image quality, as the HVS is attracted to high-frequency components of an image [144]. Therefore, the method is suboptimal under these conditions, due to the perceptual degradation of image quality in the high importance area (refer to Figure 6.5). This explains the need to only apply the method to texels with a texel to pixel size ratio greater than or equal to the maximum support size of the texture filter.

Figure 6.5 Example of image which has a texel to pixel size ratio less than the maximum support size of the filter being importance-biased in its sampling (left), compared to a point sampled texture (right). Note that the regions of high importance around the small altar (highlighted with a white rectangle) appear worse due to excessive blurring caused by a larger support for the texture filter function.

Analysis shows that the complexity costs of this technique are: one compare, 2 multiplies and a round per pixel. However, the multiply involving the importance value I_{reg} and α can be precomputed for an entire region, thereby removing the need to calculate the filter support size for every pixel. In addition, the texture to pixel ratio is available as a part of most texture mapping schemes [61], thereby removing the divide from calculations. These minimal costs are weighed against savings of up to eight texture memory accesses for each ray that is cast. This therefore makes the extra computations to be O(n) with regards to the number of regions in the image. In this case, the savings in texture accesses more than make up for the overhead of the importance calculations.

For example, consider the case of a region with only one 8×8 pixel subdivision, containing 64 pixels to supersample at 16 times per pixel. Each ray will at worst require 9 texture samples per ray (filter support 3) and at best 1 texture sample per ray (filter support 1). Texture samples per subdivision range from the worst case of 9,216 texture samples (64 pixels $\times$ 16 supersamples $\times$ 9 texture samples), to 1,024

texture samples (64 pixels × 16 supersamples × 1 texture sample), a maximum difference of 8,192 texture samples. This difference in memory accesses is further multiplied by 3 for the 3 bytes representing the RGB values of the texel.

This theoretical evidence is further born out by the savings in relative time taken to render the scenes. Empirical results shown in Section 6.2.1 indicate an approximate decrease of 10-20% in the time taken to render a textured image, even with simple scenes containing low geometric complexity. The following section reports these test results in detail.

6.2.1 Texture Importance Mapping Evaluation

The following approach was taken to evaluate the texture importance mapping approach. The images were rendered at a fixed ray-traced supersampling rate of one subdivision for each pixel. This is to allow the texture mapping sampling to have the most influence on the quality of the image. An image is then rendered with or without importance-biased texture resampling. A comparison is then made of the time taken to render each image and relative L1 / L2 error norms are generated to give a measure of the difference between the two images. A difference image is also generated to give a visual reference of where the major changes in the image have occurred. Four images were chosen as test scenes.

The first three range from a cloth texture that is highly structured; to a kitchen image that has a less regular structure; to a final garden scene that contains large regions of noisy foliage (refer to Figure 6.6). Each image was constructed by rendering a 513 × 513 scene containing one orthogonally projected textured square. This construction allowed fine control over the texture samples made in each region, thereby facilitating comparison of texturing parameters and visual comparison of image quality. These images were chosen to cover a broad range of structure and frequency content, to ascertain the capabilities of the importance-based texture sampling approach.

A fourth scene of a room with a desk was constructed as a test for more complex texturing scenarios. This provided information about the utility of the texturing

technique within a more realistic application containing perspective projections of the textured polygon surfaces (refer to Figure 6.7).

Figure 6.6 The three textures used in the tests, from left to right: cloth, kitchen and garden.

Figure 6.7 Illustration of a more complex texture test scene.

The next four sections show the results for each scene. The scenes have been rendered with and without texture importance sampling applied at a 513 × 513 image-space resolution. In the first three test images the size of the texture is varied to have the values 257 × 257, 513 × 513, 1025 × 1025, 1537 × 1537 and 2049 × 2049 pixels. The texture sizes map to having a texel to pixel size ratio of: 1:2, 1:1, 2:1, 3:1 and 4:1 respectively. These sizes allow the comparison of results for textures which range over the values of the threshold previously mentioned at the maximum size of the filter (4:1), to the case of having more then one image-space pixel per texture-space texel (1:2). This series then indicates how the texturing method performs over the differing texture scales. It is expected that the method will work best with a texel to pixel ratio of 3:1 or greater, due to each pixel having exactly the same size as the

maximum size of the filter in texture-space (refer to the discussion in Section 6.4). The final room test scene was rendered at a 513×513 pixel resolution using a flat, high quality sampling rate of 4 subdivisions per pixel.

An error image has also been presented to give a visual indication of the locations of the differences between the images rendered with and without a visual importance bias, as was done in Chapter 5. The error values in the image have been negated and thresholded, in order to aid visualisation and reproduction. That is, the dark pixels indicate locations where image differences occur, but they do not illustrate the magnitude of the differences. Rectangular portions of the images have also been cropped and magnified to highlight examples of differences between the images. A table of results for each image is presented containing L1 / L2 error ratios, the number of texture samples made and the relative time taken to render each image.

Cloth Texture Results

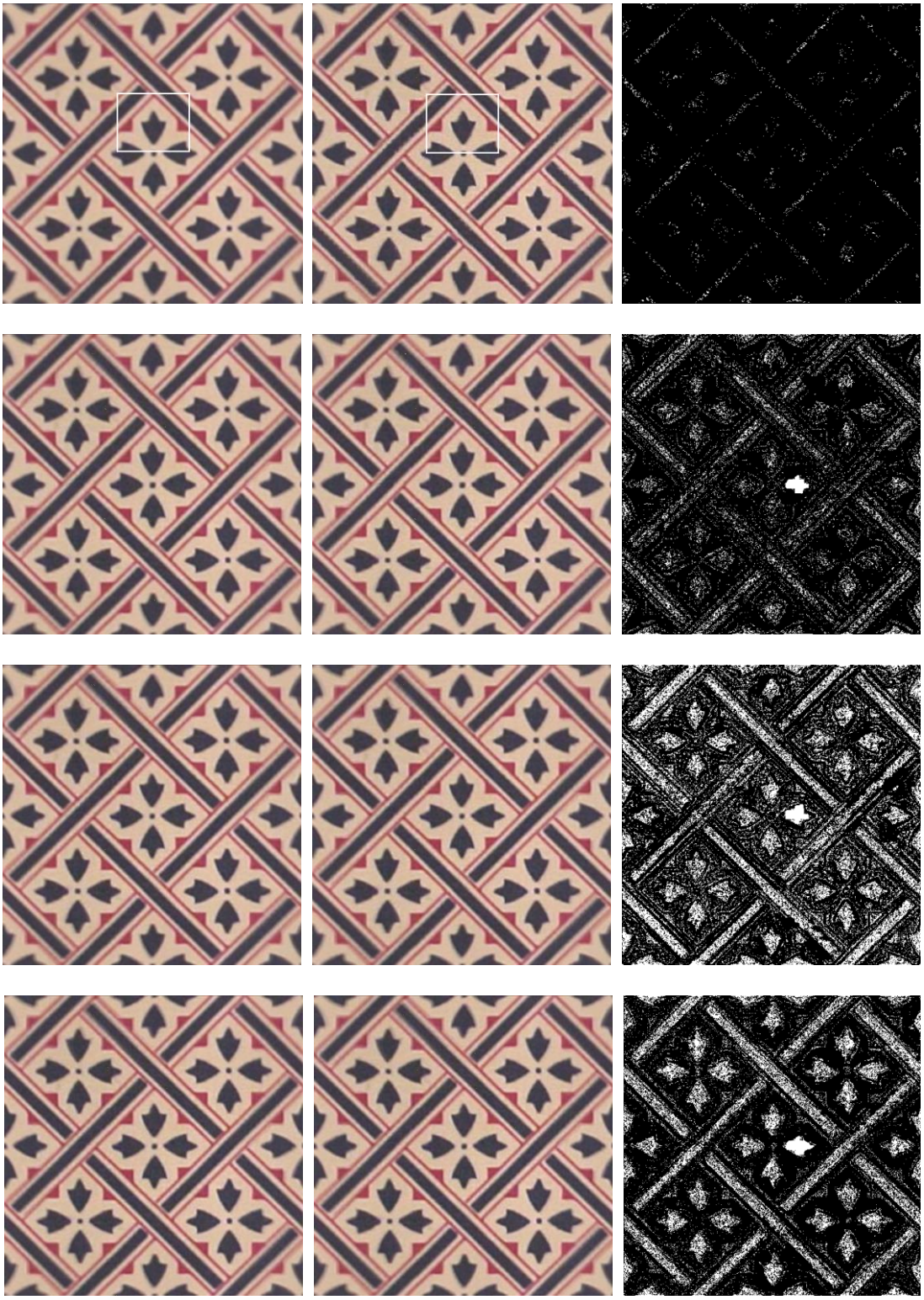


Figure 6.8 Example cloth images which have been produced using the adaptive texture mapping method (left) and without the adaptive texture method (middle). The difference between the two images is shown on the right. The rows represent, from top to bottom, textures resolutions of 257×257, 513×513, 1537×1537 and 2049×2049 pixels. The white regions within the difference images on the right represent pixels that have no difference between the biased and unbiased images. Thus the relatively important regions are shown as white blotches because of the minimal difference between the images in that location.

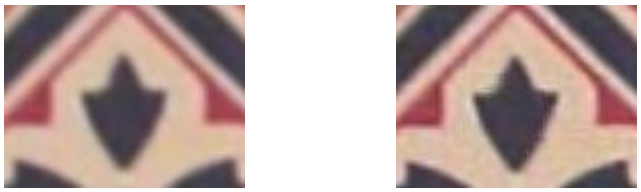


Figure 6.9 Illustration of the level of difference between subimages which contain differences induced by importance-biased sampling. The images are drawn from the white rectangles in Figure 6.8. The base image is on the left while the importance-biased image is on the right.

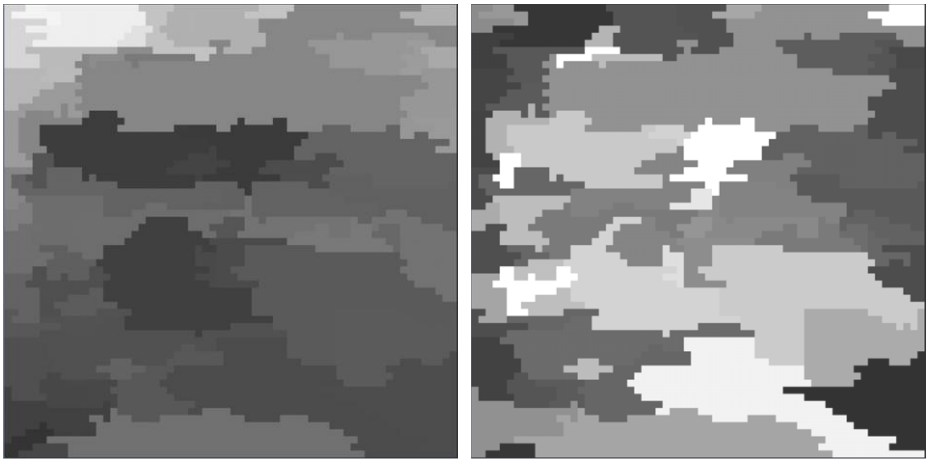


Figure 6.10 Region segementation (left) and importance (right) images for the room texture sampling scene.

Cloth Image	Texture Samples	L1 Ratio (L2 Ratio)	Relative Time Difference With Respect to Base Image
Base 257x257	7,721,584	-	-
Biased 257x257	1,652,171	0.0751 (0.0648)	0.8
Base 513x513	7,721,024	-	-
Biased 513x513	1,592,220	0.0278 (0.0194)	0.8
Base 1025x1025	7,713,832	-	-
Biased 1025x1025	1,602,168	0.0125 (0.0095)	0.9
Base 1537x1537	7,728,976	-	-
Biased 1537x1537	1,644,463	0.0114 (0.0096)	0.9
Base 2049x2049	7,728,912	-	-
Biased 2049x2049	1,528,971	0.0093 (0.0060)	0.9

Table 6.1 Table of results for the cloth texture image.

Kitchen Texture Results

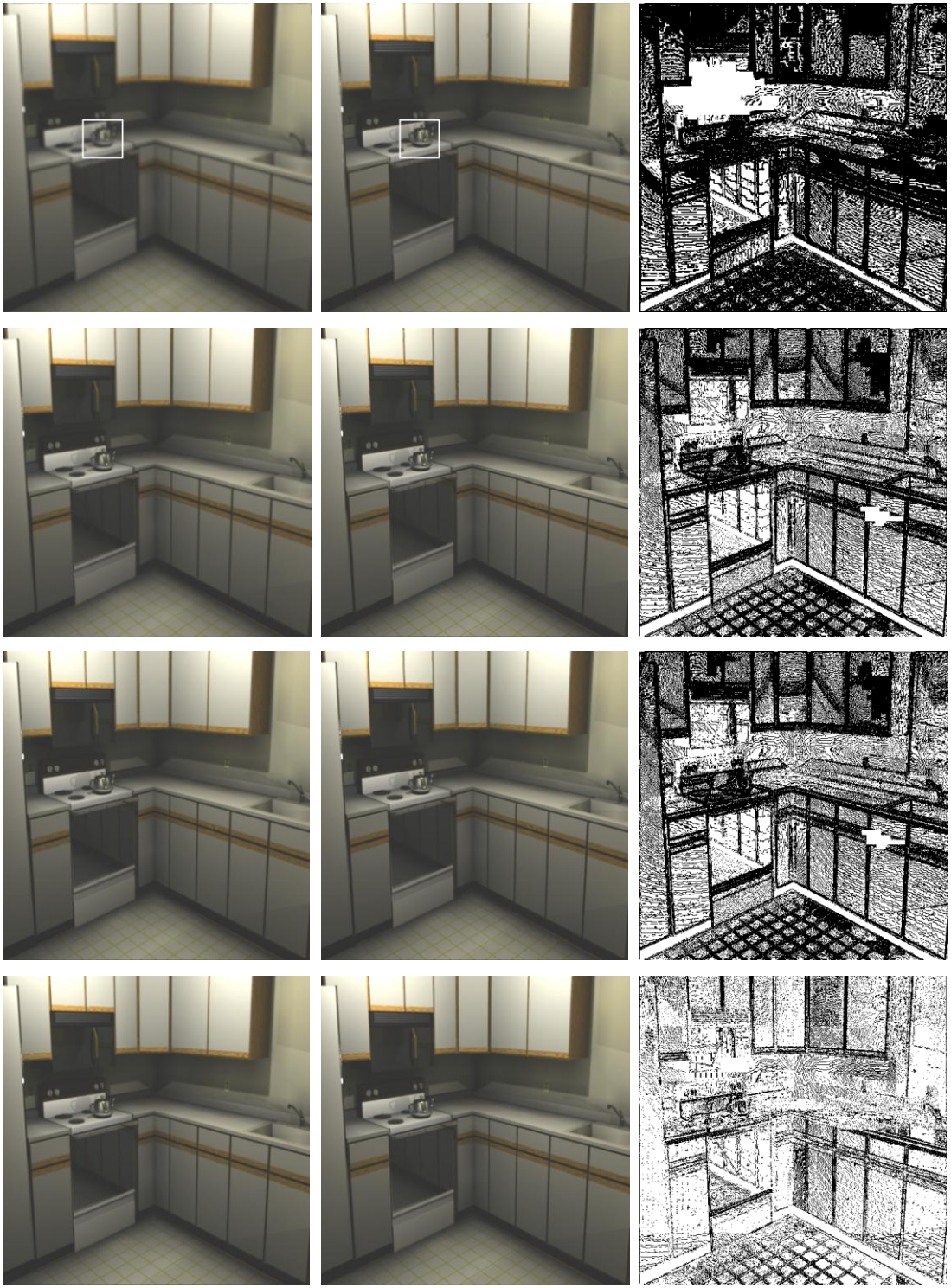


Figure 6.11 Example kitchen images which have been produced without (left) and with (middle) importance-biased texture mapping. The difference between the two images is shown on the far right. The rows represent, from top to bottom, textures resolutions of 257 × 257, 513 × 513, 1025 × 1025, 1537 × 1537 and 2049 × 2049 pixels. The white regions within the difference images on the right represent pixels that have no difference between the biased and unbiased images. Thus the relatively important regions are shown as white blotches because of the minimal difference between the images in that location.

Figure 6.12 Illustration of the level of difference in a subimage which contains differences induced by importance-biased sampled. The images are drawn from the white rectangles in Figure 6.11. The base image is on the left while the importance sampled image is on the right.

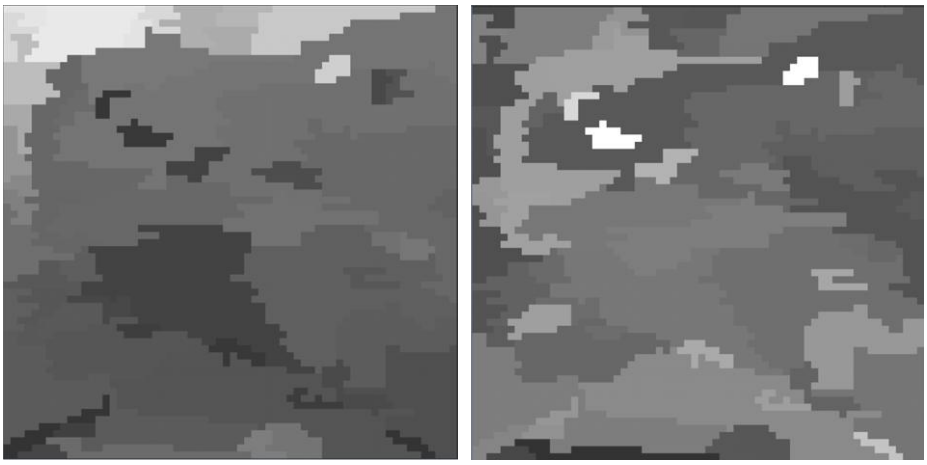


Figure 6.13 Region segementation (left) and importance (right) images for the kitchen texture sampling scene.

Kitchen Image	Texture Samples	L1 Ratio (L2 Ratio)	Relative Time Difference With Respect to Base Image
Base 257x257	8,105,712	-	-
Biased 257x257	2,031,767	0.0712 (0.0391)	0.8
Base 513x513	8,119,128	-	-
Biased 513x513	1,836,411	0.0448 (0.0208)	1.0
Base 1025x1025	8,125,200	-	-
Biased 1025x1025	2,400,637	0.0239 (0.0102)	0.7
Base 1539x1539	8,113,352	-	-
Biased 1539x1539	1,103,255	0.0198 (0.0055)	1.0
Base 2049x2049	8,119,056	-	-
Biased 2049x2049	1,834,178	0.0206 (0.0056)	0.8

Table 6.2 Table of results for the kitchen texture image.

Garden Texture Results

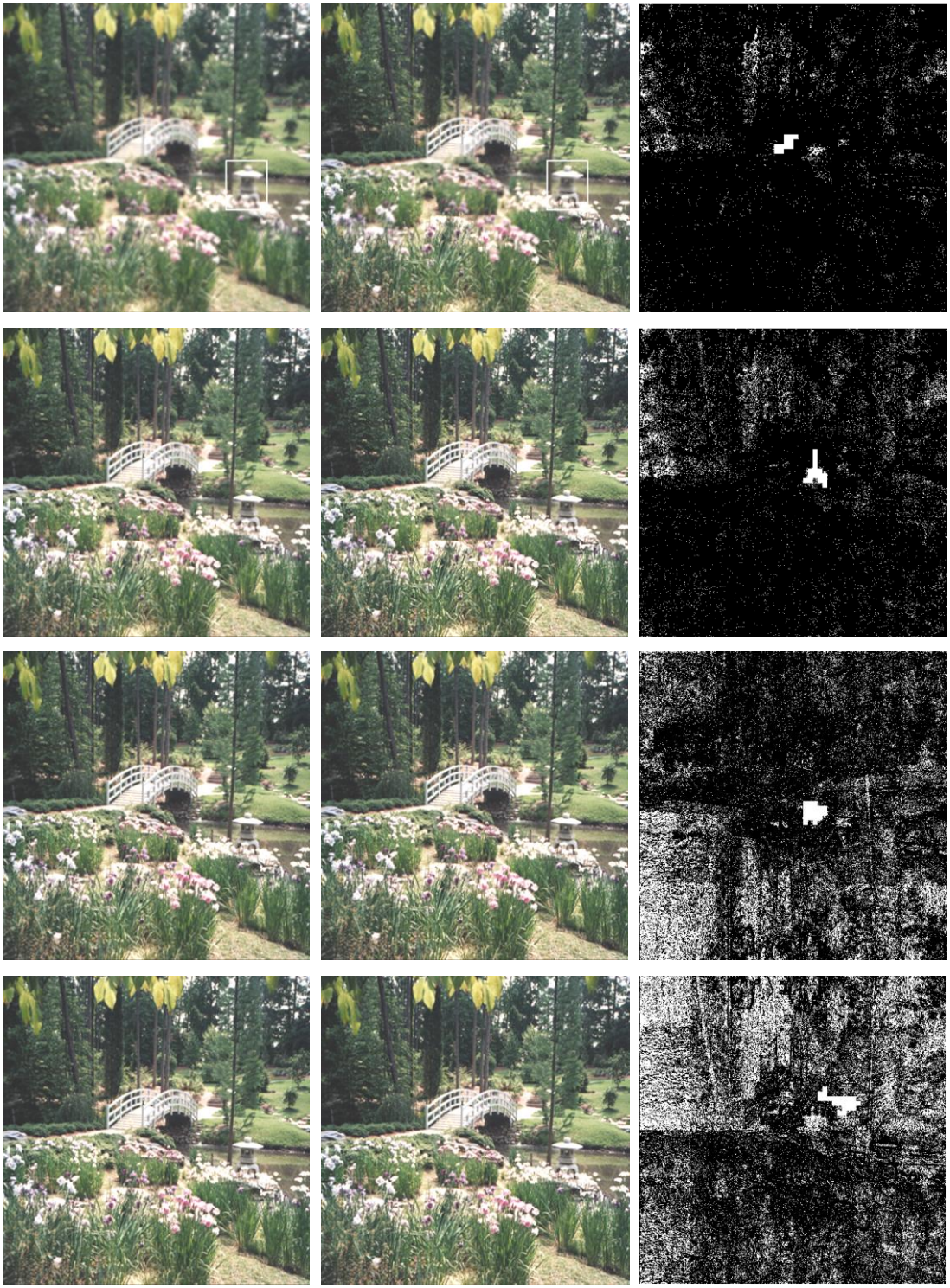


Figure 6.14 Example garden images which have been produced using the adaptive texture mapping method (left) and without the adaptive texture method (middle). The difference between the two images is shown on the right. The rows represent, from top to bottom, textures resolutions of 257 × 257, 513 × 513, 1025 × 1025, 1537 × 1537 and 2049 × 2049 pixels. The white regions within the difference images on the right represent pixels that have no difference between the biased and unbiased images. Thus the relatively important regions are shown as white blotches because of the minimal difference between the images in that location.

