INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

FINDING SALIENT OBJECTS IN AN IMAGE

Anthony Hang Fai Lau

Department of Electrical Engineering
McGill University

May 2000

A Thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of
Master of Engineering

Abstract

Many computer vision applications, such as object recognition, active vision, and content based image retrieval (CBIR) could be made both more efficient and effective if the objects of interest could be segmented from the background. This thesis discusses the development and implementation of a complete unsupervised object-based attention system for locating salient objects in an image.

The major components of this system are the segmentation and the attention process. Considerable research has been done in these two areas, but unfortunately, there is still not a single method that can be applied reliably under all situations. We have analysed the attention model proposed by Osberger and have found that their method fails to identify some important regions that are salient to humans. Modifications to this model are proposed to correct some of these problems. For the segmentation process, one important aspect is the measurement of the quality of a particular segmentation, since the attention process depends solely on the segmentation output. In particular, three different cluster validity measures are considered: a simple threshold-based index, a non-parameter index, and the modified Hubert index. From the experimental results, the simple threshold-based index is shown to outperform the other indices on most test images. We believe that the success of the threshold-based index is largely related to the incorporation of human preference in the selection of the threshold parameter.

Résumé

De nombreuses applications en vision artificielle telles que la vision active et l'indexage d'images basé sur le contenu pourraient être rendues plus efficaces si les objets d'intérêt pouvaient segmentés du fond de l'image. Cette thèse discute du développement et de l'implémentation d'un système d'attention non-supervisé basé sur des objets pour localiser des objets saillants dans une image.

Les composantes majeures de ce système sont la segmentation et le mécanimsme d'attention. Bien que que ces deux sujets aient été l'objet de nombreuses recherches. il n'existe toujours à ce jour pas de méthode fiable qui puisse être appliquée dans toutes les situations. Nous avons analysé le modèle d'attention propoé par Osberger et nous avons trouvé qu'elle ne réussi pas à identifier quelques unes des régions saillantes évidentes pour des humains. Des modifications à ce modèle sont proposées pour corriger certains de ces problèmes. Un des aspects importants pour la segmentation est la mesure de la qualité d'une méthode en particulier puisque le processus d'attention repose uniquement sur le résultat de la segmentation. Plus particulièrement. trois différentes méthodes de mesure de validité sont considérées: un index déterminé par un seuillage simple. un index non-paramétrique et une version modifiée de l'index d'Hubert. D'après les résultats expérimentaux. l' index déterminé par un seuillage simple surpasse les autres méthodes pour la plupart des images testées. Nous croyons que le succès de l'index déterminé par un seuillage simple est largement lié a l'incorporation de préfrences humaines dans la sélection du seuil utilisé.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. M.D. Levine, for his enthusiastic guidance and support. He is always available when needed and is willing to discuss with his students any difficulties encountered during the research. I must also thank Gilbert Soucy for translating the abstract to French, all the people at CIM for providing a good working atmosphere, and my family for their unfailing support and encouragement throughout the period of this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

This thesis discusses the development and implementation of a complete object-based attention system for locating salient *objects* in an image. In this chapter, the need and motivation for this approach is presented. An overview of the thesis follows, including a brief outline of each of the remaining chapters.

1.1. The Need for Object-based Attention

Many computer vision applications, such as object recognition [**60, 39**], active vision [**17**], and content based image retrieval (CBIR) [**2, 37**] can be made both more efficient and effective if the objects of interest can be segmented from the background. In the case of object recognition, especially in a complex scene, the recognition process can be more efficient and robust if even a rough estimation of the location and size of the salient objects can be obtained [**39**]. A ranking of perceptual saliency or closeness to the target model is then required to determine which region should be processed first. As a result, expensive computational resources can be focused mainly on those regions that are worthy of more detailed examination.

This kind of attention system can also be applied to CBIR to improve the retrieval accuracy. The first generation of image retrieval systems relied solely on keywords entered by a human when the image was entered into the database. The strength of this approach comes from the high accuracy of the identification of major objects present in each image and its image type (such as a scenic picture or art work). For example, if one wants to retrieve images that contain a polar bear, he just needs to type in the keyword "polar bear" to retrieve all images that have at least one polar

bear with 100% accuracy. However, there are several major drawbacks that constrain its applicability and usefulness. These disadvantages include the requirement for manual annotation and the inherent limitation of words in expressing abstract ideas. For instance, it is very difficult to describe precisely the content of some images, such as modern paintings, with a limited number of keywords. As a result, image retrieval based on image content has been proposed as a new approach to organise the huge and ever-expanding image databases (e.g., online museums and databases of medical images). Besides, the classical image retrieval system can be further improved by enabling the system to mimic the identification of salient objects in an image as in the keyword-based system.

1.2. Motivation

Object-based CBIR has been investigated by several researchers [2] [13][115]. In these approaches, although features of local regions instead of global properties are used, each region is still treated with equal importance. As a result, an irrelevant image can be retrieved just because it contains a background that is visually similar to the query image. Hence, it is desirable to have a complete and fully automatic attention system for segmenting and locating salient objects in an image. Methods for determining the saliency of regions have been investigated by Osberger and Maeder [86]. However, only initial results have been presented and no in-depth analysis of their method has been carried out. As stated in [86], the performance of an object-based attention system depends largely on the quality of the segmentation results. Hence, it is desirable to analyse their method and to select an image segmentation technique best suited to the attention algorithm.

1.3. An Overview of the Approach

Each process involved in the detection of salient objects in an image will be discussed in this thesis. The overall system is summarised in a block diagram in Figure 1.1. The system input is a single colour image. A set of biologically motivated feature maps are extracted from the image and then used in the image segmentation process. Before the region information of the "objects" can be generated, the definition of "object" must be defined precisely. To be of general use, no context-dependent information is assumed and an object is defined simply as a coherent and homogenous

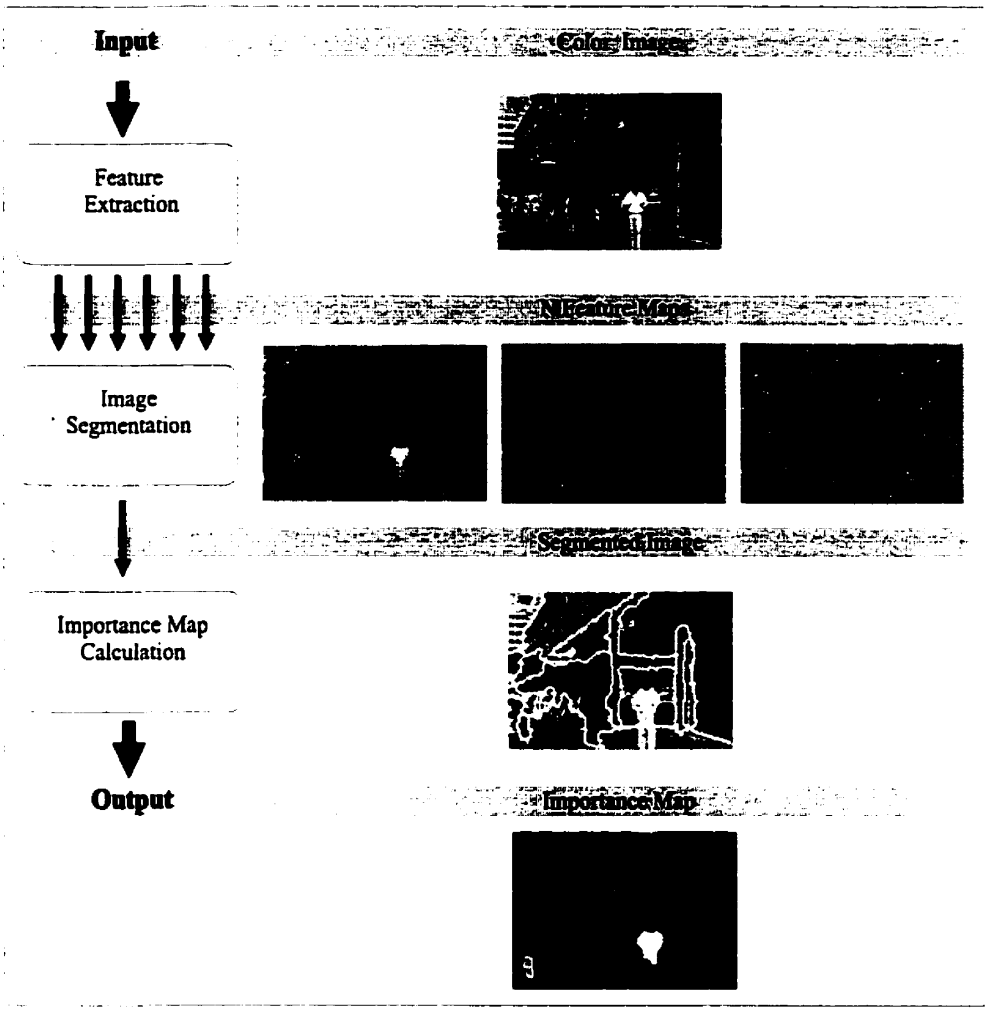


FIGURE 1.1. System Block Diagram. The method consists of three computational processes shown in the left-hand column. The data transferred between each process is indicated in the shaded strips. Examples of the input, intermediate data, and final output are shown in the right-hand column