Figure 6.15 Illustration of the level of difference in a subimage which contains differences induced by importance-biased sampling. The images are drawn from the white rectangles in Figure 6.14. The base image is on the left while the importance sampled image is on the right.

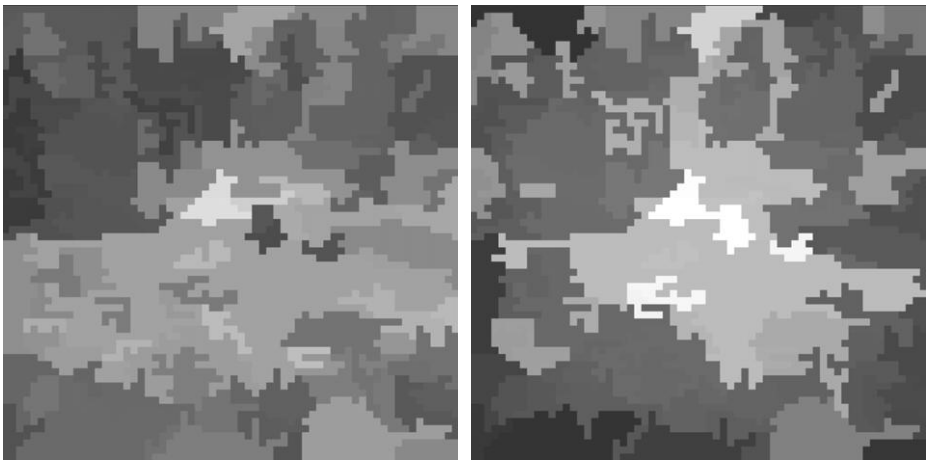


Figure 6.16 Region segementation (left) and importance (right) images for the garden texture sampling scene.

Garden Image	Texture Samples	L1 Ratio (L2 Ratio)	Relative Time Difference With Respect to Base Image
Base 257x257	7,629,824	-	-
Biased 257x257	1,786,451	0.1445 (0.1046)	0.9
Base 513x513	7,681,752	-	-
Biased 513x513	1,729,007	0.0975 (0.0794)	0.9
Base 1025x1025	7,661,960	-	-
Biased 1025x1025	1,771,196	0.0452 (0.0353)	0.9
Base 1537x1537	7,666,496	-	-
Biased 1537x1537	1,580,673	0.0457 (0.0226)	0.8
Base 2049x2049	7,668,744	-	-
Biased 2049x2049	1,570,569	0.0445 (0.0179)	0.9

Table 6.3 Table of results for the garden texture image.

Room Results

Figure 6.17 Results of room scene rendering with the base image (left), importance-biased image (middle) and a difference image (right). The white regions within the difference images on the right represent pixels that have no difference between the biased and unbiased images. Thus the relatively important regions are shown as white blotches because of the minimal difference between the images in that location.

Figure 6.18 Illustration of the level of difference in a subimage which contains differences induced by importance-biased sampling. The images are drawn from the white rectangles in Figure 6.17. The base image is on the left while the importance sampled image is on the right.

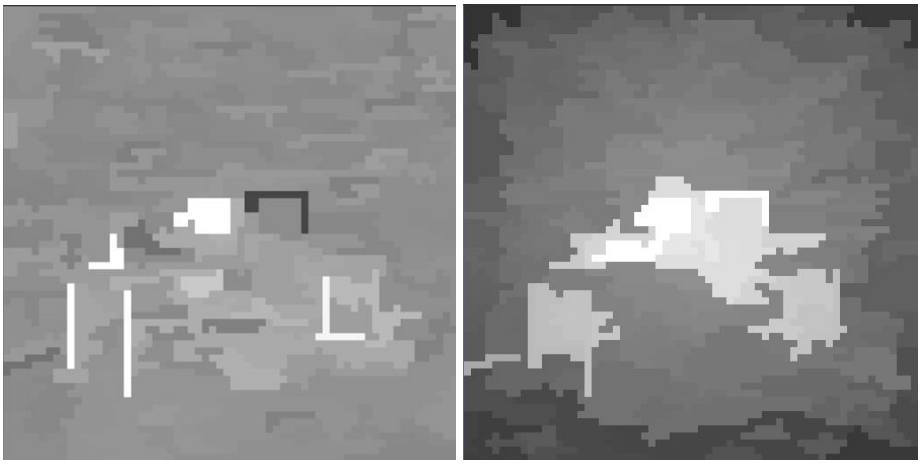


Figure 6.19 Region segementation (left) and importance (right) images for the room scene.

Room Image	Texture Samples	L1 Ratio (L2 Ratio)	Relative Time Difference With Respect to Base Image
Base	55,506,299	-	-
Biased	7,681,298	0.0404 (0.0429)	0.9

Figure 6.20 Table of results for the room scene.

Discussion

Even with a modest texture mapping method, such as used in these experiments, the importance-based texture sampling approach is still able to render the image 10-20% quicker than for a base image with a constant texture sampling rate. In addition, the images which were rendered with a texel to pixel size ratio of 3:1 and greater had lower distortion levels, as was expected. The distortion levels as indicated by the L1 / L2 norms were effectively halved once the 3:1 threshold had been approached–that is, the 1537 × 1537 size textures. This restricts the use of this form of box filter-based importance sampling to high texel to pixel size ratios. It should also be noted that the subjective quality of images generated just short of the threshold (1025 × 1025 images) was still high, indicating potential room to modify the α value in the sampling expression, to allow user control of the final image quality.

Furthermore, the room scene exhibited similar timesavings, with a small loss of image quality. This adds support to the utility of this technique in increasing the efficiency of image synthesis with non-trivial textured scenes.

It should also be noted that this adaptive texturing method is constrained by the quality of the segmentation of the image. The merge segmentation approach struggled to match the structure of some of the scenes used in this chapter. The segmentation of the room scene was, however, of a better quality. Some possible improvements to this merge algorithm are discussed in more detail in Section 9.2.

6.3 TEXTURE ADAPTIVE MESHING

The Discontinuity Coherence Map (DCM) is the contour analysis approach used in Chapter 4. Contrast and geometry measures are used by the DCM to ascertain the presence of a contour within a subdivision in the scene, effectively giving a piece-wise linear approximation of the projected luminance function for the image [57]. These measures are used in the first step of the sampling process (refer to Chapter 5), when only four samples per subdivision have been made. The DCM approach uses information from both object-space and image-space to capture contours early in the refinement process: visible lines using a hardware rendering system, polygon tags indicating the presence of different objects in a subdivision and contrast at the four corners of the subdivision [57]. Other systems of progressive rendering use a texture mapped polygon to save on the need to sample the texture map at every location. Instead the texture-mapped polygons are pre-rendered and then blended with the radiosity samples made in the scene [130]. The DCM, however, does not use a merging of polygonal and radiosity values, and does not capture all the possible contours in a scene. There is the possibility of texture information modifying the luminance function in a scene, as shown in Figure 6.21, and thus introducing contours indiscernible to the other methods until later subdivisions are performed.

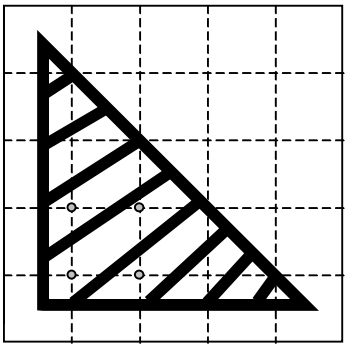


Figure 6.21 Illustration of bump map checking algorithm. The four sampled points (grey circles) would normally return no contrast difference, even though the bump/texture map has contour information (thick lines). In the new technique, the sampled points have their texture coordinates checked between them for large luminance deviations, indicating a plausible contour in image-space.

The ray-tracing implementation used in this thesis receives geometry information from the ray intercept in the form of an ID tag. The ID tag is used to index shader information for the intercepted polygon, thus indicating the presence of a bump map within the subdivision. The two points sampled at the edge of the subdivision form a line in the bump map space. If there has been no contour detected by normal means, and if the samples at each point contain bump map information, then the texture-space is scanned for contours along the subdivision edge. If the subdivision contains an edge of a large enough magnitude to possibly form a contour, then the system classifies this subdivision as non-smooth and performs further sampling and subdivision accordingly, as per the other methods of contour detection.

Furthermore, if there is only one sample that contains bump map information, or the two points access different bump maps, then the system automatically flags the subdivision for further analysis. This is similar to the cautious marking of subdivisions with different polygon hit tags as being likely to contain contours [57].

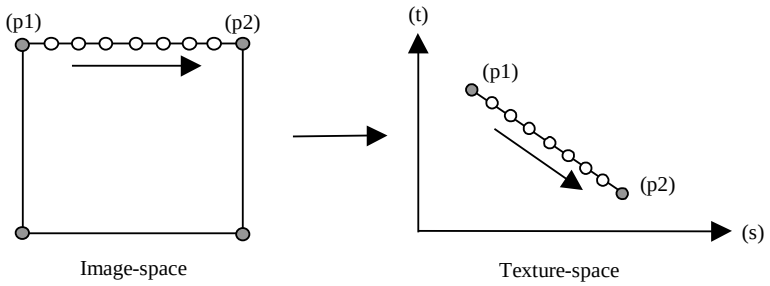


Figure 6.22 Illustration of process involved in ascertaining the locations of possible contours within a bump map. The grey corners are ray traced pixels, with the other white circles the pixel locations yet to be sampled. If a deviation is found then the subdivision is marked for further sampling according to the original DCM algorithm.

The *Bresenham* line rasterisation algorithm is used to form the sample line through the bump map texture-space (refer to Figure 6.22) [18]. The Bresenham algorithm is used due to its efficiency in utilising integer-based arithmetic to generate the texel locations. The aim here is not to ascertain the strength of the contour, this is better left to the luminance evaluation functions in the ray-tracer. The goal is to find likely image-space contours, and to flag them for later processing by the sample generator. This gives an added metric to the contour analysis system, with minimal computational effort.

A polygon with a brick bump map is shown below in Figure 6.24, with and without the bump map sampling metric. The bump map sampling metric finds the contours in the edges of a subdivision and thus alerts the subdivision algorithm to possible luminance changes that should be processed by a subdivision of the quadtree. The image is thus improved at any stage of the subdivision by being alerted to the presence of possible contours.

This method can be applied to the use of coloured textures as well. The most efficient way to handle textures is to use hardware rendering to pre-render the polygons and then blend the textured geometry with the luminance information in the frame buffer. However, in the absence of hardware rendering capabilities, one can use the texture-space search to discover contours that may appear from texture mapping in a similar manner. The method is slightly different to bump mapping, as a bump map is inherently an achromatic channel. Here, the technique has to convert

the values in the texture map to grey levels to enable the searching of the texture for contour information.

The technique preprocesses the texture map to create a simple edge detected version of the texture map, in grey scale, which allows the DCM to sample the texture map along the axis indicated. Any changes in the edge map will signify a potential edge within this side of the subdivision and, again, the DCM is alerted to the presence of a contour within the subdivision. Further processing as per the DCM algorithm is carried out for the subdivided contour subdivision.

The method is general enough that other progressive approaches will benefit from this technique, as it only requires that a line exist between two samples in the same texture-space, in some form of subdivision scheme. This means that progressive rendering methods using regular adaptive sampling [129] or Delaunay triangulation methods [125] could also benefit from this approach.

6.3.1 Texture Adaptive Meshing Evaluation

Using the same error measure as used in the objective evaluations in Chapter 5, the images were progressively rendered with L1 and L2 error norm values generated for each image. Figure 6.21 illustrates an optimal scenario, where the contours within the subdivision miss sample points, but will be detected by the bump map checking technique. Bump maps having a regular structure are expected to benefit most from this technique, as their contours fall within the subdivision edge without causing a contrast at the sample points.

Scaling of the texture is therefore an issue important to the success of this technique. The brick bump map texture used was generated at texture coordinates that doubled in size for every frame. This in effect reduced the size of the bump map texture by half each time, giving an indication of the effectiveness of the algorithm at different texture scales. The results have been plotted as L1 and L2 ratio graphs, with selected images displayed to visualise the effect on image quality.

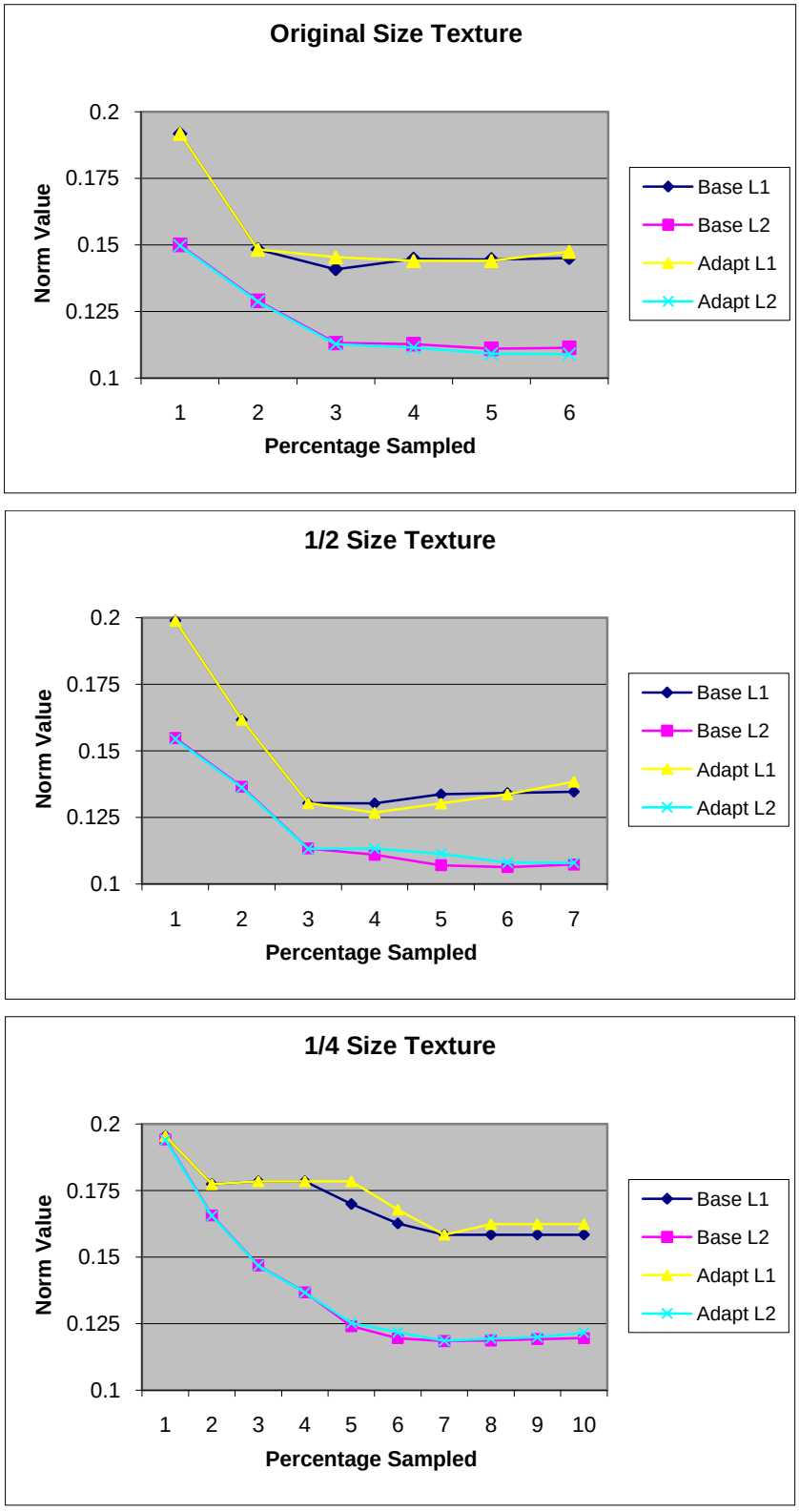


Figure 6.23 Graphs of L1 and L2 norms for progressive rendering of textures scaled to 1, ½ and ¼ their original size.

Original Size Texture Percentage Sampled	Base L1	Base L2	Adapt L1	Adapt L2
1	0.1918	0.1500	0.1918	0.1497
2	0.1483	0.1290	0.1483	0.1287
3	0.1409	0.1131	0.1454	0.1127
4	0.1447	0.1127	0.1440	0.1114
5	0.1445	0.1110	0.1440	0.1091
6	0.1451	0.1113	0.1475	0.1089

½ Size Texture Percentage Sampled	Base L1	Base L2	Adapt L1	Adapt L2
1	0.1990	0.1548	0.1990	0.1544
2	0.1617	0.1365	0.1617	0.1362
3	0.1303	0.1133	0.1303	0.1132
4	0.1302	0.1110	0.1268	0.1133
5	0.1337	0.1070	0.1302	0.1114
6	0.1341	0.1063	0.1337	0.1080
7	0.1346	0.1073	0.1385	0.1079

¼ Size Texture Percentage Sampled	Base L1	Base L2	Adapt L1	Adapt L2
1	0.1955	0.1942	0.1955	0.1940
2	0.1774	0.1657	0.1774	0.1655
3	0.1785	0.1469	0.1785	0.1468
4	0.1785	0.1368	0.1785	0.1367
5	0.1699	0.1240	0.1785	0.1252
6	0.1626	0.1195	0.1678	0.1218
7	0.1584	0.1184	0.1584	0.1185
8	0.1584	0.1187	0.1624	0.1194
9	0.1584	0.1192	0.1624	0.1199
10	0.1584	0.1196	0.1624	0.1216

Table 6.4 Table of values for the renderings of the brick bump map with original, ½ and ¼ size textures.

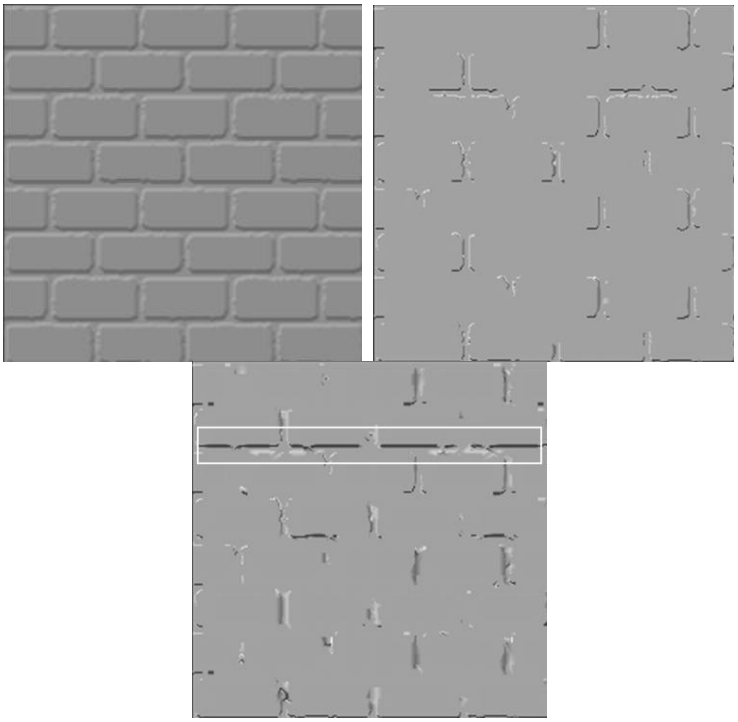


Figure 6.24 Illustration of the ability of the technique to discover contours not found by the base DCM method. The image on the left is the final rendering, the middle image is the base image 7% sampled, the right image is the texture adaptive method at 7% sampled. All images are for the ½ scaled texture example. Examples of extra contours in the far right image are highlighted by the white rectangle.

Discussion

As can be seen from the graph values, the L1 and L2 ratios are not greatly affected by the texture adaptive technique. However, the images shown in Figure 6.24 illustrate the ability of the method to discover contours within the bump map texture applied to the polygon. This ensures that contours which will influence subdivision in the progressive refinement process are discovered earlier than in the base DCM. A positive example has been shown for a regularly structured bump map.

The time complexity cost to the algorithm is the tracing of the edge of a subdivision through texture-space. This is only performed once after the initial subdivision, and only on subdivisions that have not already been flagged as containing a contour. It should also be noted that the worst case scenario of having to search every subdivision for bump map information would be unusual, as only a proportion of the subdivisions will contain bump map information for an arbitrary scene. Therefore, the time complexity is O(n), with the coefficient for the time complexity expression being less than one for most cases. The space complexity is the storing of two (s and

t) texture coordinates as machine words–for each of the four corner points of the elementary subdivisions in the image. This space complexity expression is O(n) with respect to the size of the image in pixels, with a coefficient of 2 / 64–only two samples per 8 × 8 pixel subdivision needed, due to shared subdivision vertex samples.

In addition, the effectiveness of the method is predicated on the progressive method used to render the scene. Other progressive ray-tracing methods are not aware of texture-space information in subdivision decisions [125, 129, 130], and so this general technique should be of benefit to these methodologies as well.

6.4 DISCUSSION

Texture sampling is a computationally intensive task. In this chapter, texture resampling techniques have been developed which account for the visual importance of the textured regions. This has been achieved by modulating the support of filters used to resample textures, by calculated visual importance values. The approach has been shown to work for the box texture filtering method, with a texel to pixel size ratio greater than the maximum size of the filter.

The progressive rendering of bump mapped polygons has also benefited from techniques devised to search for contours within texture-space, to help uncover contours sooner. Progressive rendering of certain bump mapped surfaces have been shown to benefit from this bump map searching approach.

Both approaches have been evaluated with objective methods of assessment, and have been shown to be beneficial in improving either the efficiency or the quality of images in a progressive ray-tracing scenario.

Chapter 7
Adaptive Image Synthesis Animation

Previous chapters have dealt with the development of an overall approach to the application of visual attention to progressive and adaptive ray tracing techniques. This chapter extends these ideas by incorporating temporal changes into the models and techniques developed.

Research indicates that motion is a strong attractor of visual attention [116, 128, 144, 151, 177, 185]. There is also physiological evidence for the pre-eminence of motion in the hierarchy of visual features, due to the presence of receptors sensitive to moving contours [20] and the magnocellular pathway in the visual system [92]. These results are consistent with psychophysical evidence showing that motion very strongly attracts visual attention [177, 183].

Closely related to motion is the abrupt onset of a stimulus [177], generally considered to be due to changes in luminance values across the scene–for example, the turning on of a spotlight. Experiments have shown that in non-attentive modes the sudden onset of stimuli within the periphery brings about attentional capture [183]. In addition to this is the discovery that the onset tends to be more effective at attracting attention when the stimulus is aligned with the appearance of an actual object, and not just a change in the visual features of a region [62]

Research has also shown the high correlation of points of regard between viewers when observing images containing movement [151]. It can be concluded that much rendering effort can be saved by further exploiting the visual attention principles used in Chapter 5 and Chapter 6. Given the attention capturing ability of moving objects [62], it is expected that the best results will be gained from adding motion to the newly developed visual attention model. Furthermore, it can be seen that any complete temporal change model must account for both motion and stimulus onset factors.