region. If higher-level, top-down information is known a priori, this information can be used to group the regions into a logical entity that resembles the original physical object. The final stage involves the computation of the Importance Map based on a number of factors, such as contrast and eccentricity, that have been able to draw attention. This importance map represents the perceived saliency of the regions.

1.4. Organisation of the Thesis

In Chapter 2, a review of the literature of the biological basis of perceptual grouping and attention will be presented. The current state of machine vision simulating these two tasks will also be described. Evidence from psychophysical experiments shows that objects can exist preattentively and can affect covert attention. However, not much research has been focused on developing an object-based model of attention. Hence, it is desirable to investigate this topic in detail.

Chapter 3 begins with a discussion of the only object-based attention model that has been developed for computer vision applications [**86**]. In this model, five factors are identified and formulated mathematically. Situations where these factors fail and solutions to these problems will be discussed in this chapter.

In Chapter 4, the details of selecting a particular representation scheme for each feature are discussed. Transformations on the feature spaces to improve the perceptual uniformity will also be presented.

In Chapter 5, the first section reviews the major image segmentation techniques. Reasons for selecting a particular image segmentation method and some implementation issues will be described in the remainder of this Chapter.

Finally, Chapter 6 presents a variety of results of the system applied to real world images. This includes an examination of the selection of various model parameters and the feasibility of using this system as a pre-processor to a face-finding system.

1.5. Contributions

The major contributions of this thesis are.

- Lots of work has been done on image segmentation. However, there is still no "off-the-shelf" solution that can be applied to all types of images. One of the major problems is the lack of a good measure of the quality of a particular segmentation. In this thesis, three different measures are considered and we find that a simple threshold-based measure with a manually selected threshold give consistently better results than other more complex, statistics-based measures.
- Parameters are a significant aspect of any mathematical formulation of an algorithm. Some parameters can be obtained through theoretical arguments.

However, the optimum values for some other parameters depend on subjective judgements, such as the importance or saliency of different objects in a scene. To reduce the bias on any particular image type or subjective opinion, systematic and extensive experimentation has been performed to find suitable parameter values.

- The complete system for locating salient objects is implemented in Microsoft Visual C++ with Microsoft Fundation Class (MFC) for a stand-alone application. Appendix A provides a brief description of the system with images of the graphical user interface.

CHAPTER 2

Literature Review

David Marr has written [**73**]:

> "What does it mean, to see? The plain man's answer (and Aristotle's, too) would be, to know what is where by looking. In other words, vision is the *process* of discovering from images what is present in the world, and where it is."

Visual perception is a natural and native ability of humans and animals. Using an abundant amount of information about colour and form, we can sense the environment in its original 3-dimensions, or 4-dimensions if time is included. Not only can we see the 3-dimensional world, but we can also recognise the objects and understand their positional, structural, and contextual relationships. In nature, the ability to detect and recognise objects effectively and efficiently is vital to survival. Animals must be able to distinguish their food from other less edible alternatives. They must also be able to detect camouflaged or occluded predators. The seemingly straightforward and effortless task of object detection and recognition for both humans and animals is extremely difficult to simulate in the computer. One reason for this difficulty is the incomplete and unclear definition of *object* in the field of computer vision. If we want a computer to recognise an object, the definition of *object* must be precise and without ambiguity. However, even for humans, there does not exist a fixed and universally held definition of object. Both Ullman [**134**] and Marr [**72**] raise the question about the goal of segmentation, particularly in a bottom-up manner. Marr asks: "What, for example, is an object, and what makes it so special that it should be recoverable as a region in an image? Is a nose an object? Is a head one?..." They both conclude that it is extremely difficult, if not impossible, either to formulate

what should be recovered as a region from an image or to separate complete objects, such as a car or a house, from a complex scene. Although the problem of unclear definition of object or goal of segmentation seems to be unsolvable, the task of object detection and recognition is performed smoothly and accurately within the human visual system, without any sign of ambiguity. In this chapter, both psychophysical and physiological aspects of the mechanisms used by humans in perceptual grouping and attention will be reviewed. An overview of the current state of machine vision will then be presented.