In this chapter a more complete model of the effects of motion upon visual attention is developed. In particular, the new model improves on others by incorporating the following factors:

- relative forms of motion are used to ascertain the importance of the region, as opposed to absolute measures used in present models;
- magnitude and directional factors are treated as separate factors contributing to the final motion-based visual importance of a region;
- global effects are also incorporated, to model the enhancing and suppressing influences of surrounding motion in a scene
- parameters for the model are gained from psychophysical research, instead of using arbitrary values;
- differentiation of onset effects from those caused by the motion of objects in the scene–for example, lighting changes.

Furthermore, new animation techniques are developed to exploit the visual importance evaluation offered by the temporal change model. A region-based motion detection approach is developed which has the following features:

- the ability to account for gross and local motion effects of regions explicitly–eg. both translation and internal rotation of objects;
- the ability to account for non-affine transformations of the regions being analysed–ie. non-linear region deformations;
- the ability to remove camera motion effects from the derived region motion vectors.

The chapter details the theoretical basis and the design of the major components of the approach. Implementation issues are also discussed at the end of this chapter. Implementation has been left as future work due to it being a major task, adding to the workload of an already large project. Furthermore, due to the good results from work performed with still images in Chapter 5, it is expected that the application of

motion to both the visual importance model and the rendering techniques should give the same, if not better, results.

Structurally the rest of the chapter is organised as follows. An analysis of present research into the effects of motion upon viewer gaze positions is presented in Section 7.1. Section 7.2 details the development of extensions to the visual attention model developed in Chapter 4. Section 7.3 then details the incorporation of the new temporal change model into an adaptive rendering system. Finally, the chapter concludes with a discussion of achievements in the design of the model in Section 7.4.

7.1 EFFECTS OF MOTION ON EYE MOVEMENTS

The physiology of the HVS contains constructs for the detection of motion. The retina (see Chapter 2) contains receptors which are only sensitive to moving contours within their receptive field [20]. Further into the visual system there is a construct named the magnocellular pathway, which is an achromatic channel sensitive to motion. This is confirmed by psychophysical tests, which indicate the achromatic nature of motion perception [92].

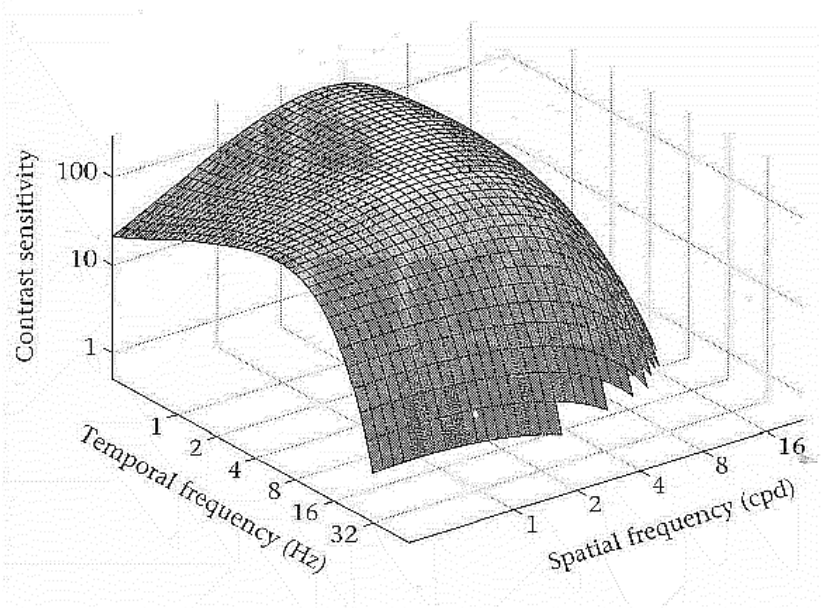


Figure 7.1 Spatiotemporal sensitivity curve for the HVS from Wandell [166]. The vertical axis represents the magnitude of contrast required to detect a contrast reversing signal at the specified spatial frequency (in cycles per degree subtended–cpd) and temporal frequency (in cycles per second–Hz). Note the asymmetry in the curves introduced by the use of logrithimic scales on each axis.

This ability of the HVS to be strongly attracted to motion has it limits. Figure 7.1 shows the spatio-temporal contrast sensitivity curve, indicating in a manner similar to the contrast sensitivity curve, the sensitivity of the HVS to combination of spatial and temporal frequencies.

The above surface indicates the sensitivity of the HVS to changes in spatial frequencies which contrast reverse at the temporal frequencies labelled on one of the axes. The upright axis indicates the inverse of the amount of contrast required before the viewer perceives the grating change. Of interest is the falloff in sensitivity after certain temporal and spatial frequency values are reached. Two commonly quoted limits derived from such psychophysical data, are the perceptual fusion of sinusoidal luminance gratings at above 16 cpd and light flashes at anything less than 15-20ms intervals [166].

Moving on from simple spatial frequencies, an even more relevant issue is the tracking capability of the HVS. This ability to track a moving object is called smooth pursuit (refer to Section 2.1.3), and has been analysed by a number of

researchers [51, 138, 164, 171]. Work performed by Daly [31] has also derived the following graph of the pursuit capabilities of a test subject.

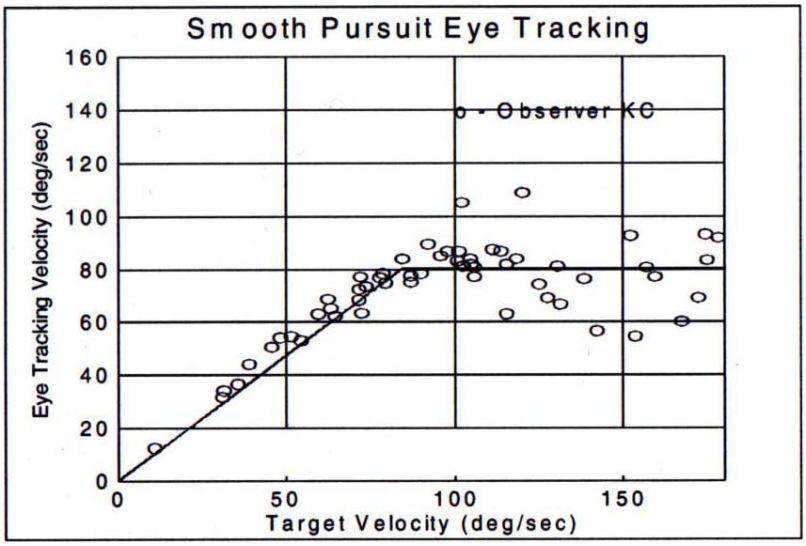


Figure 7.2 Graph of smooth pursuit capability of the human visual system [31].

These two experimental results become important when evaluating the importance of a region within the visual field of a viewer. Firstly, the change in luminance must be visible to the viewer. Secondly, the motion perceived must be able to be tracked, in order to attract attention. Motion magnitudes beyond the tracking capabilities of the HVS will reduce the correlation between the location of the viewed object and the locations of the points of regard.

In addition, results from experiments by Nothdurft also indicate a sigmoidal pop-out effect from local motion differences, with a saturation effect past a high level of local motion difference [122, 123]. As with luminance and colour, the effect is suppressed by surrounding motion differences [122].

Evidence therefore indicates that as well as temporal magnitude change, there needs to be consideration of the vector nature of motion within the visual field, based upon local and global region motion differences. Thus, a region-based model of motion importance has been developed to incorporate temporal changes using both direction

and magnitude of motion. In addition, the model also accommodates abrupt onset effects.

Previously, the majority of the application research work has been carried out into detecting changes in an image, for compression purposes, in order to reduce the amount of data needing to be sent for low bandwidth video applications [87]. Recently, in addition to this raw detection and compensation for change in an image, there has been the application of the previous psychophysical experimental results to the determination of the importance and visibility of changes occurring within a video stream [34, 128]. Models have been developed to simulate the visual importance of motion within the application areas of video processing and image synthesis, in order to further reap efficiency gains not possible through raw change detection.

A multiresolution motion model has been developed by Yee [185, 186], as an addition to the visual saliency model of Koch and Ullman. [83] and Itti and Koch [72]. The model uses a magnitude value to ascertain the visual importance of pixels in the region, using an object ID-based method of pixel displacement calculation developed by Agrawala [1]. These pixel-based velocity measures are then fed into a centre-surround mechanism, which processes local differences in motion. The values are normalised across the image by an operator accounting for the overall activation of the motion feature dimension. The approach did not, however, allow for camera movements in its motion estimation. This motion extension was used, along with other spatial features (refer to Section 3.1), to modify global illumination parameters in computer animation. The rendering system applied more sampling to those regions, which via motion differences attracted the attention of the viewer.

Region-based approaches have also been used to model motion importance within a series of images. Osberger [128] has devised an effective threshold model to incorporate motion into a video processing system. The function is also adaptive to the overall quantity of motion in the image and compensates for camera motion using an undocumented mechanism. While the motion model reports good results for determining motion importance, the parameters are not referenced to any

psychophysical data. An arbitrary upper threshold of 20 deg/sec. is used on the motion importance function. Motion importance in this approach is calculated by the absolute magnitude of the motion. Even though this approach allows for global distribution of motion, it does not allow for local difference effects due to directional or magnitude values.

A simple fuzzy logic motion importance model has been developed by Marichal et al. [102] and De Vleeschouwer et al. [34, 35] for applications to low-bandwidth video. The motion estimation is based upon absolute magnitudes of motion, which do not account for global quantities of motion. In addition, the model does not account for the direction of motion, nor does it allow for local differences in motion effects. A major component of the model is a user-defined region of interest parameter, which effectively removes a large component of its automatic importance calculation capabilities. It must be understood, though, that the model was developed for low-bitrate video applications–where the user does have considerable input into the parameter settings for the application.

The fuzzy membership functions comprising the motion component of the model are arbitrary in nature, being trapezoidal functions arranged over the antecedent universe of discourse. They do not account for any psychophysical results in the relevant literature. Furthermore, no results for the effectiveness of the model are listed, except for preliminary results of subjective video quality. The system is incorporated into a low-bandwidth video encoding and transmission scheme, which facilitates user control of quality parameters.

The issue of appropriate parameter values for the motion importance system is crucial to its potential modelling of visual importance. One important parameter is the upper limit on the smooth pursuit motion of the HVS. Daly [31] reports a value of 80 deg/sec[13], while other researchers report values ranging from 20 to 30 deg/sec by Wesheimer and Robinson [138, 171] to 50 deg/sec by Verstraten et al. [164]. The latter value being considered of dubious application to this thesis application area,

[13] Degrees per second being the number of degrees subtended per second by the motion of the region. This value is therefore dependent upon the distance of the viewer from the scene.

due to the results being gained from wilful direction selection of ambiguously rotating stimuli and not the tracking of real moving objects in the visual field. The values reported by Daley have added weight due to the natural viewing nature of the experimental conditions.

These region-based models only treat motion as a magnitude, and do not include its vector component in their calculations. The models also do not differentiate between motion and abrupt onset in any fashion. Therefore, these approaches can be improved by incorporating a region-based measure of local differences, which accounts for global suppression and enhancement effects. As motion is a vector quantity, this should be considered in the model. If a region is moving in an opposite direction at the same velocity as other objects, then it will still be noticeable due to the local difference in motion direction [122]. The next section will detail a temporal change model that incorporates these improvements.

7.2 A VISUAL ATTENTION MODEL INCORPORATING TEMPORAL CHANGES

Temporal changes involve two major categories, actual motion of objects in a scene, and sudden changes occurring due to luminance changes unrelated to object motion. Both of these have been characterised within this temporal change model, to accommodate most effects occurring within an animated scene.

7.2.1 Motion Membership Functions

In a similar fashion to the membership functions in the spatial visual attention system, the motion membership function is adaptive to the magnitude of motion detected within the visual field. Another membership function models the importance effects of the direction component of the region motion. Both of these factors contribute to the motion importance of regions. An illustration of this concept is shown in Figure 7.3.

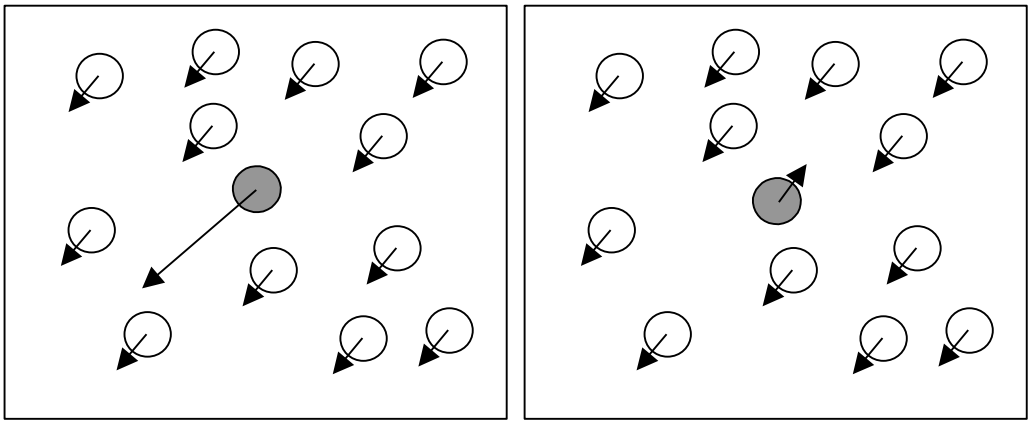


Figure 7.3 Illustration of the concepts of motion magnitude importance and motion direction importance. Both images show regions with vectors attached, indicating their direction and magnitude of motion. The left diagram show a grey region standing out due to a difference in velocity magnitude–indicated by the longer arrow–while proceeding in the same direction. The right diagram shows a grey region standing out because of its relative difference in direction–indicated by the reversed direction vector–while proceeding at the same speed.

The membership functions derived from these also exhibit threshold and saturation effects as uncovered in psychophysical research [122]. Therefore, the membership function follows the design outlined in Figure 7.4.

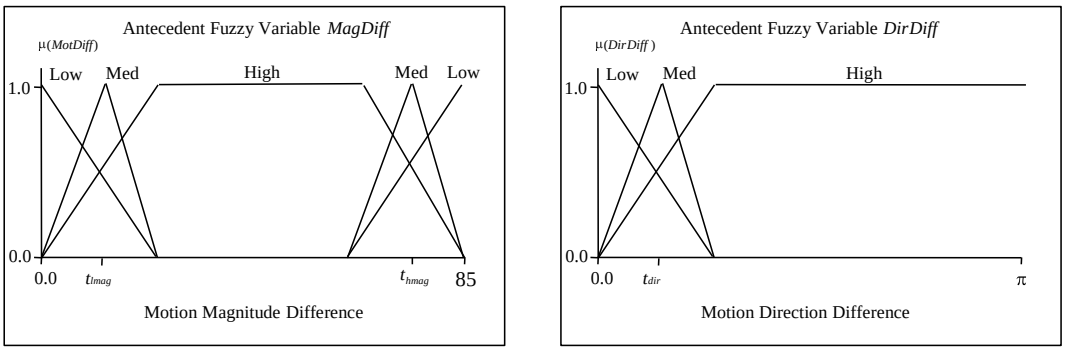


Figure 7.4 Diagram of the motion evaluation membership functions for the Magnitude of the motion (left) and the Direction of the motion (right).

The functions adapt themselves by fuzzifying the threshold of pop-out for the motion feature. The membership function for the magnitude difference variable has two thresholds. The first t_{lmag} is derived from the average number of motion magnitude differences occurring between the regions within the scene. The second threshold t_{hmag} is the practical upper limit for the tracking ability of the human visual system. According to Daly [31] this value is approximately 80 deg/sec. Therefore, the visual

importance of regions above the final threshold falls off to zero quickly-over 5 degrees to be zero at 85 deg/sec.

Research into the tracking capabilities of the HVS reports upper thresholds with respect to the visual angle subtended by the target per sec. [31]. However, it has to be noted that the falloff in tracking ability has not been modelled by the research. Therefore, the ad hoc value of 5 degrees has been chosen to account for two possible threshold factors. Firstly, to model a potentially fast falloff in tracking ability that is most likely sigmoidal in nature [122]. Secondly, the possible threshold differences between subjects fuzzifies the actual values of the threshold–in a similar manner to the other thresholds in the visual importance model (refer to Section 4.2). This final upper threshold is non-adaptive, as it models the physical tracking limit of the HVS. Based upon the above factors, the universe of discourse for the motion magnitude importance function ranges over [0.0, 85.0]. The following equations formally show the thresholds for the fuzzy membership function:

$$t_{lmag} = max(5.0, M_{avg}) \quad (7.1)$$

$$t_{hmag} = 80.0 \quad (7.2)$$

where:

t_{lmag} is the low magnitude threshold for the motion magnitude importance function, ranging over [5.0, 80.0];
t_{hmag} is the high magnitude threshold, set to 80.0 degrees per second, as per Daly [31];
M_{avg} is the average magnitude differences between regions segmented from the whole scene.

No research has been performed to indicate relative motion effects, except in characterising pop-out [122], which only included observations of motion direction differences, not actual motion magnitudes. For example, one object could be moving at 100 deg/sec, with surrounds moving at 90 deg/sec. It can be hypothesised that both of these regions are unable to be tracked by the HVS. From this it can then be

surmised that the local difference would not attract attention due to temporal masking and blurring effects obscuring the local difference in motion values, due to the movement of the untracked regions across the retina [51]. Subsequently, the absolute motion value of the region being examined is thresholded to 80 deg/sec before being processed for relative motion analysis, to prevent these circumstances influencing the final motion importance of the object.

In addition, research has indicated asymmetry in the pop-out induced by motion. That is, a moving object on a stationary background is far easier to see than a stationary object against a moving background [36]. Similarly, a slow moving object against a fast background is not as easy to distinguish as a fast object on a slow background [75]. Therefore, in order to model this effect, only moving regions are considered for the motion importance calculations. This removes the case of stationary objects having large relative motion differences causing inappropriate pop-out. This is implemented by simply assigning a zero value to the motion differences for a stationary region.

In the case of an object changing speed and having its absolute velocity fall under the threshold of visibility, then the model is able to respond due to the previously mentioned threshold. When it has been established that the object is moving under the 80 deg/sec threshold, then the object will pass through the filter and contribute to the importance calculations due to motion.

Motion direction vectors are handled in a similar manner, with only one adaptive threshold t_{dir} derived from the mean value of the direction differences, measured as θ (radians)–with absolute values being taken with reference to the x axis. The bottom threshold is set to 0.35 radians (20 degrees), if there is no background activity [122]. The motion threshold value is derived using the following equation:

$$t_{dir} = max(0.35, M_{dir}) \tag{7.3}$$

where:

t_{dir} is the adaptive threshold for the motion direction membership function;
M_{dir} is the average value of the motion direction differences between regions within the scene, in radians.

7.2.2 Onset Membership Functions

The handling of sudden onsets with regards to pop-out is not treated as an adaptive luminance change function. The membership function is not adaptive in nature, as the formula used allows for any adaptation directly (refer to Figure 7.5).

Essentially, the effect of abrupt onset is reduced to luminance amplitude change effects, regarded to be the proportion of segments per region that have changed luminance, as a ratio over how many segments have changed in the entire image. The amplitude of change is treated in a similar fashion to luminance contrast, with there needing to be greater than 1% contrast from frame to frame before it is considered a noticeable change [166].

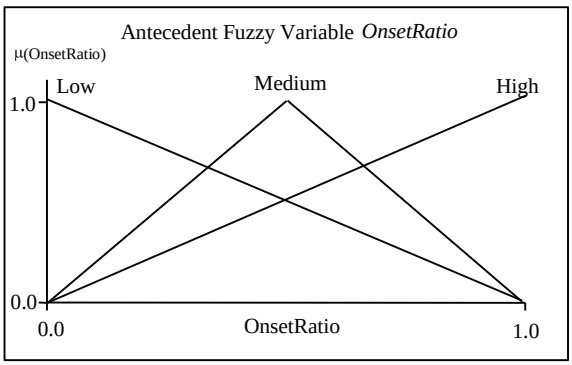


Figure 7.5 Illustration of abrupt onset membership function.

The onset value is calculated according to the following equation:

$$OnsetRatio = NumRegSegCh / NumImageSegCh \quad (7.4)$$

where:

OnsetRatio is the fuzzified onset value;

NumRegSegCh is the number of segments within the region that have changed;

NumImageSegCh is the number of segments within the entire image that have changed.

7.2.3 Temporal Change Evaluation Rules

In a similar manner to the spatial system developed in Chapter 3, the system uses the following rules to evaluate the temporal importance of a segmented region:

IF *MagDiff*	IS High	THEN *FinImp* IS High
IF *MagDiff*	IS Med	THEN *FinImp* IS Med
IF *MagDiff*	IS Low	THEN *FinImp* IS Low
IF *DirDiff*	IS High	THEN *FinImp* IS High
IF *DirDiff*	IS Med	THEN *FinImp* IS Med
IF *DirDiff*	IS Low	THEN *FinImp* IS Low
IF *OnsRatio*	IS High	THEN *FinImp* IS High
IF *OnsRatio*	IS Med	THEN *FinImp* IS Med
IF *OnsRatio*	IS Low	THEN *FinImp* IS Low

These rules are integrated into the region-based importance mechanism. Therefore, the aggregation and defuzzification methodologies are the same as used in the spatial importance system.

7.2.4 Integration into Spatial Visual Attention Model

The temporal importance rules are used in the same way as the other spatial importance rules, in a multiple-additive manner [7]. The temporal changes determined by object motion and abrupt onset are combined into the final importance value for the regions. This concurs with present thinking on the close relationship between onset and motion effects in visual search [177].

The only anomaly in the rule base is the case of object-based motion occurring when there is no luminance change in the image segmentation. For this scenario, it would be inappropriate to allow the motion importance from object information to influence the visual importance of the region, as no visible change has taken place in the image. If the object that has moved has no subdivision changes with a temporal luminance contrast above 1%, then the motion vectors associated with the region are set to zero.

The only modification to the implication process lies in the weightings applied to the spatial and temporal rules. The temporal rules receive a higher weighting value due to the importance of temporal image changes. In a manner similar to other models, the implementation here uses 0.6 for the temporal rules and 0.4 for the other spatial rules [116, 128]. The next section describes how the temporal visual attention model is integrated with the spatial visual attention model.

7.3 A MOTION-BASED ADAPTIVE ANIMATION RENDERING APPROACH

In order to incorporate the above model into an adaptive rendering approach a number of stages must take place. The system must make some segmentation of the scene based upon motion information, using previous frames and region importance maps. The approach must also compensate for ego motion caused by camera movement. Next, the temporal model is applied to the motion vector estimates from the segmented regions to produce a relative visual importance value for moving region. Finally, the importance value is used to control the adaptive rendering system. The major components of this approach are depicted in Figure 7.6.