2.1. Perceptual Grouping

In the literature, perceptual grouping is sometimes described in other terms, such as segmentation, clustering, association, and figure-ground separation, depending on the point-of-view from which this problem is viewed. In [**66**], Lowe states that "Perceptual organisation refers to a basic capability of the human visual system to derive relevant groupings and structures from an image without prior knowledge of its contents". Similarly, Sarkar [**106**] defines the term perceptual grouping or perceptual organisation as the ability to impose structural organisation on sensory data, so as to group sensory primitives arising from a common underlying cause. If a person is asked to segment an image into different regions, the answer may not be unique and varies from person to person. For the image in Figure 2.1, one may segment the image into two distinct groups: the baby and the background. Another possible segmentation could be the baby, the beach, the water, and the sky. However, one can also further segment the baby's head from the body. This variation in complexity may arise because of different general grouping systems. However, it is more likely due to a difference in the level of abstraction rather than the overall system. Such a hierarchical framework for representing objects has been used in many computer vision systems for deriving higher level concepts of objects from lower level primitives [**73**] [**95**] [**78**] [**42**] [**107**]. In the first chapter of Marr's book "Vision" [**73**], he described four levels of abstraction for deriving shape information from images. The lowest level is the image itself and the primitive at this level is the intensity value (either in grey scale or colour) at each pixel in the image. The second level is the primal sketch. At this level, a set of low level features is extracted from the intensity or colour map of the first level. The primitives at this stage are zero-crossings, blobs, terminations

FIGURE 2.1. A sample picture of a baby girl

and discontinuities. edge segments. virtual lines. groups. curvilinear organisation and boundaries. The third level of abstraction is the 2-1/2 D sketch. The purpose of this stage is to organise and represent the primal sketch in a viewer-centred coordinate frame with a rough description in terms of surfaces. The primitives now become local surface orientation. distance from the viewer. discontinuities in depth and surface orientation. The highest level of abstraction is the actual 3-D model representation. The purpose of this stage is to derive and represent the objects in an object-centred coordinate frame so that recognition can be achieved with viewpoint invariance. The primitives are 3-D shape models with the corresponding surface properties and their spatial organisation. This representational framework is mainly object-centred. On the other hand. viewer-based representation has also been proposed for explaining how information is stored in the human visual system [**3**]. In a viewer-based framework. different views of the object rather than its 3-D model are extracted and stored. The advantage of this approach is that it is not necessary to build an explicit model of every object intended to be recognised.

Although any object can be described by different levels of abstraction as suggested by Marr. it is still not clear how the grouping process works or how it can terminate. The first theory for explaining perceptual grouping is the Gestalt Theory proposed by Wertheimer in 1912 [**140**]. This theory proposes that the geometrical

relationships that humans use in perceptual grouping can be categorised as follows [**141**]:

- Similarity: Similar elements are grouped together.
- Proximity: Elements that are close together tend to be grouped together.
- Continuation: Elements that lie along a common line or smooth curve are grouped together.
- Symmetry: Symmetric curves are grouped together.
- Closure: Curves are connected to enclose regions.
- Familiarity: Elements are grouped into familiar structures.

This theory implies that there is a tendency for humans to seek the most unambiguous and simple interpretation of the world. This principle of simplicity of form is similar to the law of least action or the minimum principle discovered by ancient Greek geometers. This theory has fostered many other theories and continues to exert significant influence on the psychology of perception. Although introduced at the beginning of the 20^{th} century, these six principles are still valid and are the basis of most grouping methods. It should be noted that these rules are not exclusive, and groupings may be formed using combinations of subsets of these relationships. Unfortunately, the algorithmic implementation of these rules is very difficult because they have been obtained through observation and they often conflict, even for simple stimuli, as shown by Lowe [**67**]. Moreover, the theory is usually demonstrated using simple visual patterns, which may not always occur in the real world, the world of unreliable, uncertain stimuli. Therefore, only a relatively few aspects of the Gestalt theory have been incorporated into computer vision systems, such as similarity, proximity, and continuity [**106**]. When these principles are used together, higher level meta-rules are employed either explicitly or implicitly, to guide their application.