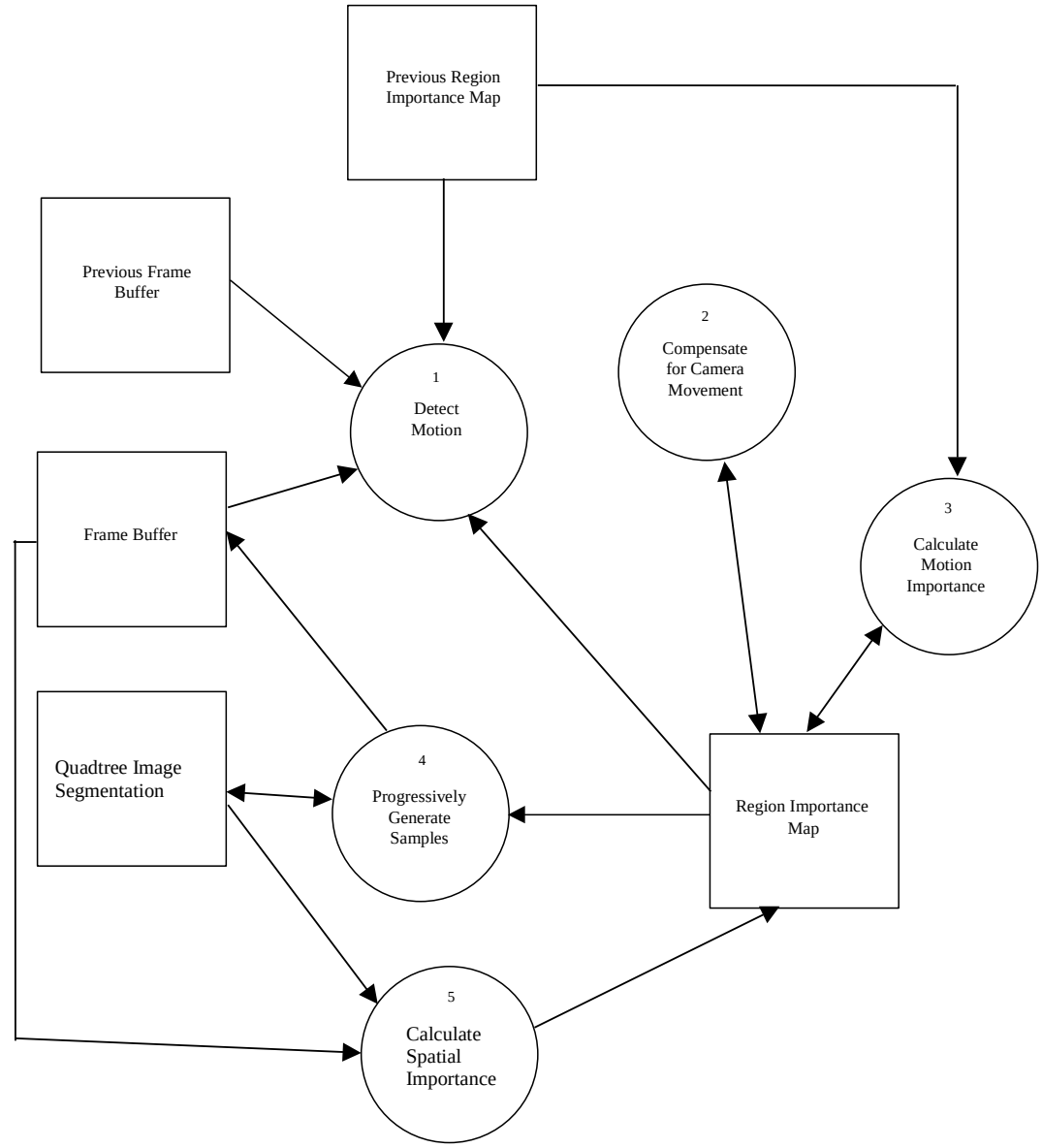


Figure 7.6 Flow diagram of the major stages in the temporal change approach.

In the newly developed approach, the calculations are performed from frame to frame, this requires the approach to have the *Previous Frame Buffer* stored to facilitate the change analysis. While the image is progressively sampled to the level of one sample per pixel, the present and previous frame buffers are analysed for motion. A previous region segmentation is stored in order to facilitate the motion importance calculations. Once region motion vectors have been derived for the regions in the scene, then these vectors are processed to remove any camera motion. The resultant vectors are then fed to the motion membership functions that are detailed in Section 7.2.1. The resultant motion importance values are stored in the

R*egion Importance Map*. The motion importance value is then integrated with the spatial importance value to form a spatiotemporal region importance value stored in the present region importance map. In a similar fashion to that detailed in Chapter 5, the importance value is used to modulate the supersampling performed within each pixel.

7.3.1 Motion Estimation Technique

There are two major methods for motion estimation within the area of image synthesis. The motion estimation can be performed in an image-based manner, similar to video systems, or by using object-based techniques.

The image-based techniques typically involve estimating the location of how far a subdivision or block has moved within the image. Image-based approaches often use a block-based search window to minimise the Mean Square Error (MSE) of the block being examined, in comparison with other blocks within the window. The block with the minimum error is considered the location where the block has moved. This predicted displacement is used to calculate motion vectors for compression techniques in video transmission [87].

Object-based methods exploit the object-space geometry and the associated transformation matrices to estimate where a geometry component will be translated to on the screen [1, 56, 165, 187]. Object-based approaches perform better than image-based methods in 3D animation applications, due to the unambiguous nature of the object ID information. A number of motion estimation techniques have been developed for image synthesis to facilitate compression of synthetic movies.

The motion estimation process used by Guenter et al. [56, 187] is pixel-based, utilising pixel RGB colour, object ID and depth buffer values. The estimation technique uses these parameters as an aid in determining, in the forward direction from frame n to frame n + 1, the position of a four pixel (2×2) square. Error thresholds from depth and object ID information are used to determine whether a predicted pixel matches the n + 1 frame in the sequence when transformed from frame n.

Wallach et al. [165] use hardware gouraud shading and texturing techniques to garner information about optical flow within the image being generated. The mode vector of the 16 × 16 pixel optical flow block is used as the centre of a brute force search. This is similar to the approach by Guenter et al., due to the need to accommodate more than one object being within a block being processed. Again, the technique does not allow for non-translational optical flow, but it does give a very efficient hardware-assisted method to gain optical flow vectors for a synthetic animation.

Agrawala et al. use back projection to ascertain the location of a pixel in the previous frame to the one being examined [1]. The transformation and projection matrices are used to obtain the position of a pixel in object-space in the previous frame. The difference between the two gives an object-space accurate optical flow motion vector for the pixel. Depth and object ID information is also used to help deal with occlusion problems from frame to frame, in a similar manner to Guenter et al. [55, 56]. Their results show that for their compression application, a hybrid method using brute force window methods and least squares methods for block vector estimation performed better overall. The block search method is applied to blocks containing object ID edges, or newly uncovered object IDs. A least squares estimation scheme is then used to compute the transformation matrix for the other 8 × 8 pixel blocks within the scene. The least squares scheme allows for non-translational optical flow, an improvement on previous methods.

As the temporal change approach developed here continues the region-based paradigm, it is appropriate that the motion estimation technique will be developed from a region-based perspective. Furthermore, it can be argued that the perceived motion in a scene is region-based in nature, due to a person focusing on regions in an image, and not pixels or blocks (refer to Section 3.2).

As an improvement to the segmentation techniques used in Chapter 5, the region merge algorithm has been modified to include object IDs. The motion estimation

scheme detailed here obtains regions by segmenting the scene using the object ID as a basis for the comparison operations in the subdivision merging stage. Once the image is segmented by object ID, then the regions are further segmented by luminance and hue features. This provides a two level hierarchy, where the top-level regions are segmented by object ID and the second level regions are segmented by luminance and hue features, as per the method used in Chapter 5. This facilitates the detection of gross region motion effects within the top-level, while internal motion effects are detected within the second segmentation level.

The Object ID information is gained from Attribute command information in the Renderman© file format [162]. This may be obtained from various levels of the hierarchy that constructs the objects in the scene. For instance, an object named as a bike may be made up of a number of components: wheels, frame, handle bars etc. At this stage the object IDs are obtained from near the top of the hierarchy, so complex constructs are considered objects in this model. This will match, in most cases, the construction and setting of the scene as a background with a number of objects in the foreground, for example, a room scene with furniture. However, more sophisticated methods could be employed, based upon the projected area covered by the boundary of the object ID in question [169]. If the object ID used does not refer to a large enough projected area, then an object ID can be chosen from a higher point in the hierarchy.

As well as aiding the correct segmentation of regions for more accurate motion calculation, the use of object IDs speeds up the merge segmentation algorithm. The segmentation algorithm is now divided into two main processes. The first is the merging of subdivisions that have the same object IDs. This can be performed in a serial fashion by simply scanning the subdivisions from top to bottom, left to right, placing them in subdivision lists identified by the mode of the object ID samples within the subdivisions.

The lists of subdivisions are then processed using the merge algorithm previously developed in Chapter 5. The subdivisions are divided up into regions based upon

hue and luminance differences (refer to Figure 7.7). This segmentation approach is incremental in nature. If a subdivision changes object ID, luminance or hue through further sampling, then the subdivision is reallocated to another region, triggering new importance calculations to update the importance map.

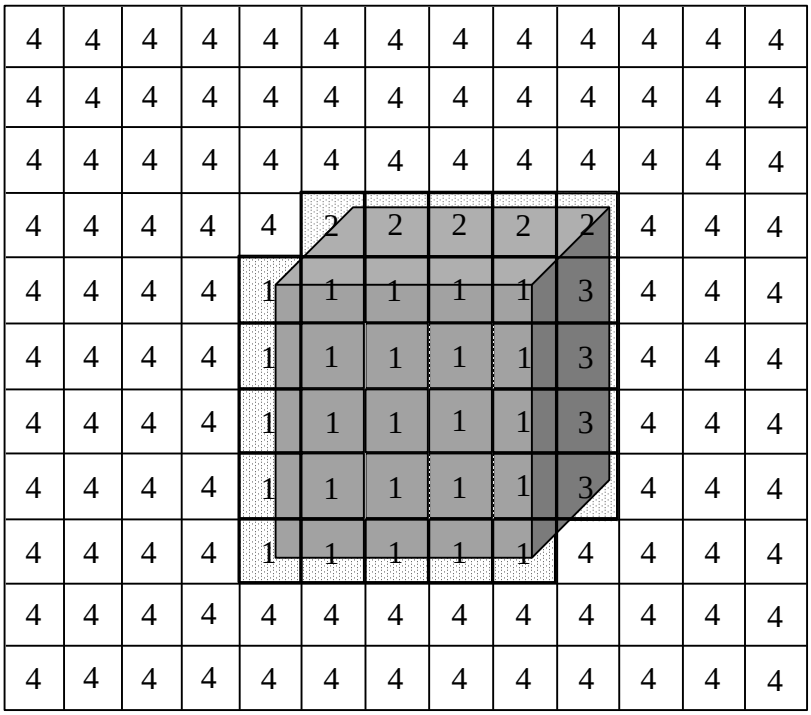


Figure 7.7 An example of the hierarchy of segmentation used in the motion esitmation system. The colour of the subdivision represents object ID segmentations. The dotted area represents one object ID, while the white background represents another object ID. The numbers represent the segmentation within the object ID segmentation. In the example, the cube region has been further subdivided into three regions (1, 2, 3).

Due to the inherent correlation of the object ID with the motion in a scene, the segmentation based on object IDs provides effective search windows for further internal motion estimation. These windows are more accurate than arbitrary sized square regions, which do not necessarily contain the blocks causing the perceived motion. This brings about better matches when performing motion prediction within an object region.

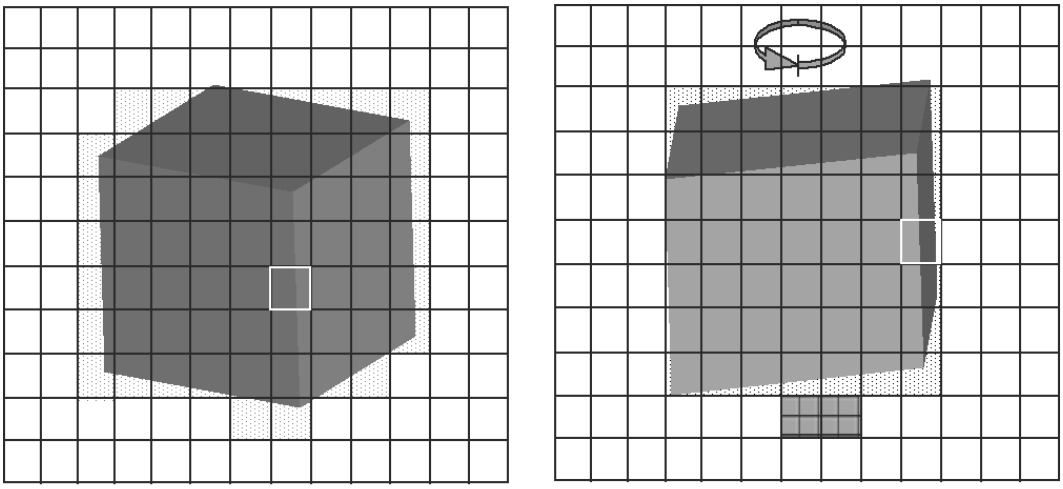


Figure 7.8 Illustration of the internal motion search method over two frames (frame n on the left, n + 1 on the right), within the regions segmented at the level of object IDs– dotted regions surrounding cube. A subdivision which changes from frame to frame is highlighted in white. A subdivision which changes across two regions is highlighted by a cross hatch pattern in the second frame.

For video compression systems, motion vectors must be collated across the whole scene for every pixel and block [87]. Transmitting the change vector, instead of the actual image data, reaps efficiency savings. As this application is the computation of region-based visual attention using motion differences, there is not such a need to search for block motion outside of the area segmented by an object ID. This approach only requires estimates of the motion of the segmented region across the image, and the internal motion of a region. Therefore, the algorithm works within the object ID region to search for internal motion. This internal motion difference can be processed in the same manner as the gross region motion, to gain a measure of internal motion importance for each region.

Subdivision changes may occur across different region segmentations, as shown in Figure 7.8 by the hatched subdivisions (marked  in Figure 7.8). They can be classified as the appearance of a new object within the segmentation, and therefore are treated as an abrupt onset change, as per the model developed in Section 7.2.

The possible object motions can be roughly divided into two categories: rigid and non-rigid. Rigid transformations preserve the shape of the object in 3D, and occur due to object translation and rotation. Non-rigid deformations occur due to

transformations that deform the shape of the object in 3D. A combination of these two may occur for any object in the scene. The method developed here will process both forms by performing a hierarchy of motion calculations. The first level is the gross motion of the object ID segmented region, while the second is the internal motion of the hue and luminance segmented regions.

The motion to be calculated for both the region levels is translational in nature. This will still effectively model the internal motion of the regions–for example, the spinning of a cube (refer to Figure 7.8). The object ID level of segmentation will compute importance for the translation of the cube through the scene. The second level, which represents the segmentation of the gross region into similar hue and luminance regions, will give an estimate of the internal motion importance of the object.

To identify the motion of the gross object ID regions is straightforward. The method searches for a corresponding region object ID from frame $n + 1$ in the previous region segmentation for frame n. The centroids of the regions in the two frames are then subtracted to form a motion difference vector $MObjectID_r$. This vector is then used to compute the motion importance for the gross region motion. If the object ID cannot be found in the previous region importance map, then the motion vector M_r is set to zero. An object entering a scene will, on the first frame, be treated as a sudden onset region. Frames occurring afterwards with this object will then process its movement in a normal fashion.

The internal motion of the regions uses a region matching technique, similar in nature to the block matching techniques used in video compression. In this case the method is modified to match regions, not blocks, for efficiency purposes. The object ID level is used as the search window to search for matching regions. An internal region in frame $n + 1$ is compared to the internal regions within the previous importance map for frame n. The method minimises the MSE of the compared regions, in order to find the closest match. In a similar manner to the object region calculations, the two internal region centroids are subtracted from each other to gain a motion vector $MInternal_r$.

This motion and onset information is then passed to the camera compensation module to remove any camera motion from the vector.

7.3.2 Camera Compensation Technique

Before a motion importance value can be computed, the component of motion due to the camera must be removed from the computed region motion values. Motion in an image can be caused by both the movements within the scene, and the spatial transformation of objects within the scene. Camera motion forms a background motion noise, which needs to be suppressed in order to ascertain correctly the true changes in the scene for motion importance purposes [128].

The advantage with image synthesis camera compensation is the availability of the world to camera space transformation T_{wc}, and the projection matrix P. These two transformation matrices enable a complete model of the contribution of the camera to region motion estimates. Before the region being examined is searched for in the previous frame region list, its centroid is transformed by the opposite of the difference in world to camera transformation matrices, thus removing any image-plane motion produced by the view camera.

The following notation is used in the equations detailed in this section:

- $T_{wc,\ n}$ represents the camera transformation matrix–from world to camera coordinates–for frame n;
- P represents the projection matrix used for the scene–orthogonal or perspective;
- $C_{r,\ n}$ represents the centroid of a segmented importance region r in frame n;
- $M^*_{r,\ n}$ represents the final image-plane 2D region motion vector prefix for region r in frame n–eg. $MFin_{r,\ n}$.

The difference between the two camera transformations ΔT_{wc} is formed by analysing the world to camera space transform matrix T_{wc}. The following equation provides the camera transformation matrix between two frames n and $n + 1$:

$$\Delta T_{wc} = T_{wc,\, n+1} \cdot T_{wc,\, n}^{-1} \tag{7.5}$$

where:

ΔT_{wc} is the camera motion transformation from frame n to frame $n + 1$;
$T_{wc,\, n+1}$ is the world to camera space transformation for the frame $n + 1$ in the animation;
$T_{wc,\, n}^{-1}$ is the inverse of the world to camera space transformation for the frame n in the animation.

The inverse of ΔT_{wc} matrix, ΔT_{wc}^{-1}, is then applied to the centroid $C_{r,\, n+1}$ of the region being camera compensated, to remove the motion caused by the camera. This means that the final 2D motion vector $M_{r,\, n+1}$ for a region r, for frame $n + 1$ is:

$$C'_{r,\, n+1} = C_{r,\, n+1} \cdot P^{-1} \cdot \Delta T^{-1}_{wc,\, n+1} \cdot P \tag{7.6}$$

$$M_{r,\, n+1} = C'_{r,\, n+1} - C'_{r,\, n} \tag{7.7}$$

where:

$C'_{r,\, n+1}$ is the camera compensated centroid for the region being examined in frame $n + 1$;
$C'_{r,\, n}$ is the camera compensated centroid of the region in the previous frame n;
P, P^{-1} are the view projection matrix and its inverse;
$M_{r,\, n+1}$ is the final 2D motion vector (Δx, Δy, z ignored) computed for region r in frame $n + 1$;
$\Delta T^{-1}_{wc,\, n+1}$ is the inverse world to camera space transformation for frame $n + 1$ in an animation.

The final calculation is the addition of the internal region motion vector with the object ID region vector to produce the overall motion of the internal region. This is accomplished by the following equation:

$$MFin_{r,\,n+1} = MObjectID_{r,\,n+1} + MInternal_{r,\,n+1} \tag{7.8}$$

where:

$MFin_{r,\,n+1}$ is the final vector combining the gross and internal region motion;
$MObjectID_{r,\,n+1}$ is the camera corrected gross region motion vector for frame $n + 1$;
$MInternal_{r,\,n+1}$ is the camera corrected internal region motion vector for frame $n + 1$.

The final 2D motion vector $MFin_{r,\,n+1}$ is passed to the motion evaluation component of the temporal change model to derive a motion-based visual importance value.

7.3.3 Adaptive Image Synthesis Animation

In order to exploit these temporal importance values drawn from the importance map, a new framework for animation rendering must be fabricated.

In the past, adaptive rendering for motion has been handled in a number of ways. Distributed rendering is a technique used to simulate motion blur caused by shutter speed effects in cameras [52]. Other temporal motion detection models have been used to perform motion compensation for video compression of image synthesis animations [1, 56, 165, 187]. They typically detect changes in pixels and create pixel flow vectors by identifying which object has been intercepted at the pixel level and then tracking the transformation of the pixel with the object ID to the next frame, using the 3D transformation matrices contained within the animation script. This method, while accurate, is restrictive as it requires the transformation of every pixel within the image to ascertain pixel flow vectors for the next scene. In addition, the methods may require hardware support in order to be efficient, due to the overhead

of performing the calculations for every pixel [165]. The region-based techniques developed in this chapter are much more efficient due to the eschewing of pixel-based motion estimation, in favour of region-based motion estimation.

A multiresolution model of temporal importance has been implemented [185, 186] in order to control sampling rates in a ray tracing system. As with the critique shown in Section 3.1, the motion model designed by Yee has the same difficulties that were addressed in the spatial importance model in Chapter 4. The model uses a pixel-based absolute value motion model, which lacks the ability to deal with the relative motion of regions and requires a hardware prerendering of the scene to provide information to the motion model used. From these observations it seems that no work has been developed for the region-based processing of motion for image synthesis efficiency purposes. In addition, the region-based method developed here is more in line with present psychophysical thinking on object-based visual attention, and should be more efficient being region-based rather than pixel-based in its calculations. Furthermore, in the same manner as the spatial model, the motion importance animation technique is truly progressive. The approach uses the early samples of the scene to make estimates of region importance, and does not require a hardware prerendering to ascertain motion importance values.

The adaptive and progressive methods used in this approach will continue the concepts developed in Chapter 5 by modifying the supersampling rate of a region according to its visual importance. The supersampling techniques to be used are the same, with flat and perceptual methods of pixel subdivision control, modulated by the importance of the region containing the pixel. Furthermore the region importance algorithm within the progressive rendering approach needs to be modified in order to obtain frame-to-frame changes in luminance and region motion.

The algorithm developed for progressive image synthesis outlined in Chapter 5 has been modified for animation and is detailed in the following: Algorithm 7.1, Algorithm 7.2 and Algorithm 7.3. The overall approach progressive rendering approach is the same, except for additions to the region importance evaluation function *EvalRegImp* that enables the calculation of motion information for visual

importance calculations. Most of the motion calculations are performed in Algorithm 7.3, with the generation of vectors for each of the regions contained within the ObjectID tag segmentations.

Procedure: EvalRegImp

Inputs: List of *regions* containing segmented region information for image.
List of importance subdivisions *elemSeg* containing 8 × 8 pixel information

Outputs: Nil

EvalRegGlobalDiff(*regions, numReg, lumDiffMean, hueDiffMean, sizeDiffMean, contDiffMean, motMagDiffMean, motDirDiffMean, onsetTotal*)

for *regNum* = 0 to *numReg* **do** { For each region do }
Set *surrLum, surrHue, surrContDens, surrSize, surrMotMag, surrMotDir* ← 0

for *surr* = 0 to *regions*[*regNum*].*bordCount* **do** { For each surrounding region do }
Set *surrReg* ← *elemSeg*[regions[regNum].*border*[*surr*]].*regNum*
Add the surrounding local feature values of region *surrReg* to *surrLum, surrContDens, surrSize, surrHue, surrMotMag, surrMotDir*

if *regions*[*regNum*].*motMag* > 0 and *regions*[*surrReg*].*motMag* > 0 **then** { Only add to differences if both regions are in motion }
Add the *surrReg* region's *motionMag* to *surrMotMag*
Add the *surrReg* region's *motionDir* to *surrMotDir*
end if
end for

{ Obtain local feature difference values, with respect to surrounding average feature values }

Set *bordCount* ← *regions*[*regNum*].*bordCount* { Set count of *regNum* surrounding regions }
Set *lumDiff* ← | *regions*[*regNum*].*lumAvg* - *surrLum* / *bordCount* |
Set *hueDiff* ← | *regions*[*regNum*].*hueAvg* - *surrHue* / *bordCount* |
Set *contDensDiff* ← | *regions*[*regNum*].*contCount* / *regions*[*regNum*].*segCount* - *surrContDens* / *bordCount* |
Set *sizeDiff* ← | regions[*regNum*].*segCount* / *numElem*) - *surrSize* / *bordCount* |

{ Obtain absolute feature values }

Set *loc* ← ((*regions*[*regNum*].*centreY* - *dimElem* / 2.0)^2 + (*regions*[*regNum*].*centreX* - *dimElem* / 2.0)^2)$^{-2}$ / ((*dimElem* / 2.0)^2 + (*dimElem* / 2.0)^2)$^{-2}$
Set *edgeProp* ← *regions*[*regNum*].*imEdgeCount* / (*dimElem* × 2.0 - 2.0)
Set *onsetRatio* ← *SegChanged(regions[regnum]) / onsetTotal* { Ratio of regNum changed subdivisions to overall number of image subdivision changes }

if *onsetRatio* <> 0 **then** { Add motion if change in region }
Set *motMagDiff* ← | regions[regNum].*motMag* - *surrMotMag* / *bordCount* |
Set *motDirDiff* ← | regions[regNum].*motDir* – *surrMotDir* / *bordCount* |
end if

{ Obtain visual importance of region using fuzzy logic system designed in Section 4.2 }

RegImp(*imp, lumDiff, lumDiffMean, hueDiff, hueDiffMean, sizeDiff, sizeDiffMean, contDensDiff, contDiffMean, motMagDiff,* motMagDiffMean, *motDirDiff, motDirDiffMean, onsetRatio, loc, edgeProp*)
Set *regions*[*regNum*].*imp* ← *imp*

if *regImpMax* < *imp* **then** { Determine the maximum and }
Set *regImpMax* ← *imp* { minimum importance values }
end if

if *regImpMin* > *imp* **then**
Set *regImpMin* ← *imp*
end if
end for

NormRegImp(*regions, regImpMax, regImpMin*) { Normalise values to [0, 1] }

Algorithm 7.1 Modified region importance algorithm EvalRegImp, incorporating new highlighted motion importance calculations.

```
Procedure: EvalRegGlobalDiff

Input:    List of segmented regions containing visual feature information
          Number of regions numReg
Output:   A number of global feature activation variables: lumDiffMean, hueDiffMean, sizeDiffMean, contDiffMean,
          motMagDiffMean, motDirDiffMean, onsetTotal.

Set diffCount ← 0
Set objListPres ← {ObjID1, ObjID2, ... , ObjIDn}          { List of ObjectID records containing
                                                            regions with same object IDs for
                                                            present frame buffer }
Set objListPast ← {ObjID1, ObjID2, ... , ObjIDn}          { List of ObjectID records containing
                                                            regions with same object IDs for past
                                                            frame buffer }
EvalRegionMotion(objListPres, objListPast)                { Obtain motion vectors for regions }

for regNum = 0 to numReg do                               { For each region do }
    for surr = 0 to regions[regNum].bordCount do          { For each surrounding region do }

        Set surrReg ← elemSeg[regions[regNum].border[surr]].regnum   { Obtain surrounding region number
                                                                       from index }
        Set revSurrReg ← 0                                { Find the link to the present region
                                                            regNum }
        while regNum <> elemSeg[regions[surrReg].border[revSurrReg]].regNum do
                                                          { from the surrounding region
                                                            surrReg }
            Increment revSurrReg by 1
        end while

        { If present region has not been compared to the surrounding region before then add feature values to global feature }
        { difference variables }

        if not regions[surrReg].checked[revSurrReg] then  { If regnum, surrReg not checked }
            Add difference between surrReg and regNum regions to total variables: lumDiffMean, contDiffMean,
                sizeDiffMean, hueDiffMean.

            if regions[regNum].motMag > 0 then
                Add motion differences between surrReg and regNum regions to total variables: motMagDiffMean,
                    motDirDiffMean.
            end if

            Set regions[regNum].checked[surr] ← True;     { Two regions have been compared }
            Increment diffCount by 1
        end if
    end for

    if regions[regNum].motMag == 0 then                   { Check for onset }
        Set onsetTotal ← SegChanged(regions[regnum])
    end if
end for

if diffCount <> 0 then                                    { Averages for feature differences }
    Divide lumDiffMean, contDiffMean, sizeDiffMean, hueDiffMean, motMagDiffMean, motDirDiffMean by diffCount
end if
```

Algorithm 7.2 Algorithm listing of modified procedure EvalRegGlobalDiff, with motion importance additions highlighted.

```
Procedure: EvalRegionMotion

Input:  List of records containing regions with same ObjectID tags for past frame buffer objListPres
        List of records containing regions with same ObjectID tags for past frame buffer objListPast

Output:

for obj = 0 to numObj do                                        { For each objectID in the image }

    if obj ∈ objListPast then                                   { If not new object (onset) }

        for regNum = 0 to numPresObjReg do                      { For each region with present objectID }
            Set presObjX ← presObjX + objListPres[obj].regions[regNum].centreX   { Obtain centre of object )
            Set presObjY ← presObjY + objListPres[obj].regions[regNum].centreY
        end for

        for regNum = 0 to numPastObjReg do                      { For each region with past objectID }
            Set pastObjX ← pastObjX + objListPast[obj].regions[regNum].centreX   { Obtain centre of object }
            Set pastObjY ← pastObjY + objListPast[obj].regions[regNum].centreY
        end for

        Set MObjectID[x] = presObjX – pastObjX                  { Calculate raw gross object vector }
        Set MObjectID[y] = presObjY – pastObjY
        Set MInverseCam ← EvalInvCamTrans()                     { Calculate inverse camera transform }
        Set MInternal ← EvalIntMSEVec()                         { Calculate raw internal object vector }
        Set MFin ← MObjectID + MInternal                        { Calculate resultant vector–gross and internal }
        Set MFin ← MFin · MInverseCam                           { Remove camera motion from final vector }

        for regNum = 0 to numPresObjReg do                      { For each region with present objectID }
            Set objListPres[obj].regions[regNum].motMag ← |MFin|          { Set magnitude of region vector }
            Set objListPres[obj].regions[regNum].motDir ← MFin / |MFin|   { Set normalised region direction vector }
        end for
    else
            Set objListPres[obj].regions[regNum].motMag ← 0     { Set to zero to allow for onset effects }
            Set objListPres[obj].regions[regNum].motDir ← 0
    end if
end for
```

Algorithm 7.3 Algorithm listing of procedure EvalRegionMotion, which performs the actual motion estimation calculations.

7.3.4 Time and Space Complexity of Approach

Most of the components in the progressive rendering algorithm remain the same as in Chapter 5. Each frame is progressively rendered to the level of single pixel size then, as before, the supersampling regime subdivides the pixel according to its visual importance. The new components to be analysed are contained within the region segmentation and region importance calculation modules. The following sections detail the time and space complexity constraints on the newly developed algorithm.

Region Segmentation

In each of the expressions derived for algorithmic complexity, the values are with respect to the number of subdivisions within the image, rather than the number of pixels within the image. This is due to algorithms being performed on a subdivision

by subdivision basis, as the subdivision forms the basis of the region importance map generated for the animation frame.

In the case of every subdivision having a unique object ID, the region segmentation algorithm performs at its most efficient. The segmentation method only requires a serial scan through the list of subdivisions to allocate a list for each of the objects, with some small overhead with regards to subdivision list maintenance. No further segmentation has to take place. In this scenario the algorithm complexity expression is n.

The worst-case scenario is the entire scene being composed of one object ID. In this case, the region segmentation reverts to luminance and hue comparisons to subdivision the scene in a less efficient manner than the serial allocation of lists of subdivisions with the same object ID. In this case, the segmentation efficiency is contingent on the variation in luminance and colour intensities within the image. In other systems the segmentation may be controlled by parameters to enforce a bottom-level size for the regions [128]. This bottom-level would be enforced in an implementation of this segmentation algorithm.

Spatial complexity is increased over previous methods of region segmentation, due to the need to store the object ID within each pixel. A machine word sized pointer is needed to identify the object intersected by the pixel. Therefore, the algorithm requires n words more of storage, compared to the original segmentation algorithm, where n represents the number of pixels within the image.

Region Importance

In the following complexity analysis of the region importance algorithm, the values are given with respect to the number of segmented regions within the image, as this is the atomic unit processed by the region-importance algorithm.

The best-case scenario is when the entire scene is only one region. In this case the expression reduces to a constant, as there are no other regions to be compared.

The worst-case scenario is that all the subdivisions have different object IDs. This means a high number of regions would occupy the image, causing large computational overheads on the motion comparison computations. As the segmentation is performed with a four-connected set of subdivisions, the expression is 5n, due to the need to compare every region with its four surrounding regions to ascertain global activity within the image, and then a final pass to gain the importance values for each region within the context of the global movement occurring.

Space complexity is larger for this method compared to the still image importance approach, due to the need to store a copy of the frame buffer, the region segmentation from the previous frame and new motion information needing to be stored for each. This means that the memory requirements are approximately double the spatial model. Therefore, the expression for space requirements in bytes for the motion model is 14n + 8r. With n being the number of pixels in the image (to store the two frame buffers) and r being the number of regions in the image.

Even though there are costs involved with maintaining the region motion information, the algorithm still scales well, being linear in nature in both time and space complexity. This is due in the major part to the algorithm being reliant on image-space information, in which the number of subdivisions and regions varies linearly with the size of the image.

7.4 DISCUSSION

This chapter has detailed the development of a novel region-based temporal change model. The major achievements of this chapter are:

- The development of a region-based temporal change model that uses region motion differences, not just absolute motion values. This more closely follows psychophysical models of visual pop-out.
- The development of a motion model that more fully characterises region motion as being a combination of the gross regional motion and the internal regional motion. This allows the model to produce an

accurate estimation of region motion, which allows for both translational motion and the internal effects from rotation and non-rigid deformation of the object in the scene.

- The development of an improved region segmentation algorithm, which utilises the object ID information returned from the rendering system. This removes ambiguity problems caused by the coarse segmentation of the scene, as the object ID has an unambiguous relationship to the image-plane region segmentation. Furthermore, this facilitates more accurate and efficient calculation of region motion.
- The modification of supersampling techniques to accommodate region-based motion importance values.
- The development of a novel image synthesis-based camera compensation model for motion estimation. Camera compensation has been used in video motion importance calculations, but this method is novel due to the use of object-space transformation information to remove camera motion from the derived motion vectors.

At this stage, the model and the associated techniques have been fully designed and analysed. The next process is their implementation and incorporation into the present visual importance rendering system. This task is relatively straightforward due to a number of factors. Firstly, the frame buffer and region importance map data structures already exist and have been implemented. The addition of motion importance map information is an incremental process. Secondly, the theory has been derived for the calculation of region-based motion importance, including a detailed derivation of camera transforms, vector calculations and modified algorithms. Finally, the framework for still image supersampling modulation has already been implemented. Therefore, the process of performing frame-to-frame modification of supersampling rates is, again, an incremental process. However, even if the groundwork has been laid, there is still plenty of scope for the

improvement of the visual importance approach by the addition of the newly developed motion importance model.

Subjective Evaluation of Approach

Previous chapters in this thesis have described objective image assessment with analysis of L1 and L2 error ratios for each of the test scenes. This objective approach has reaped useful information regarding the quantitative amounts of distortion in the images, and the locations of these distortions. Objective methods however, do not give an indication of the quality of an image from a human perspective. In order to obtain an estimate of the quality of images as presented to the human viewer, the images generated in the previous chapters will have subjective quality tests performed upon them. A number of models have been developed to simulate human image quality criteria [30, 95, 114, 126]. However, these models do not fully simulate the image quality assessment capabilities of the HVS. Due to this factor, subjective testing is still the mainstay of current research into image quality assessment, to provide a human perspective on the ability of various image processing and image synthesis algorithms. This chapter will now describe such subjective tests upon the images generated by the techniques developed in this thesis.

The structure of the chapter is as follows. Section 8.1 describes the methodology used in the subjective testing of the rendering approach. Section 8.2 discusses the results of the subjective testing performed. Finally, Section 8.3 discusses the results of the objective and subjective testing in an integrated manner, in order to draw out overall conclusions as to the success of the new importance-biased rendering approaches.

8.1 SUBJECTIVE TESTING METHODOLOGY

The experimental task is the comparison of a standard high quality image with a degraded image produced by the visual importance algorithms. Therefore, the most appropriate subjective testing methodology to use is the CCIR ITU-R BT.500-6 [26] stimulus comparison method, which is designed for just this purpose and is in common use in image processing research [5, 6, 126]. In this approach, instead of eliciting an absolute image quality value from the subject, a subjective estimation of the difference in quality between two simultaneously displayed images is recorded. This method is most useful, due to the need to ascertain the visibility of changes in

visual image quality introduced by the importance-biased methodologies. Thus, the methods used were based on the Rec. 500 standard, with modifications contingent on the availability of suitable resources.

Due to the unavailability of a CCIR Rec. 500 video quality assessment room [26], some of the parameters have been relaxed or changed. Instead of a TV monitor with appropriate contrast levels, an Silicon Graphics workstation 19 inch monitor has been set to approximate, as closely as possible, the contrast range specified by the standard [27, 28]. The values for the monitor are listed in Table 8.1.

Furthermore, the room conditions were modified. The monitor was situated in front of a neutral grey wall within a typical office. The viewer sat approximately three to four monitor heights away from the viewing screen (refer to Figure 8.1). The lighting in the room was characterised with an illuminance meter, to register the ambient light levels from the viewing position of the subject. Using a luminance meter, the monitor and its immediate surrounds were measured for luminance levels.

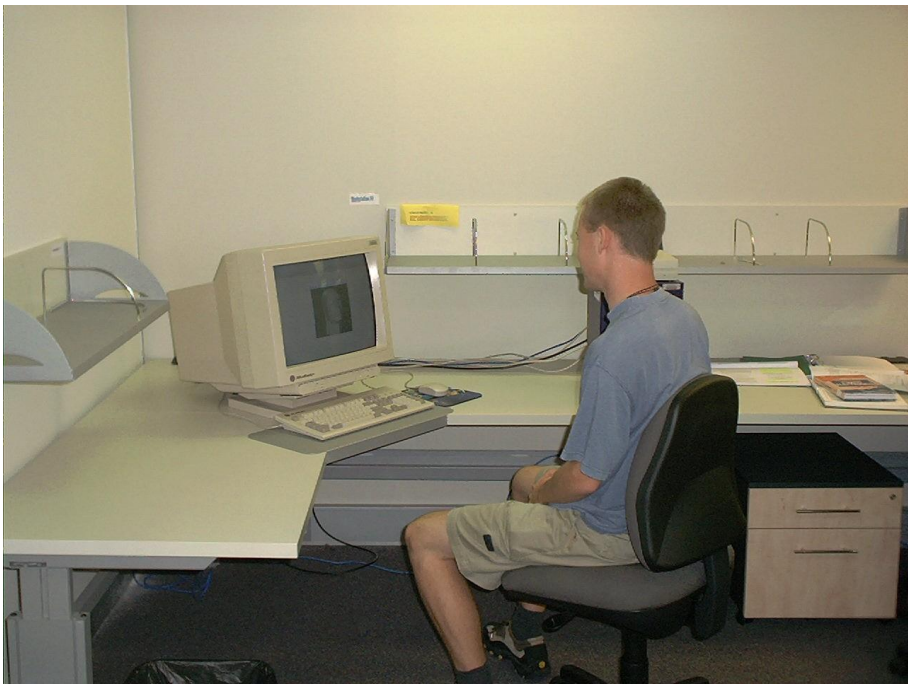


Figure 8.1 A photograph of the subjective testing setup.

Condition Variable	Value
Ambient illumination at viewer position	339 Lux
Approximate viewer distance from monitor	~3-4 screen heights
Average surrounding wall illumination	38.74 cd/m^2
Monitor contrast value	0.11
Monitor peak value (off value)	33.93 cd/m^2 (3.12 cd/m^2)
Monitor Gamma Value	2.4 (recommended by manufacturer)

Table 8.1 Table of experimental conditions for subjective viewing evaluation.

Fifteen subjects engaged in the subjective testing, as specified by the Rec. 500 standard. All subjects had full vision or corrected to full vision. The sample of subjects was drawn from students and staff members of the Faculty of Information Technology, Queensland University of Technology. Out of the fifteen, only two had previous image processing experience. The others were all computer literate, but not image processing or image assessment experts.

The subjects were placed in front of the preview monitor and shown the images. First, a series of 6 images drawn from the test group were displayed to acclimatise the viewer to the quality range of images to be displayed in the assessment tasks. No quality assessment was recorded for these practice images.

The subjective testing approach had both progressive and high quality assessment tasks as its components–matching the two approaches developed within this thesis, progressive rendering and supersampling. The images (refer to Figure 8.2) used in previous objective assessments (refer to Section 5.6) were used in the subjective assessment experimentation. Parameter settings were varied in order to obtain an understanding of what optimal parameters facilitated the use of visual importance in the implemented rendering techniques. These two subjective assessment tasks–progressive rendering and supersampling–are now presented in detail.

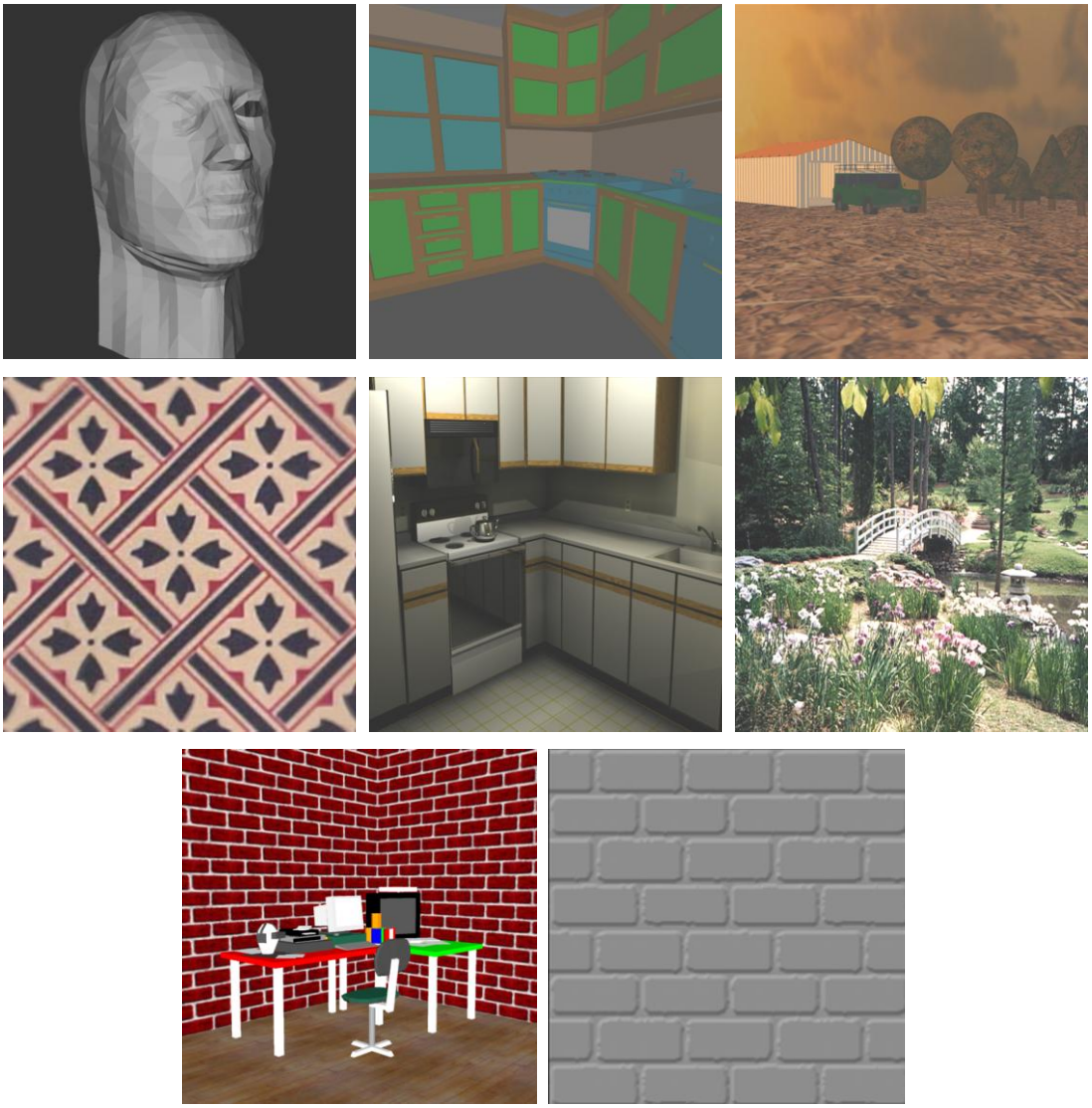


Figure 8.2 Illustration of the images used in the subjective assessment process. The images are from top to bottom, left to right: head, kitchen, farm, cloth, kitchen, garden, texture room and brick bump map.

8.1.1 Progressive Image Assessment Task

A series of four progressive images were shown for assessment. The progressive images represented were a small subset that illustrated the largest L1 and L2 ratio differences between the base and contour accelerated images. A comparison of these particular scenes indicated whether the contour importance module improved the subjective quality of the progressively rendered images. More specifically, the images presented were the following:

- the head scene 8% sampled;

- the kitchen scene 10% sampled;
- the farm scene 8% sampled;
- the brick bump map 7% sampled.

Each progressive assessment task commenced by showing the sequence number of the image to be assessed. A high quality version of the assessment image was then shown for ten seconds. Two images rendered by the base and importance accelerated algorithms were then displayed side by side for 10 seconds. The subjects were asked to evaluate which image was closer in quality to the initial high quality image. A 10 second period of time was then provided, to allow them to mark down their assessment of the relative quality of the images. This image sequence is represented diagrammatically in Figure 8.3.

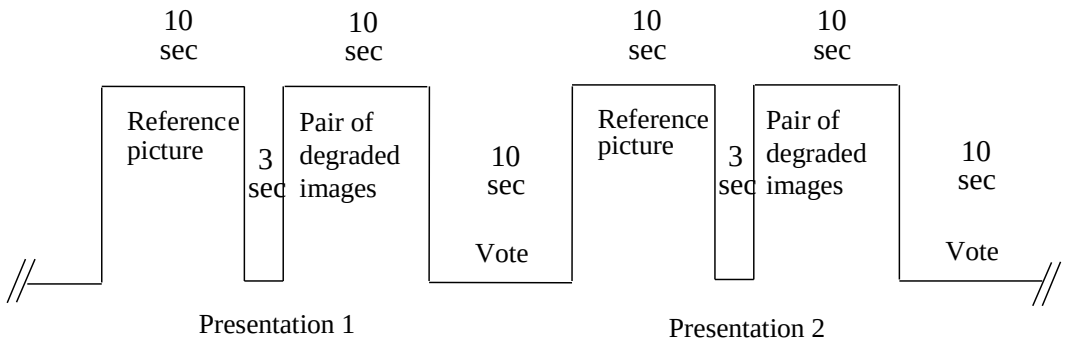


Figure 8.3 Diagram of progressive test assessment methodology, based upon the CCIR methdology for comparitive subjective testing [26].

8.1.2 Supersampling Image Assessment Task

A series of 37 supersampled images were shown for assessment. The supersampling images presented were compared to gain an indication of the visibility of any aliasing introduced by the importance-biased subdivision approach, in comparison to non-biased images. The head, kitchen and forest scenes were produced by both the flat and perceptual supersampling methods, at all the sampling rates displayed in Chapter 5. The flat supersampling scenes were rendered at maximum supersampling rates of 4, 9 and 16 samples per pixel. The perceptual supersampling images were rendered at error threshold values of 10, 20, 30, 40 and 50.

In addition, a series of texture test scenes from Chapter 6 were included into the supersampling test set. The cloth, kitchen and garden textured test scenes were displayed with texture sizes of 257 × 257, 1025 × 1025 and 1537 × 1537 pixels. The images were chosen to test the hypothesis that the quality of the texturing would fall off around the point of the projected pixel size being the same as the maximum size of the filter–in this case 1537 × 1537 (3 times the filter) sized textures.

A set of flat supersampling control images was also included. These images were included to test whether there is a need for visual attention concepts within the supersampling paradigm. During the objective tests, it became apparent that some of the images had little in the way of L1 / L2 error ratio differences. This suggested that there was possibly no subjective difference between an image supersampled once per pixel and an image supersampled further per pixel. If this is true, it would preclude the use of visual importance in the pixel supersampling, due to the pragmatic outcome of only requiring one subdivision per pixel. This, in effect, could mask any subjective differences caused by the spatial modulation of supersampling rates by the visual importance of the image region. Therefore, for each scene, a test was performed by comparing the subjective quality of an image that had only one subdivision per pixel, with an image which had four subdivisions per pixel. This would indicate whether the removal of samples within pixels actually caused subjective loss of quality.

Each supersampling assessment task commenced by displaying the number of the image to be assessed. Two high quality images with and without importance-biased supersampling were then displayed side by side for 10 seconds. The subject was asked to evaluate which image had less visual artefacts. A period of 10 seconds was then given for the subject to mark down their assessment of the relative quality of the images. This process is shown diagrammatically in Figure 8.4.

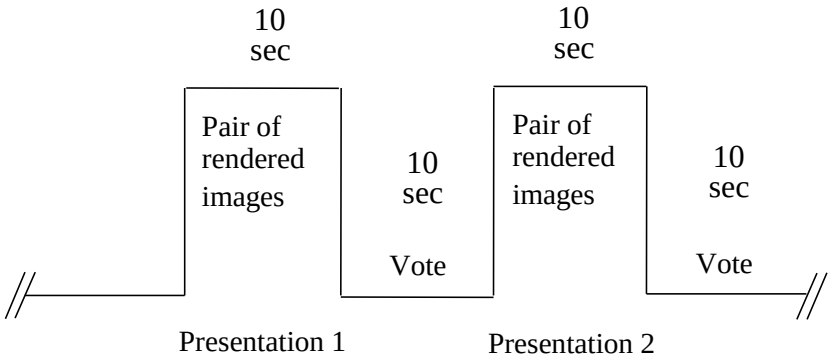


Figure 8.4 Diagram of supersampling test assessment methodology, based upon the CCIR methdology for comparitive subjective testing [26].

The order of the display of the image pairs was pseudo random in nature for both the progressive and supersampling assessment tasks. The images were not displayed in a fixed order, and the location of the impaired image varied randomly from left to right. This was so the subject was not aware of when an impaired image would appear, thus preventing any bias from prediction of the location of the degraded image.

In each task, the subject had a 10 second voting time in which to grade the quality of the images presented. In this period, the subjects marked a scale using a pen, this scale is shown in Figure 8.5.

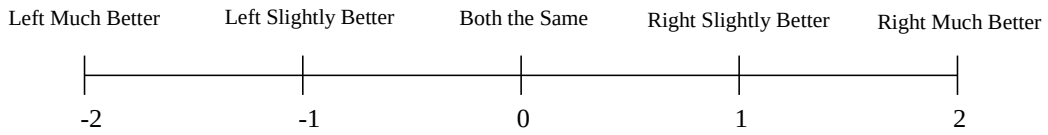


Figure 8.5 Illustration of the five point evaluation scale used by the subjects in the evaluation [26].

The subject was asked to mark the position on the line that represented the perceived amount of improvement in quality in either the left or right image. The measured distance from the centre then became a quantified measure of the difference in image quality perceived by the viewer. The following section analyses the results from these experiments.

8.2 RESULTS

The results in this section are presented for each assessment method as tables containing the mean and standard deviation of the assessment distributions for each of the images, confidence intervals for the sample means and results of hypothesis tests for means not equal to, greater than, and less than zero. Hypotheses are also presented, and then assessed using the appropriate test statistics drawn from the result tables.

Due to the continuous nature of the results from the subjects, and the small sample size, *Student's T Test* was used to accept or reject the null hypotheses presented [104]. The results were processed with reference to the base image on the left and the importance-biased images on the right–after removing the random positioning applied in the experiment. Thus, when the base image appeared to be of higher quality, the value recorded was negative. When the importance image appeared to be of higher quality, the value recorded was positive. Therefore, if no difference was detected, then the value recorded was zero. In each case the *Null Hypothesis* was that the mean of the sample values was equal to zero ($\mu = 0$). The *Alternative Hypotheses* were that the mean was greater than ($\mu > 0$), less than ($\mu < 0$) or not equal to zero ($\mu \neq 0$). Appropriate one-tail and two-tail tests were performed to a confidence level of 95% ($\alpha = 0.05$). Therefore, the one tailed value for t in each test is 1.83, while the two tailed test t value is 2.26. The following sections present, in detail, the results for each of the rendering methods developed in this thesis: progressive rendering, supersampling and texture importance mapping.

8.2.1 Progressive Rendering

Table 8.2 lists the subjective viewing results for the progressively rendered images.

Progressive Image	Sample Mean μ	Sample Standard Deviation σ	Test Statistic Student's t	Confidence Interval $\pm$value	Test H_0 with H_a: $\mu \neq 0$	Test H_0 with H_a: $\mu < 0$	Test H_0 with H_a: $\mu > 0$
Brick 7%	16.00	15.61	3.97	8.64	Reject	Accept	Reject
Head 8%	6.87	22.13	1.20	12.25	Accept	Accept	Accept
Kitchen 10%	-17.80	25.73	-2.68	14.25	Reject	Reject	Accept
Farm 8%	3.13	17.68	0.69	9.79	Accept	Accept	Accept

Table 8.2 Listing of subjective testing results for progressive images. Rejected null hypotheses are shaded in dark grey.

In this case, the hypothesis is that the new progressive method succeeded if the mean of the test sample was shown to be greater than zero ($\mu > 0$). This would indicate that the subjects considered the importance-biased image as having a greater visual quality. A number of things can be noted from the results.

Firstly, the subjects considered the image of the brick wall with bump map contour detection (Section 6.3) to be of better quality than the image generated without bump map contour detection. The null hypothesis was rejected in favour of the alternative hypothesis of $\mu > 0$. This adds support to the efficacy of the method in finding contours quickly within a progressive rendering for an appropriate image. However, one test image does not prove it effective for every case, as was indicated in the objective testing performed in Section 6.3.1. Nevertheless, evidence is presented here to show that images containing these forms of contours will benefit from the contour detection technique.

The other progressive tests were surprising in that they showed the opposite of the expected results. The head image was expected to show a positive improvement, as the objective tests revealed it to have the largest difference in L1 and L2 values. However, the subjects considered the biased and non-biased head images to be equal in quality ($\mu = 0$). This also occurred for the farm image.

The importance-biased kitchen image was expected to be the same in quality as the base image, but was regarded as inferior by the students ($\mu < 0$). The raw scores of the subjects were also quite consistent in this assessment. It seems that the acceleration of certain contours is deleterious to subjective quality, even though the

L1 and L2 norms indicate otherwise. This probably indicates that contour refinement based on contour information alone is not enough, that is, the approach needs to account for region importance as well. This issue is discussed further in Section 9.2.2.

8.2.2 Supersampling

Table 8.2 lists the subjective viewing results for the flat method supersampling images.

Flat Method Test Images	Sample Mean μ	Sample Standard Deviation σ	Test Statistic Student's t	Confidence Interval $\pm$value	Test H_0 with H_a: $\mu \neq 0$	Test H_0 with H_a: $\mu < 0$	Test H_0 with H_a: $\mu > 0$
Head 4	-12.40	20.52	-2.34	11.36	Reject	Reject	Accept
Head 9	-19.27	21.79	-3.43	12.06	Reject	Reject	Accept
Head 16	-17.93	21.08	-3.30	11.67	Reject	Reject	Accept
Kitchen 4	-6.67	21.84	-1.18	12.10	Accept	Accept	Accept
Kitchen 9	-30.20	22.76	-5.14	12.60	Reject	Reject	Accept
Kitchen 16	-23.07	27.15	-3.29	15.04	Reject	Reject	Accept
Farm 4	0.20	22.42	0.03	12.41	Accept	Accept	Accept
Farm 9	-3.60	15.09	-0.92	8.36	Accept	Accept	Accept
Farm 16	-10.13	21.47	-1.83	11.89	Accept	Reject	Accept
Head Control	8.33	22.13	1.46	12.25	Accept	Accept	Accept
Kitchen Control	16.53	28.34	2.26	15.70	Reject	Accept	Reject
Farm Control	17.00	11.40	5.78	6.31	Reject	Accept	Reject

Table 8.3 Table containing the subjective supersampling results for the flat-rate method with 4, 9 and 16 supersamples per pixel. Flat method control image results are also included. Rejected null hypotheses are shaded in dark grey.

In the case of the flat-rate supersampling subjective tests, the rendering algorithm succeeds if the subjects have a mean quality score of zero ($\mu = 0$). This indicates that the visual importance degraded image is of the same subjective quality as a more expensively rendered flat-rate image.

The flat methodology table exhibits a number of interesting results that have in the majority occurred according to expectations. The objective assessment in Section 5.6 indicated that the flat methodology of supersampling would be the least successful, due to the introduction of aliasing effects from the pixels not being subdivided when the importance value is zero. The above results bear this statement

out, as the null hypothesis is rejected for most of the images ($\mu \neq 0$). However, it should be noted that for the spatially complex scenes like the farm, the subjects regarded the scenes as being of equal quality. This result was expected, as the high count of edges in the scene should mask the aliasing introduced by lack of sampling. This is due to the background noise in the images hiding the lost of quality in the low sampled regions. The head and kitchen scenes contained less edges, therefore any loss of quality was more likely to be noticed against a less noisy background.

An anomaly occurred with the flat sampled kitchen images. The four supersamples per pixel images were considered by the subjects to be of equal quality ($\mu = 0$). This may be explained by the nature of the kitchen image, as it consists of flat coloured regions with sharp edges and little spatial noise. Therefore, the high frequencies would not be effectively sampled by either regime, and so the images would appear to be of a similar quality.

The control experiments were performed to test whether the subjects would consider any images with at least one subdivision per pixel as being of the same visual quality. In two out of the three scenes the null hypothesis was rejected, and the images were considered to be of different quality ($\mu > 0$). If the locations of the supersampling did not matter past one subdivision, then the images should have been seen as the same in quality. This supports the notion that there is a place for the use of visual attention in supersampling methods, as subjects can tell the difference between images sampled with one or greater pixel subdivisions with images containing high frequency components.

On the other hand, the head scene disagreed with these results, with both images being considered the same quality ($\mu = 0$). This results can be explained by observing that the head image contains softer edges, which have frequencies easily antialiased by one or more subdivisions. Therefore, it is reasonable to expect that this image would be considered equal, whereas the other images, with more high frequency content, would be considered to be of differing qualities. These results give support to the notion that the use of visual importance modulated supersampling

is predicated on the spatial frequency content of the image. So it can be concluded that importance biasing is probably more suited to perceptual rendering algorithms, due to their ability to account for the visibility of spatial frequencies within the image. The following Table 8.4 lists the results for the perceptual supersampling tests.

The perceptual images performed as expected. In each case the null hypothesis was accepted, with the subjects being unable to discern the difference between the images. This would be expected of a methodology that accounts for the sensitivities of the human visual system in its image comparison calculations. In addition, no discernable movement occurred with the mean of the samples as the quality threshold was lowered. These results compare favourably with the objective results in Section 5.6, where the objective error norms showed only small differences between importance-biased and unbiased images over the different error thresholds.

Nevertheless, the kitchen images rendered at an error threshold of thirty were considered to be different in quality ($\mu < 0$). This result is hard to explain, as the kitchen threshold values before and afterwards all accepted the null hypothesis. Possibly this is an anomalous result, due to consistency of the other results for the perceptual method. In addition, the objective assessment results in Table 5.8 do not record deviating values for this image, adding evidence to the suggestion that this result is a data recording error.

Perceptual Test Images	Sample Mean μ	Sample Standard Deviation σ	Test Statistic Student's t	Confidence Interval $\pm$value	Test H_0 with H_a: $\mu \neq 0$	Test H_0 with H_a: $\mu < 0$	Test H_0 with H_a: $\mu > 0$
Head 10	0.53	9.06	0.23	5.02	Accept	Accept	Accept
Head 20	2.47	17.07	0.56	9.45	Accept	Accept	Accept
Head 30	-11.67	25.76	-1.75	14.26	Accept	Accept	Accept
Head 40	-1.40	21.91	-0.25	12.14	Accept	Accept	Accept
Head 50	-9.13	17.03	-2.08	9.43	Accept	Reject	Accept
Kitchen 10	3.07	16.38	0.73	9.07	Accept	Accept	Accept
Kitchen 20	3.73	14.08	1.03	7.80	Accept	Accept	Accept
Kitchen 30	-7.67	9.49	-3.13	5.26	Reject	Reject	Accept
Kitchen 40	-5.47	12.86	-1.65	7.12	Accept	Accept	Accept
Kitchen 50	-3.67	8.72	-1.63	4.83	Accept	Accept	Accept
Farm 10	0.07	18.53	0.01	10.26	Accept	Accept	Accept
Farm 20	6.47	14.69	1.71	8.13	Accept	Accept	Accept
Farm 30	0.73	8.36	0.34	4.63	Accept	Accept	Accept
Farm 40	-2.73	12.89	-0.82	7.14	Accept	Accept	Accept
Farm 50	-3.20	11.87	-1.04	6.57	Accept	Accept	Accept

Table 8.4 Table containing the subjective supersampling results for the perceptual method with a 10, 20, 30, 40 and 50 error threshold per region. Note that the null hypothesis was accepted for each of the images. Rejected null hypotheses are shaded in dark grey.

8.2.3 Texture Importance Mapping

Table 8.5 and lists the subjective viewing results for the texture images.

Texture Mapping Images	Sample Mean μ	Sample Standard Deviation σ	Test Statistic Student's t	Confidence Interval $\pm$value	Test H_0 with H_a: $\mu \neq 0$	Test H_0 with H_a: $\mu < 0$	Test H_0 with H_a: $\mu > 0$
Cloth 1537	5.33	22.81	0.91	12.63	Accept	Accept	Accept
Cloth 1025	6.93	15.32	1.75	8.48	Accept	Accept	Accept
Cloth 257	21.33	17.54	4.71	9.71	Reject	Accept	Reject
Kitchen 1537	-0.47	15.48	-0.12	8.57	Accept	Accept	Accept
Kitchen 1025	-1.53	23.52	-0.25	13.03	Accept	Accept	Accept
Kitchen 257	18.47	36.75	1.95	20.35	Accept	Accept	Reject
Garden 1537	4.73	16.49	1.11	9.13	Accept	Accept	Accept
Garden 1025	5.27	13.57	1.50	7.52	Accept	Accept	Accept
Garden 257	31.00	24.01	5.00	13.30	Reject	Accept	Reject
Textured Room	5.87	12.61	1.80	6.98	Accept	Accept	Reject

Table 8.5 Table containing the subjective texture importance mapping results for the perceptual method with a 1537, 1025, 257 pixel square textures error threshold per region. Note that the null hypothesis was rejected only for two of the 257 $\times$ 257 texture images. Rejected null hypotheses are shaded in dark grey.

The above table again compares well with the objective results gained from Section 6.2.1. It was found that two of the images that had textures of size 257 $\times$ 257 pixels

were considered to be of lower quality when importance biasing was introduced. The kitchen image was the only exception to the rule, however its t value was close to the threshold, and it actually rejected the null hypothesis ($\mu > 0$). These figures suggest a possible Type II error caused by visual masking of the blurring in the image.

When the size of the textures was 1025 or 1537 pixels square, then the subjects accepted the images as being of the same quality ($\mu = 0$). The fact that they accepted this for a texture of 1025×1025 pixels in size gives support to having a parameter α in the texture/filter relation described in Section 6.2, which allows the relationship to relax. For these cases, the projected pixels of size less than the maximum size of the support filter may have their texels sampled using an importance-biased method, without losing much visual quality.

Finally, the acceptance of the textured room as being of equivalent quality ($\mu = 0$) gave evidence for the use of adaptive texture sampling in non-trivial scenes. This usefulness is further reinforced by the 10% timesavings gained from modulating the support of the texture filter (refer to Section 6.2.1).

8.3 DISCUSSION

Overall the subjective results provided evidence that the importance-biased images could be discerned as being of similar quality to the unbiased images. The results indicated that it works better for the perceptual supersampling method, compared to the flat-rate supersampling. This was consistent with results obtained from the objective evaluations performed in Chapter 5 and Chapter 6.

In addition, the texture results gave evidence of the efficacy of using importance-biased resampling. In particular, the results indicated that there was some room for relaxing the strict relationship between the size of the filter used and the projected pixel size.

Furthermore, the control experiments supported the application of visual attention to supersampling rendering systems, as the viewers were able to discern differences between images with one subdivision, and images with more than one subdivision. This indicates, for this set of images, the perceptual significance of spatially modulating the pixel supersample rates across the image.

The images in Figure 8.6 are *Fast Fourier Transform* (FFT) frequency-space images derived from the head, kitchen and farm images respectively. Taking an FFT of both the flat rate biased and unbiased scenes, and then plotting the absolute value of the difference between them formed each of the images. They give a qualitative indication of the frequency structure of the differences between the biased and unbiased images.

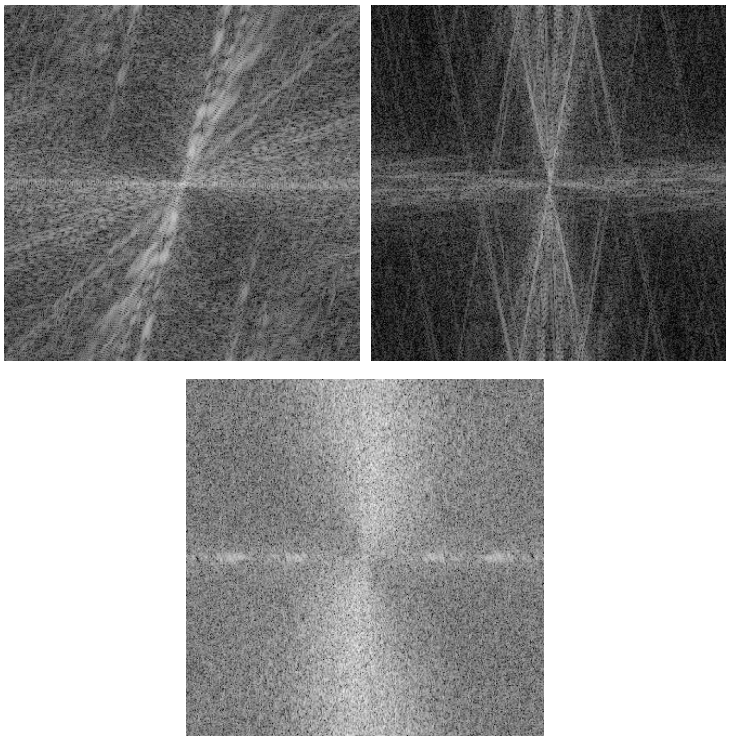


Figure 8.6 FFT diagrams of the differences between the frequency components of the biased and unbiased images– from left to right head, kitchen and farm.

It is interesting to note the wedge shapes that appear in the FFT images. In each case this indicates the removal of coherent edge structure from the images [168]. This concurs with the difference images shown in Section 5.6, as they exhibit a coherent

structure similar to the overall shape of the scene. In addition, the nature of this coherency changes for each image should be noted. The head and kitchen images both show strong location tendency for the differences in frequency space. Inspection of the images reveals lots of straight edges and structures, thus any differences will be strongly aligned in frequency space. The farm FFT image shows less of a linear nature in frequency space, suggesting less of an edge like nature in the image differences. This is concurred by the raw difference image for the farm, which is noisy in character containing less noticeable edges.

The only negative results came from the progressive rendering subjective tests. However, it was expected that the small differences in error ratios would not be significant enough to cause visual improvement, so the overall results for the progressive images was not unexpected. Possible improvements mentioned in Chapter 9 may give better results with subjective testing.

It should be noted that the tests performed in this chapter are preliminary in nature, and should be accepted with the understanding that further experimentation is required to more precisely assess the capabilities of the importance-biased approaches to rendering. These tests would be improved by larger sample sizes and a more diverse image database, to better quantify the differences, and to more effectively characterise the images that perform best or worst with the rendering methods. Nevertheless, these preliminary evaluation results encourage the further development of the importance-biased rendering approaches developed in this thesis.

Discussion and Conclusions

Image synthesis is a computationally intensive task. Many techniques have been developed to improve the efficiency of image synthesis methods, but none have applied a region-based visual importance model to the task of reducing the computations required. Two goals were then set in regards to this problem of image synthesis efficiency. The first was the development of an efficient and improved model of visual importance, designed for the application area of image synthesis. The second was the modification of relevant image synthesis techniques to exploit visual importance information for efficiency gains.

An investigation of research conducted into visual attention in humans yielded information for the development of a region-based fuzzy logic model of visual importance. The model improved on previous fuzzy models by allowing for differences in the visual features, by incorporating contour information and by modelling background variations in feature differences. Furthermore, the model has been applied to progressive image synthesis, allowing it to control the progressive rendering of an image and the supersampling subdivision rates for each pixel. The approach was extended to texture mapping efficiency problems. Further work on the visual importance model incorporated object motion and temporal change effects. These additions improved on other region-based models by separating temporal change from actual object motion, and by introducing motion magnitude and vector differences as an improvement over the absolute motion value calculations performed by other models.

The rendering approach was objectively and subjectively tested for image quality and efficiency gains. The system was able to obtain large savings in image rendering times, without unduly affecting the quality of the rendered images.

Section 9.1 presents a brief overview of the contents and contributions of each of the chapters in this thesis. Section 9.2 describes possible extensions that could be made to both the visual importance model and the progressive rendering approach, to improve both their performance and flexibility. Section 9.3 concludes with potential

application areas for the visual importance model and progressive rendering approach.

9.1 DISCUSSION OF ACHIEVEMENTS

The research conducted for this thesis had two foci. The first was the development of an improved fuzzy logic model of visual importance. The second was the development of progressive rendering techniques that would use this visual importance model to reap efficiency gains.

A literature review of the operation of the HVS from both physiological and psychophysical perspectives was presented in Chapter 2. A brief overview of relevant physiological components of the HVS was presented. Evidence was shown for the existence of physiological constructs sensitive to contrast for colour, luminance, oriented edges and motion. Psychophysical experimentation was reviewed showing evidence for a list of visual features, which through local spatial differences cause regions of the visual field to become visually salient. Psychophysical models of visual attention-based upon these observations were then detailed. Research reviewed also uncovered viewing behaviour patterns indicating that certain regions were regarded repeatedly, and that these regions only constituted a small area of the image. These regions were then shown to contain visual feature differences, which in a bottom-up fashion attracted the attention of the viewer. The chapter concluded with an investigation of bottom-up and top-down factors affecting visual attention. The literature reviewed showed a lack of models completely characterising the relationships and weightings between these features, though some work has produced evidence for a general hierarchy of visual features. General principles were also related to the development of a computational visual importance model.

Chapter 3 presented the development of a new region-based fuzzy model of visual importance, specifically designed for image synthesis applications. A literature review was conducted of computational visual attention models, with the two main approaches being multiresolution and region-based. A region-based model was developed, due to the evidence for object-based viewing of images and the efficiency

constraints of the image synthesis application. A fuzzy logic approach was also chosen, due to the imprecision in the decisions to be made regarding the visual importance of regions within the image. The model was developed into two main modules. The first was a contour importance module, based upon information provided by a progressive rendering system called the DCM. This provided a novel approach to the assessment of the visual importance of contours contained within the progressive rendering of a scene. The second was a region-based visual importance module, containing improvements such as: adaptive membership functions, consistent feature difference processing and the incorporation of contour information.

Progressive rendering techniques were then modified in Chapter 5 to incorporate visual importance features. Existing techniques were identified which did not account for the visual importance of regions being rendered. Progressive rendering techniques were modified to incorporate contour importance, which met with some success. Supersampling techniques were modified to allow for visual importance of the regions being viewed. Objective results showed that the method saved approximately half of the time required to render the scene using normal supersampling techniques, for both flat and perceptual supersampling methods. The approach has also been shown to be efficient and scalable, with both time and space complexity being O(n), with respect to the size of the image.

The techniques developed in the previous chapter have been extended to texture and bump mapping in Chapter 6. Techniques were developed which modified the support of texture resampling filters according to the visual importance of the region in which the samples were being made, reaping savings of 10-20% in rendering times. These techniques were found to be particularly successful where the size of the projected pixel was greater than or equal to the size of the maximum support in pixels of the texture resampling filter. Another technique was developed to improve the efficiency of the progressive rendering of bump mapped polygons. Principles of texture coherence were exploited to enable the rendering algorithm to search texture-space for potential contours. This enabled the progressive rendering algorithm to

detect contours early on in the rendering process, to produce superior quality early images. A subset of images was identified which benefited from this approach.

In Chapter 7, the visual importance model developed in Chapter 3 had motion and temporal change capabilities added. Previous motion importance models were identified, and were shown to only include absolute magnitude estimates of motion. The newly developed model extended this to motion differences in the image, and motion direction estimates. In addition, an abrupt onset model was developed which allowed the model to separate effectively object motion from simple temporal change. The new model of image synthesis motion also differed from others by being region-based, and not block or pixel-based. This facilitated efficiency increases without loss of segmentation accuracy, due to the use of object IDs to control the region segmentation. This use of region segmentation based on object IDs is a novel addition to image synthesis techniques. The algorithm has been theoretically evaluated via complexity analysis, showing the method to be efficient and scalable, with a time and space complexity of O(n), with respect to the size of the image.

Chapter 8 concludes the developmental work carried out in this thesis with an analysis of subjective image quality tests performed with a cohort of viewers. A selection of images generated by the techniques developed within this thesis were shown and compared using a stimulus comparison continuous-scale method. This quantified the amount of perceptual degradation that occurred with images rendered using the new visual importance techniques.

Overall the subjective results provided evidence that the importance-biased images could be discerned as being of similar quality to the unbiased images. The results indicated that it works better for the perceptual supersampling method, compared to the flat-rate supersampling. The texture results also gave evidence of the efficacy of using importance- biased resampling. Furthermore, the control experiments supported the application of visual attention to supersampling rendering systems, as the viewers were able to discern differences between images with one subdivision, and images with more than one subdivision. This indicated the importance of

controlling pixel supersampling via a visual importance model, as the changes in pixel supersampling are visually significant.

9.2 EXTENSIONS TO THE APPROACH

This visual importance approach has been useful in facilitating rendering efficiency gains, and has extended present models of visual attention. However, various areas can be developed further in both the visual importance model and the progressive rendering approaches.

9.2.1 Extensions to the Visual Importance Model

Listed here are a number of extensions proposed for the visual importance model developed as a part of this thesis:

- The membership functions are based upon a model of visual feature differences that captures effects unable to be modelled by absolute values. However, some of the functions are able to capture large magnitudes of differences as visually salient, but they do not model some effects. For example, a small difference in values will stand out against a background of large differences in values uniformly distributed across a screen. Further characterisation of such behaviour would make the model more flexible.
- The use of perceptually uniform colour spaces in the membership functions would give a more accurate estimation of the hue contrast within the image.
- The parameters used in the membership functions are set to arbitrary values, in the absence of empirical psychophysical data. Experiments could be performed in order to find thresholds for the saliency of different features with background variations.
- Experiments could be performed to model more closely the change in shape of adaptive membership functions over a range of background variation values.
- The ad hoc feature combination weights for the contour model should be replaced by empirical values.

- A more global measure of contour importance could be developed to process nearby subdivisions when a contour has been detected. This would improve the detection of curvature and junctions within the image.
- The region and contour importance maps could be integrated, due to the evidence of region-based contour concentration effects upon the search patterns of a viewer. The region-based model, instead of just incorporating a count of the number of contours within the region, could include a measure of the importance of the contours within a segmented region. This combination of contours and other region feature information would allow the model to obtain a more complete measure of the texture information within a segmented region, as the contour importance measures include measures of curvature and concentrations.
- Texture information could be incorporated into the segmentation algorithms, to allow for texture segregation effects, thus obtaining more effective region segmentations.
- Improvements could also be introduced to the integration of different features within the implication method of the fuzzy model. Two major areas needing further work are the weights used for each feature, and the interactions of each feature in the final importance value.

9.2.2 Extensions to Progressive Rendering Approach

The progressive rendering algorithms could be extended in a number of ways, to better utilise the visual importance information available to the rendering system:

- The visual importance system, as well as controlling the principle rays fired into the scene and the texture resampling, could weight a number of other ray-tracing techniques. Recursive ray tracing of reflections and refraction could have the termination criteria modified by the visual importance of image regions. Secondly, the global illumination

algorithms could have their integral error thresholds modulated by the region-based visual importance of the pixel being sampled.

- The progressive contour rendering system could be made region-based by the integration of the region and contour importance systems, to allow important contours within important regions to be refined first.
- The animation techniques developed in Chapter 7 could be extended to more effectively integrate progressive rendering techniques across multiple frames of animations. This would require the integration of the region-based importance model with the contour importance model to enable it to perform region-based progressive rendering. The system could then copy the regions that have not changed over to new frames to accelerate the preview of the final animation.
- To determine the relative importance of the texels within the texture, a visual importance model could preprocess the image. The importance could be stored within an alpha channel for the image file. The texel alpha value could then be read and used to modify the support of the sampling function, as a further adaptive resampling parameter.
- The contour searching technique could be made more efficient by including MIP mapping into the texture-space search. When bump mapping the polygon, the scaled texture to be chosen would have the texture pixel dimensions closest in magnitude to the projected image-space pixel dimensions. This would save on the number of pixels needing to be sampled on the edge of the subdivision in texture-space.

9.3 POTENTIAL APPLICATIONS

The visual importance model developed within the thesis could be applied to the following areas:

- *Compression*-the newly developed model could guide other perceptually-based image compression algorithms which utilise models of early human vision, to place compression artifacts within areas not regarded by the viewer, making the image appear to be of a

higher quality. This would especially suit low bandwidth methods used in video compression applications and progressive transmission applications.

- *Image and video databases*-the algorithms used to search images for particular objects can benefit from restricting their search windows to regions regarded by the viewer. This reduces the computational overhead of having to search the entire image.
- *Machine vision*-due to the relatively low overhead offered by this visual importance system, it is expected that real-time active vision systems could benefit from this approach, enabling the efficient acquisition of targets from the viewing field of the image capture device.
- *Virtual reality*-with the increase of real-time rendering speeds has occurred the concomitant increase in geometric modelling complexity used in virtual reality and visualisation systems. The visual importance level of the rendering region could modulate the level of detail of the mesh. A simplified version of the model presented could enable a real-time system to judiciously apply geometric complexity to visually important regions within the scene to be rendered.
- *Progressive mesh transmission*-applications can be found in the progressive transmission of meshes in low bandwidth applications, for example, in a web-based medical atlas of 3D isosurfaces. These meshes could be preprocessed to ascertain the visual importance of regions within the mesh. These visually important regions can then be sent first, to aid the progressive visual quality of the mesh.

These extensions, and other possible application areas, indicate the area of visual importance has potential for future theoretical and applied research.

Glossary

Adaptive Sampling	Ray tracing algorithm that is sensitive to visual features of a scene, thus modifying the sampling rate to correctly sample high frequency image components.
Antialiasing	Antialiasing is the process of increasing the discrete sampling rate of a signal beyond the Nyquist limit to prevent the generation of unwanted frequencies.
Attentive	Pertaining to visual processed occurring after application of attention.
Bottom-up	Visual effects proceeding from the stimuli alone.
Bump Mapping	Process of applying an image over geometry to modify the lighting function and represent fine surface details.
Cones	Chromatic light sensitive cells in eye.
Defuzzification	Calculation of single value which represents a fuzzy set.
Degree Of Fulfillment	Level of activation in fuzzy logic membership function.
Feature Detectors	Physiological constructs in the HVS designed to respond to particular spatial features.
Feature Integration Theory	Major theory of visual attention, proposes that feature differences attract attention of viewer due to pop-out.
Fovea	High acuity region in centre of visual field.
Fuzzy Logic	Mathematical approach which allows truth value to be any value between 0.0 and 1.0.
Guided Search Model	Major theory of visual attention, proposes top-down improvements to FIT.
Hidden Surface Removal	Rendering approach used to find visible surfaces in a scene.
Implication	Process of fuzzifying, applying rule-base, and defuzzifying in fuzzy logic system.
Importance Map	Spatial map of the visual importance of regions in an image.

Just Noticeable Difference	Threshold value where visual feature difference is perceivable.
L1 Norm	Maximum column sum value of matrix, gives an indication of size of the matrix in an absolute sense–in this thesis used to indicate the absolute amount of error.
L2 Norm	Maximum eigenvalue for matrix, which is the closest matrix form of a Euclidean inner product between functions–in this thesis used to indicate an overall distance measurement between functions.
Lateral Geniculate Nucleus	Physiological construct connecting Optic Nerve to Visual Cortex.
Nyquist Limit	Defines a sampling lower bound at which an algorithm must exceed in order to antialias the frequencies of the signal. Defined as being twice the highest frequency component contained in the signal.
Optic Nerve	Nerve connections from retina to LGN.
Photo-realistic	Pertaining to appearing like a photo of a real object.
Photopic	Bright light levels (daylight).
Pop-out	Phenomenon of region salience caused by feature differences.
Preattentive	Pertaining to visual processes occurring before application of attention.
Progressive Rendering	Rendering which refines an image over a temporal period.
Quadtree	Data structure which represents the recursive decomposition of an image into quadrants.
Ray-tracing	Method of hidden surface removal where vectors are fired into scene to determine visibility.
Resampling	Application of filter to resample an image (often a texture).
Retina	Light sensitive layer at back of eye.

Rods	Achromatic light sensitive cells in the eye, distributed more in the periphery than around the centre of the visual field.
Saccades	Ballistic eye movements to move visual attention from fixation to fixation.
Scotopic	Dull light levels (night).
Student's T Test	Statistical test suited to small sample size hypothesis testing.
Supersampling	Antialiasing performed by taking more than one sample per pixel.
Texel	Texture Pixel–texture image atomic element.
Texture Mapping	Process of applying an image over geometry to represent fine surface details.
Top-down	Visual effects proceding from viewing task factors.
Universe of Discourse	Domain over which fuzzy membership function is defined.
Visual Cortex	Region of brain at rear containing major early visual functions.

References

[1] M. Agrawala, A. Beers, and N. Chaddha, Model-based motion estimation for synthetic animations, in *Proceedings of Third International Conference on Multimedia*, 1995, Lausanne, Switzerland, pp. 477-488.

[2] J. Antes and J. Pentland, Picture context effects on eye movement patterns, in *Eye movements: Cognition and visual perception*, D. Fisher, R. Monty, and J. Senders, Editors, 1976, Lawrence Erlbaum, Hillsdale, USA, pp. 157-179.

[3] ASL, *Asl home page*, www.a-s-l.com, Jan, 2002.

[4] F. Attneave, Some informational aspects of visual perception, *Psychological Review*, 1954, vol. 61(3), pp. 183-193.

[5] M. Baker, *Video compression using a region-based motion model*, Ph.D. Thesis, School of Engineering, University of Ballarat, Ballarat, 1997.

[6] D. Bell, *A region-based progressive image compression technique: Repic*, Ph.D. Thesis, School of Engineering, University of Ballarat, Ballarat, 2000.

[7] R. Berkan and S. Trubatch, *Fuzzy systems design principles, building fuzzy if-then rule bases*, 1997, New York, U.S.A., IEEE Press.

[8] D. Berlyne, The influence of complexity and novelty in visual figures on orienting responses, *Journal of Experimental Psychology*, 1958, vol. 55(3), pp. 289-296.

[9] N. Bichot, J. Schall, and K. Thompson, Visual feature selectivity in frontal eye fields induced by experienced in mature macaques, *Nature*, 1996, vol. 381, pp. 697-699.

[10] I. Biederman, Recognition-by-components: A theory of human image understanding, *Psychological Review*, 1987, vol. 94(2), pp. 115-147.

[11] T. Binford, Inferring surfaces from images, *Artificial Intelligence*, 1981, vol. 17, pp. 205-244.

[12] J. Blinn, Simulation of wrinkled surfaces, in *Proceedings of SIGGRAPH 78*, 1978, pp. 286-292.

[13] J. Blinn and M. Newell, Texture and reflection in computer generated images, *Communications of the ACM*, 1976, vol. 19(10), pp. 542-547.

[14] M. Bolin and G. Meyer, A frequency based ray tracer, in *Proceedings of SIGGRAPH*, 1995, Los Angeles, USA, pp. 409-418.

[15] M. Bolin and G. Meyer, Visual difference metric for realistic image synthesis, in *Proceedings of Human Vision and Electronic Imaging IV*, 1999, San Jose, USA, pp. 106-120.

[16] R. Boyd-Merritt, What's on the frontier of 3d graphics?, *Electronic Engineering Times*, 1997, vol. 958, pp. 85-87.

[17] J. Braun, *Natural scenes upset the visual applecart*, Plymouth Institute of Neuroscience, Devon, 2002.

[18] J. Bresenham, Algorithm for computer control of a digital plotter, *IBM Systems Journal*, 1965, vol. 4(1), pp. 25-30.

[19] R. Brown, B. Pham, E. Aidman, and A. Maeder, Efficient image rendering using a fuzzy logic model of visual attention, in *Proceedings of Advances in Intelligent Systems: Theory and Applications (AISTA)*, 2000, Canberra, Aust., pp. 314-319.

[20] V. Bruce and P. Green, *Visual perception: Physiology, psychology and ecology*, 2nd ed, 1990, Hove, UK, Lawrence Erlbaum.

[21] G. Buswell, *How people look at pictures- a study of the psychology of perception in art*, 1935, Chicago, USA, University of Chicago Press.

[22] T. Callaghan, Interference and dominance in texture segregation: Hue geometric form and line orientation, *Perception and Psychophysics*, 1984, vol. 46(4), pp. 299-311.

[23] T. Callaghan, Interference and dominance in texture segregation: Hue, geometric form, and line orientation, *Perception and Psychophysics*, 1989, vol. 46(4), pp. 299-311.

[24] E. Catmull, *A subdivision algorithm for computer display for curved surfaces*, Ph.D. Thesis, Computer Science Department, University of Utah, Salt Lake City, USA, 1974.

[25] K. Cave and J. Wolfe, Modeling the role of parallel processing in visual search, *Cognitive Psychology*, 1990, vol. 22, pp. 225-271.

[26] CCIR, *Methodology for the subjective assessment of the quality of television pictures*, 1994.

[27] CCIR, *Specification and alignment procedures for setting of brightness and contrast of displays*, 1994.

[28] CCIR, *Specification of a signal for measurement of the contrast ratio of displays*, 1994.

[29] R. Cook, T. Porter, and L. Carpenter, Distributed ray tracing, in *Proceedings of SIGGRAPH 85*, 1984, Minneapolis, USA, pp. 137-45.

[30] S. Daly, The visible difference predictor: An algorithm for the assessment fo image fidelity, in *Digital images and human vision*, A. Watson, Editor, 1993, The MIT Press, Cambridge, USA, pp. 179-206.

[31] S. Daly, Engineering observations from spatiovelocity and spatiotemporal visual models, in *Proceedings of Human Vision and Electronic Imaging III*, 1998, San Jose, USA, pp. 179-206.

[32] J. Daugman, Entropy reduction and decorrelation in visual coding by oriented neural receptive fields, *IEEE Transactions on Biomedical Engineering*, 1989, vol. 36(1), pp. 107-114.

[33] L. De Grandis, *Theory and use of color*, 1986, Englewood Cliffs, USA, Prentice Hall Abrams.

[34] C. De Vleeschouwer, T. Delmot, X. Marichal, and B. Macq, A fuzzy logic system for content-based bit-rate allocation, *Signal Processing. Image Communication*, 1997, vol. 10(1-3), pp. 115-141.

[35] C. De Vleeschouwer, X. Marichal, T. Delmot, and B. Macq, A fuzzy logic system able to detect interesting areas of a video sequence, in *Proceedings of Human Vision and Electronic Imaging II*, 1997, pp. 234-245.

[36] M. Dick, S. Ullman, and D. Sagi, Parallel and serial processes in motion detection, *Science*, 1987, vol. 237, pp. 400-402.

[37] A. Duchowski, 3d wavelet analysis of eye movements, in *Proceedings of Wavelet Applications V*, 1998, Orlando, USA, pp. 435-446.

[38] J. Duncan and G. Humphreys, Visual search and stimulus similarity, *Psychological Review*, 1989, vol. 96, pp. 433-458.

[39] M. D'Zmura, Color in visual search, *Vision Research*, 1991, vol. 31, pp. 951-966.

[40] D. Ebert, ed. *Texturing and modelling: A procedural approach*. 1994, Academic Press Professional, Boston, USA.

[41] J. Enoch, Effect of the size of a complex display on visual search, *Journal of the Optical Society of America*, 1959, vol. 49(3), pp. 208-286.

[42] J. Ewins, M. Waller, M. White, and P. Lister, Implementing an anisotropic texture filter, *Computers & Graphics*, 2000, vol. 24, pp. 253-267.

[43] J. Findlay, The visual stimulus for saccadic eye movements in human observers, *Perception*, 1980, vol. 9, pp. 7-21.

[44] J. Findlay, Saccade target selection during visual search, *Vision Research*, 1997, vol. 37, pp. 617-631.

[45] J. Findlay and I. Gilchrist, Eye guidance and visual search, in *Eye guidance in reading and scene perception*, G. Underwood, Editor, 1998, Elsevier Science, Amsterdam, pp. 295-312.

[46] J. Foley, A. van Dam, S. Feiner, and J. Hughes, *Computer graphics: Principles and practice*, 2nd ed, 1992, Reading, USA, Addison Wesley.

[47] A. Gaddipatti, R. Machiraju, and R. Yagel, Steering image generation with wavelet based perceptual metric, in *Proceedings of EuroGraphics*, 1997, Oxford, UK, pp. 241-251.

[48] A. Gale, Human response to visual stimuli, in *The perception of visual information*, W. Hendee and P. Wells, Editors, 1993, Springer, New York, USA, pp. 115-133.

[49] C. Gerald and P. Wheatley, *Applied numerical analysis*, 1989, Reading, USA, Addison Wesley.

[50] J. Gerrissen, On the network-based emulation of human visual search, *Neural Networks*, 1991, vol. 4, pp. 543-564.

[51] B. Girod, Eye movements and coding of video sequences, in *Proceedings of Visual Communications and Image Processing*, 1988, Cambridge, USA., pp. 398-405.

[52] A. Glassner, Adaptive precision in texture mapping, in *Proceedings of SIGGRAPH 86*, 1986, pp. 297-306.

[53] A. Glenstrup and T. Engell-Nielsen, *Eye controlled media: Present and future state*, Undergraduate Thesis, University of Copenhagen, Copenhagen, Denmark, 1995.

[54] J. Gottlieb, M. Kusonoki, and M. Goldberg, The representation of visual salience in monkey parietal cortex, *Nature*, 1998, vol. 391, pp. 481-484.

[55] B. Guenter and J. Tumblin, Quadrature prefiltering for high quality antialiasing, *ACM Transactions on Graphics*, 1996, vol. 15(4), pp. 332- 353.

[56] B. Guenter, H. Yun, and R. Mersereau, Motion compensated compression of computer animation frames, in *Proceedings of SIGGRAPH 1993*, 1993, Anaheim, USA, pp. 279-288.

[57] B. Guo, Progressive radiance evaluation using directional coherence maps, in *Proceedings of SIGGRAPH*, 1998, Orlando, USA, pp. 255-266.

[58] R. Hayasaka, J. Zhao, and Y. Matsushita, Outstanding objects-oriented color image segmentation using fuzzy logic, in *Proceedings of Multimedia Storage and Archiving Systems II*, 1997, Dallas, USA, pp. 303-314.

[59] R. Hayasaka, J. Zhao, Y. Shimazu, K. Ohta, and Y. Matsushita, Automatic determination of region importance and jpeg codec reflecting human sense, in *Proceedings of International Workshop on Image and Signal Processing*, 1996, Manchester, UK, pp. 231-234.

[60] C. Healey and J. Enns, Large dataset at a glance: Combining textures and colors in scientific visualisation, *IEEE Transactions on Visualization and Computer Graphics*, 1999, vol. 5(2), pp. 145-167.
[61] P. Heckbert, Survey of texture mapping, *IEEE Computer Graphics and Applications*, 1986, vol. 6(11), pp. 56-67.
[62] A. Hillstrom and S. Yantis, Visual motion and attentional capture, *Perception and Psychophysics*, 1994, vol. 55(4), pp. 399-411.
[63] J. Hoffman, Visual attention and eye movements, in *Attention*, H. Pashler, Editor, 1998, University College London Press, London, UK.
[64] D. Hubel, *Eye, brain, and vision*, 1995, New York, USA, Scientific American Library.
[65] D. Hubel and T. Wiesel, Receptive fields and functional achitecture of monkey striate cortex, *Journal of Physiology*, 1968, vol. 195, pp. 215-243.
[66] R. Hunt, *Measuring colour*, 2nd ed, 1991, New York, Ellis Horwood.
[67] T. Ikedo and J. Ma, The truga001: A scalable rendering processor, *IEEE Computer Graphics and Applications*, 1998, vol. 18(2), pp. 59-79.
[68] N. Ito, Y. Shimazu, T. Yokoyama, and Y. Matsushita, Fuzzy logic based non-parametric color image segmentation with optional block processing, in *Proceedings of 23rd ACM Computer Science Conference*, 1995, Nashville, USA, pp. 119-126.
[69] J. Itten, *The art of color - the subjective experience and objective rationale of color*, 1973, New York, USA, Van Nostrand Reinhold Company.
[70] L. Itti and C. Koch, A comparison of feature combination strategies for saliency-based visual attention systems, in *Proceedings of Proceedings of Human Vision and Electronic Imaging IV*, 1999, San Jose, USA, pp. 473-482.
[71] L. Itti and C. Koch, Learning to detect salient objects in natural scenes using visual attention, in *Proceedings of Image Understanding Workshop, (in press: a preprint version of this article is available from http://www.klab.caltech.edu/~itti/attention)*, 1999, pp. N/A.
[72] L. Itti and C. Koch, Computational modelling of visual attention, *Nature Reviews: Neuroscience*, 2001, vol. 2(3), pp. 194-203.
[73] L. Itti and C. Koch, Feature combination strategies for saliency-based visual attention systems, *Electronic Imaging*, 2001, vol. 10(1), pp. 161-169.
[74] L. Itti, C. Koch, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis*, in *IEEE Transaction on Pattern Analysis and Machine Intelligence*. 1998. pp. 1254-1259.
[75] R. Ivry, Asymmetry in visual search for targets defined by differences in movement speed, *Journal of Experimental Psychology: Human Perception and Performance*, 1992, vol. 18(4), pp. 1045-1057.
[76] R. Jacob, Eye tracking in advanced interface design, in *Advanced interface design and virtual environments*, W. Barfield and T. Furness, Editors, 1995.
[77] B. Julesz, Textons, the elements of texture perception, and their interactions, *Nature*, 1981, vol. 290, pp. 91-97.
[78] B. Julesz, A brief outline of the texton theory of human vision, *Trends in NeuroScience*, 1984, vol. 7, pp. 41-45.
[79] B. Julesz, Vision : The early warning system, in *Oxford companion to the mind*, 1988, Oxford University Press, Oxford, UK, pp. 786-793.

[80] B. Julesz and J. Bergen, Textons, the fundamental elements in preattentive vision and perception of textures, *The Bell System Technical Journal*, 1983, vol. 62(6), pp. 1619-1645.

[81] S. Kastner, H. Nothdurft, and I. Pigarov, Neuronal correlates of pop-out in cat striate cortex, *Vision Research*, 1997, vol. 37, pp. 371-376.

[82] C. Kelsey, Detection of visual information, in *The perception of visual information*, W. Hendee and P. Wells, Editors, 1993, Springer, New York, USA.

[83] C. Koch and S. Ullman, Shifts in selective visual attention: Towards the underlying circuitry, *Human Neurobiology*, 1985, vol. 4, pp. 219-227.

[84] A. Kugler, High-performance texture decompression hardware, *The Visual Computer*, 1997, vol. 13, pp. 51-63.

[85] H. Kwak and H. Egeth, Consequences of allocating attention to locations and to other attributes, *Perception and Psychophysics*, 1992, vol. 51(5), pp. 455-464.

[86] W. Leekwijck and E. Kerre, Defuzzification: Criteria and classification, *Fuzzy Sets and Systems*, 1999, vol. 108(2), pp. 159-178.

[87] D. LeGall, Mpeg: A video compression standard for multimedia applications, *Communications of the ACM*, 1991, vol. 34(4), pp. 30-44.

[88] F. Li, R. VanRullen, C. Koch, and P. Perona, Rapid natural scene categorisation in the near absence fo attention, *Proceedings of National Academy of Sciences*, 2002, vol. 99, pp. 9596-9601.

[89] R. Linkser, From basic network principles to neural architecture: Emergence of orientation columns, *Proceedings of National Academy of Sciences*, 1986, vol. 83, pp. 8779-8783.

[90] R. Linkser, From basic network principles to neural architecture: Emergence of orientation-selective cells, *Proceedings of National Academy of Sciences*, 1986, vol. 83, pp. 8390-8394.

[91] R. Linkser, From basic network principles to neural architecture: Emergence of spatial-opponent cells, *Proceedings of National Academy of Sciences*, 1986, vol. 83, pp. 7508-7512.

[92] M. Livingstone and D. Hubel, Segregation of form, color, movement and depth: Anatomy, physiology, and perception, *Science*, 1988, vol. 240, pp. 740-749.

[93] G. Loftus and N. Mackworth, Cognitive determinants fo fixation location during picture viewing, *Journal of Experimental Pychology*, 1978, vol. 4(4), pp. 565-572.

[94] G. Lohse, Consumer eye movement patterns on yellow pages advertising, *Journal of Advertising*, 1997, vol. 26, pp. 61-73.

[95] J. Lubin, A visual discrimination model for imaging system development, in *Vision models for target detection and recognition*, E. Peli, Editor, 1995, World Scientific, New Jersey, USA, pp. 245-283.

[96] N. Mackworth and A. Morandi, The gaze selects informative details within pictures, *Perception and Psychophysics*, 1967, vol. 2(11), pp. 547-552.

[97] A. Maeder, Importance maps for adaptive information reduction in visual system, in *Proceedings of 3rd ANZIIS conference*, 1995, Perth, Aus, pp. 24-29.

[98] A. Maeder, J. Diedrich, and E. Niebur, Limiting human perception for image sequences, in *Proceedings of Human Vision and Electronic Imaging*, 1996, San Jose, USA, pp. 330-337.

[99] A. Maeder and B. Pham, A colour importance measure for colour image analysis, in *Proceedings of IS&T and SID's Color Imaging Conference*, 1993, Scottsdale, Arizona, pp. 233-237.

[100] J. Maillot, L. Carraro, and B. Peroche, Progressive ray tracing, in *Proceedings of 3rd Eurographics Workshop on Rendering*, 1992, Bristol, UK, pp. 9-19.

[101] V. Maljkovic and K. Nakayama, Priming of pop-out: 1. Role of features, *Memory and Cognition*, 1994, vol. 22(6), pp. 657-672.

[102] X. Marichal, T. Delmot, C. De Vleeschouwer, V. Warscotte, and B. Macq, Automatic detection of interest areas of an image or of a sequence of images, in *Proceedings of IEEE International Conference on Image Processing*, 1996, Lausanne, Switz., pp. 371-374.

[103] D. Marr, *Vision*, 1982, San Francisco, USA, W. H. Freeman.

[104] W. Mendenhall, *Introduction to probability and statistics*, 1983, Boston, USA, Prindle, Weber and Schmidt.

[105] G. Meyer, Wavelength selection for synthetic image generation, *Computer Vision, Graphics and Image Processing*, 1998, vol. 41, pp. 57-79.

[106] G. Meyer and A. Liu, Color spatial acuity control of a screen subdivision image synthesis algorithm, in *Proceedings of Human Vision, Visual Processing and Digital Display*, 1992, San Jose, USA, pp. 387-399.

[107] R. Milanese, J.-M. Bost, and T. Pun, A bottom-up attention system for active vision, in *Proceedings of 10th European Conference on Artificial Intelligence*, 1992, Vienna, Austria, pp. 808-810.

[108] R. Milanese, S. Gil, and T. Pun, Attentive mechanisms for dynamic and static scene analysis, *Optical Engineering*, 1995, vol. 34(8), pp. 2428-2434.

[109] R. Milanese, T. Pun, and H. Wechsler, A non-linear integration process for the selection of visual information, in *Intelligent perceptual systems: New directions in computational perception*, R. V., Editor, 1993, Springer, Berlin, Germ, pp. 323-336.

[110] D. Mitchell, Generating antialiased images at low sampling densities, in *Proceedings of SIGGRAPH 87*, 1987, Anaheim, USA, pp. 65-72.

[111] K. Myszkowski, Visible differences predictor: Applications to global illumination problems, in *Proceedings of Ninth Eurographics Workshop on Rendering*, 1998, Vienna, Austria, pp. 223-236.

[112] J. Neider, T. Davis, and M. Woo, *Opengl programming guide*, 1993, Reading, USA, Addiso Wesley.

[113] U. Neisser, *Cognitive psychology*, 1967, New York, USA, Meridith Publishing.

[114] L. Neumann, K. Matkovic, and W. Purgathofer, Perception based color image difference, *Computer Graphics Forum*, 1998, vol. 17(3), pp. 233-241.

[115] E. Niebur and C. Koch, Control of selective visual attention: Modelling the "where" pathway, in *Proceedings of Advances in Neural Information Processing Systems*, 1995, Denver, USA, pp. 802-808.

[116] E. Niebur and C. Koch, *Modeling the where visual pathway*, Caltech-UCSD Institute for Neural Computation, LosAngeles, USA, 1995.

[117] E. Niebur and C. Koch, Computational architectures for attention, in *The attentive brain*, R. Parasuraman, Editor, 1998, MIT Press, Cambridge, USA, pp. 163-186.
[118] E. Niebur, C. Koch, and C. Rosin, An oscillation-based model for the neuronal basis of attention, *Vision Research*, 1993, vol. 33(18), pp. 2789-2802.
[119] J. Nimeroff, N. Badler, and D. Metaxas, *Texture resampling while ray-tracing: Approximating the convolution region using caching*, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, 1995.
[120] D. Notan and L. Stark, Scanpaths in saccadic eye movements while viewing and recognizing patterns, *Vision Research*, 1970, vol. 11, pp. 929-942.
[121] D. Notan and L. Stark, Eye movements and visual perception, *Scientific American*, 1971, vol. 224, pp. 34-43.
[122] H. Nothdurft, The conspicuousness of orientation and motion contrast, *Spatial Vision*, 1993, vol. 7, pp. 341-363.
[123] H. Nothdurft, The role of features in preattentive vision: Comparison of orientation, motion and color cues, *Vision Research*, 1993, vol. 33(14), pp. 1937-1958.
[124] H. Nothdurft, Salience from feature contrast: Additivity across dimensions, *Vision-Research*, 2000, vol. 40(10-12), pp. 1183-1201.
[125] I. Notkin and C. Gotsman, Parallel progressive ray-tracing, *Computer Graphics Forum*, 1997, vol. 16(1), pp. 43-55.
[126] W. Osberger, *Perceptual vision models for picture quality assessment and compression applications*, Ph.D. Thesis, Space Centre for Satellite Navigation, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, Brisbane, 1999.
[127] W. Osberger and A. Maeder, Automatic identification of perceptually important regions in an image using a model of the human visual system, in *Proceedings of ICPR*, 1998, Brisbane, Aus., pp. 701-704.
[128] W. Osberger and A. Rohaly, Automatic detection of regions of interest in complex video sequences, in *Proceedings of Human Vision and Electronic Imaging VI*, 2001, San Jose, USA, pp. 361-372.
[129] J. Painter and K. Sloan, Antialiased ray tracing by adaptive progressive refinement, *Computer Graphics*, 1989, vol. 23, pp. 281-288.
[130] F. Pighin, D. Lischinski, and D. Salesin, Progressive previewing of ray-traced images using image-plane discontinuity meshing, in *Proceedings of Eurographics Workshop: Rendering Techniques*, 1997, St. Etienne, France, pp. 115-126.
[131] Pixar, *Pixar company website*, www.pixar.com, Sept, 2001.
[132] M. Posner, Orienting of attention, *The Quarterly Journal of Experimental Psychology*, 1980, vol. 32, pp. 3-25.
[133] M. Posner and M. Raichle, *Images of mind*, 1997, New York, USA, Scientific American Library.
[134] J. Prikryl and W. Pugathofer, *Overview of perceptually-driven radiosity methods*, Insititute of Computer Graphics, Vienna University of Technology, Vienna, 1999.
[135] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard, Modelling saccadic targeting in visual search, in *Advances in neural information processing*

systems, D. Touretzky, M. Mozer, and M. Hasselmo, Editors, 1996, MIT Press, Cambridge, USA, pp. 830-836.

[136] R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard, *Eye movements in visual cognition*, National Resource Laboratory for the Study of Brain and Behavior, Computer Science Dept., University of Rochester, Rochester, USA, 1997.

[137] T. Reed, Local frequency representations for image sequence processing and coding, in *Digital images and human vision*, A. Watson, Editor, 1993, The MIT Press, Cambridge, USA, pp. 3-12.

[138] D. Robinson, The mechanics of human smooth pursuit eye movements, *Journal of Physiology*, 1965, vol. 180, pp. 569-591.

[139] B. Rogowitz, D. Rabenhorst, J. Gerth, and E. Kalin, Visual cues for data mining, in *Proceedings of Human Vision and Electronic Imaging*, 1996, San Jose, USA, pp. 275-300.

[140] J. Russ, *The image processing handbook*, 1992, Boca Raton, USA, CRC Press.

[141] H. Samet, *The design and analysis of spatial data structures*, 1990, Reading, USA, Addison Wesley.

[142] H. Schwarz, *Numerical analysis: A comprehensive introduction*, 1989, Chichester, UK, John Wiley & Sons.

[143] J. Senders, Speculations and notions, in *Eye movements and psychological processes*, M. R. and S. J., Editors, 1976, Lawrence Erlbaum, Hillsdale, USA.

[144] J. Senders, Distribution of visual attention in static and dynamic displays, in *Proceedings of Human Vision and Electronic Imaging II*, 1997, San Jose, USA, pp. 186-194.

[145] F. Sharp and R. Philips, Physiological optics, in *The perception of visual information*, W. Hendee and P. Wells, Editors, 1993, Springer, New York, USA, pp. 1-30.

[146] D. Sheena and J. Borah, Compensation for some second order effects to improve eye position measurements, in *Eye movements: Cognition and visual perception*, D. Fisher, R. Monty, and J. Senders, Editors, 1981, Laurence Erlbaum, New Jersey, pp. 257-268.

[147] M. Shepherd, J. Findlay, and R. Hockey, The relationship between eye movements and spatial attention, *The Quarterly Journal of Experimental Psychology*, 1986, vol. 38A, pp. 475-491.

[148] L. Stark, Top-down vision in humans and robots, in *Proceedings of Human Vision, Visual Processing, and Digital Display IV*, 1993, San Jose, USA, pp. 613-621.

[149] L. Stark and S. Ellis, Scanpaths revisited: Cognitive models direct active looking, in *Eye movements: Cognition and visual perception*, D. Fisher, R. Monty, and J. Senders, Editors, 1981, Lawrence Erlbaum, New Jersey, USA.

[150] L. Stark, K. Ezumi, T. Nguyen, R. Paul, G. Tharp, and H. Yamashita, Visual search in virtual environments, in *Proceedings of Human Vision, Visual Processing and Digital Display II*, 1992, San Jose, USA, pp. 577-589.

[151] L. Stelmach, W. Tam, and P. Hearty, Static and dynamic spatial resolution in image coding: An investigation of eye movements, in *Proceedings of Human Vision, Visual Processing and Digital Display II*, 1991, San Jose, USA, pp. 147-152.

[152] T. Syeda-Mahmood, Detecting perceptually salient texture regions in images, *Computer Vision and Image Understanding*, 1999, vol. 76(1), pp. 93-108.
[153] M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy, Integration of visual and linguistic information in spoken language comprehension, *Science*, 1995, vol. 268, pp. 1632-1634.
[154] J. Theeuwes, Visual selective attention: A theoretical analysis, *Acta Psychologica*, 1993, vol. 83(2), pp. 93-154.
[155] A. Treisman, Features and objects: The fourteenth bartlett memorial lecture, *The Quarterly Journal of Experimental Psychology*, 1988, vol. 40A, pp. 201-237.
[156] A. Treisman, The perception of features and objects, in *Attention: Selection, awareness, and control*, A. Baddeley and L. Weiskrantz, Editors, 1993, Clarendon Press, Oxford, UK.
[157] A. Treisman, Modularity and attention: Is the binding problem real?, in *Visual selective attention*, C. Bundesen and H. Shibuya, Editors, 1995, Hillsdale, USA, Lawrence Erlbaum.
[158] A. Treisman and G. Gelade, A feature-integration theory of attention, *Cognitive Psychology*, 1980, vol. 12, pp. 97-136.
[159] A. Treisman and S. Sato, Conjunction search revisited, *Journal of Experimental Psychology: Human Perception and Performance*, 1990, vol. 16, pp. 459-478.
[160] J. Tsotsos, An inhibitory beam for attentional selection, in *Proceedings of York Conference on Spatial Vision in Humans and Robots*, 1991, Toronto, Can, pp. 313-331.
[161] N. Tsumura, C. Endo, H. Haneshi, and Y. Miyake, Image compression and decompression based on gazing area, in *Proceedings of Human Vision and Electronic Imaging*, 1996, San Jose, USA, pp. 361-367.
[162] S. Upstill, *The renderman companion*, 1992, Reading, USA, Addison Wesley, Reading.
[163] M. Usher and E. Niebur, Modeling the temporal dynamics of it neurons in visual search: A mechanism for top-down selective attention, *Journal of Cognitive Neuroscience*, 1996, vol. 8(4), pp. 311-327.
[164] F. Verstraten, P. Cavanagh, and A. Labianca, Limits of attentive tracking reveal temporal properties fo attention, *Vision Research*, 2000, vol. 40, pp. 3651-3664.
[165] D. Wallach, S. Kunapalli, and M. Cohen, Accelerated mpeg compression of dynamic polygonal scenes, in *Proceedings of SIGGRAPH 1994*, 1994, Orlando, USA, pp. 193-196.
[166] B. Wandell, *Foundations of human vision*, 1st ed, 1995, Sunderland, USA, Sinauer.
[167] G. Ward, F. Rubenstein, and R. Clear, A ray tracing solution for diffuse interreflection, in *Proceedings of SIGGRAPH*, 1988, pp. 85-92.
[168] A. Watt and F. Policarpo, *The computer image*, 1998, New York, ACM Press.
[169] J. Wernecke, *The inventor mentor: Programming object-oriented 3d graphics with open inventor, release 2*, 1994, Reading, USA, Addison Wesley.
[170] C. Westelius, *Dissertation no. 379: Focus of attention and gaze control for robot vision*, Ph.D. Thesis, Linkoping Studies in Science and Technology, University of Linkoping, Linkoping, Sweden, 1995.

[171] G. Westheimer, Eye movement responses to horizontally moving visual stimuli, *Arch. Opthamology*, 1954, vol. 52, pp. 932-941.
[172] T. Whitted, The hacker's guide to making pretty pictures, in *Proceedings of SIGGRAPH 85, Image Rendering Tricks, Course Notes 12*, 1985, New York, USA, pp.
[173] L. Williams, Pyramidal parametrics, in *Proceedings of SIGGRAPH 83*, 1983, pp. 1-11.
[174] J. Wise, *Eye movements while viewing commercial ntsc format television*, SMPTE psychophysics subcommittee white paper, 1984.
[175] G. Wolberg, *Digital image warping*, 1990, Los Alamitos, USA, IEEE Computer Society Press.
[176] J. Wolfe, Guided search 2.0 a revised model of visual search, *Psychonomic Bulletin & Review*, 1994, vol. 1, pp. 202-238.
[177] J. Wolfe, Visual search: A review, in *Attention*, H. Pashler, Editor, 1996, University College London Press, London, UK, pp. 13-73.
[178] J. Wolfe, The deployment of visual attention: Two surprises, in *Search and target acquisition*, Nato-Rto, Editor, 2000, NATO-RTO, Utrecht, Netherlands, pp. 20.21-20.11.
[179] J. Wolfe and K. Cave, Deploying visual attention: The guided search model, in *Ai and the eye*, A. Blake and T. Troscianko, Editors, 1990, Wiley, New York, USA, pp. 79-103.
[180] J. Wolfe and G. Gancarz, Guided search 3.0 a model of visual search catches up with jay enoch 40 years later, in *Basic and clinical applications of vision science*, V. Lakshiminarayanan, Editor, 1996, Kluwer Academic, Dordrecht, Netherlands.
[181] J. Wolfe, P. O'Neill, and S. Bennett, Why are there eccentricity effects in visual search? Visual and attentional hypotheses, *Perception and Psychophysics*, 1998, vol. 60(1), pp. 140-156.
[182] R. Yager and D. Filev, *Essentials of fuzzy modeling and control*, 1st ed, 1994, New York, USA, John Wiley and Sons.
[183] S. Yantis and J. Jonides, Abrupt visual onsets and selective attention: Voluntary versus automatic allocation, *Journal of Experimental Psychology: Human Perception and Performance*, 1990, vol. 16, pp. 121-134.
[184] A. Yarbus, *Eye movements and vision*. 1967, Plenum Press: New York, USA.
[185] H. Yee, S. Pattanaik, and D. Greenberg, Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments, *ACM Transactions on Graphics*, 2001, vol. 20(1), pp. 47-54.
[186] Y. Yee, *Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic scenes*, Masters Thesis, Program of Computer Graphics, Cornell, Ithaca, USA, 2000.
[187] H. Yun, B. Guenter, and R. Mersereau, Lossless compression of computer generated animation frames, *ACM Transactions on Graphics*, 1997, vol. 16(4), pp. 359-396.
[188] H. Zabrodsky and S. Peleg, Attentive transmission, *Journal of Visual Communications and Image Representation*, 1990, vol. 1, pp. 189-198.
[189] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, and Y. Matsushita, An outstandingness oriented image segmentation and its application, in *Proceedings of International Symposium On Signal Processing and its Applications, ISSPA*, 1996, Gold Coast, Australia, pp. 45-48.

[190] U. Zimmer, Connectionist decision systems for a visual search problem, in *Proceedings of IIZUKA*, 1994, Fukuoka, Japan, pp. 1-3.