Since perceptual grouping can be defined at many different levels of abstraction, a variety of specific goals has been selected and pursued by researchers. Numerous interesting computational approaches have been proposed over a wide range of abstraction levels. A classificatory structure in perceptual organisation is proposed by Sarkar and Boyer [**106**] to organise these algorithms and as a standard nomenclature with which to discuss existing and future research. In their classification scheme, algorithms are classified based on two characteristics. The first is the type of feature being organised or the level of abstraction : signal level, primitive level, structural level, and assembly level. The second is the dimensions over which the organisations

are sought : 2-D, 3-D, 2-D plus time and 3-D plus time. A grey scale image is in 2-D while a range image is in 3-D. With this classification scheme, since the total number of categories is just 16, some categories may contain more than one algorithm. To further differentiate these algorithms, additional classification schemes have been suggested by Sarkar and Boyer, such as the computational technique. This classification structure is useful for comparing and visualising the similarities and differences between algorithms and thus will be used here. However, another possible classification scheme can be based on whether top-level knowledge of objects is utilised or not.

Since the emphasis of this thesis is on 2-D images, the review will be focused on those algorithms designed for grey-scale or colour images. Readers are referred to Sarkar's paper [**106**] for methods involving higher dimensions.

2.1.1. Signal Level

This level involves the lowest and most basic form of organisation, and the input to the algorithms are local point properties.

Zahn [**148**] has proposed the use of graphs to extract and detect Gestalt clusters in dot-clustering problems. He uses a family of graph-theoretical techniques based on the minimal spanning tree to segment several kinds of dot clusters. A minimal spanning tree retains both the information of the local neighbourhood and the overall structures of the clusters and thus is suitable for data clustering problems. Zucker [**151**] approached the problem of dot clustering with a probabilistic model for clusters. Each pixel is classified according to one of three labels: edge, interior, and noise with the corresponding probability. A relaxation process is used to relabel the pixels iteratively until no more pixels are relabelled. A similar method is used by Spann [**116**] for figure-ground separation. He approached the problem using global optimisation of a function representing the local error fit of an assumed model describing the variation of the luminance over the local regions in the image. To minimise the effect of variance in scale and noise, a multi-scalar pyramid was used with interconnections between the layers. The optimisation is carried out using simulated annealing. The use of a model and global optimisation removes the necessity of selecting parameters and thresholds. However, choosing a suitable model may even be more difficult than setting thresholds or parameters depending on the problem domain.

Image segmentation also belongs to this category. In a review paper [**87**] published in 1993, 173 papers are quoted in the references. Since then, more than ten

new algorithms have been published [**26, 111, 5, 61, 64, 110, 20, 90**]. The major contributions of these methods are twofold. The first is a better definition of coherent regions or boundaries, especially for complex scenes . For example, Deng et al. [**26**] propose a new measure J for region uniformity that evaluates the spatial distribution of colour in an image. To reduce the overall complexity and to improve the stability of the distribution estimation, the image is pre-quantised to reduce the number of distinct colours. An interesting aspect of this measure is that both texture and colour information are preserved and encoded in the distribution. Shi and Malik [**111**] propose a new feature distance derived to reduce the instability of a similarity matrix. Feature distance was previously defined either arbitrarily, such as equal weighting on all features, or from the statistics in the test image set. Since this new distance is based solely on the image data, there is no need to pre-define the significance of each feature. For measuring texture, a set of filters is usually applied to the image. Belongie and Malik [**5**] find that the filter responses inside textured regions are generally spatially inhomogeneous. Thus, they have developed a new method for reducing these inhomogeneities by a method called area completion. The main idea behind this method is to increase the similarities between pixels if they are close to each other in the spatial domain and have neighbours that are close in the feature domain. As a result, a non-uniform region having a repetitive pattern of features can still be classified as one region. Lambert and Carron [**61**] define a new colour space symbolically, where hue is explicitly defined and processed according to its relevance to chroma. A fuzzy classifier is used to classify the relevance of hue based on the following rules: 1. Hue is not relevant and cannot be utilised in segmentation for small chroma values. 2. Hue is approximately as relevant as chroma and intensity for medium chroma values. 3. Hue is very relevant for large chroma values. Leung and Malik [**64**] define a new definition of texture as repeated scene elements. To be invariant to scale and perspective, affine transformation is used when measuring the similarity between different regions.

The second contribution of recently proposed segmentation algorithms is a more effective or efficient way of region merging and clustering in feature space. Shi and Malik [**110**] propose a novel approach to solve the perceptual grouping problem by treating image segmentation as a graph partitioning problem. A global criterion, normalised cut, is proposed by them for segmenting the graph. Comaniciu and Meer [**20**] propose a general technique for image segmentation based on feature density.

A technique called *mean shift algorithm* is used for estimating density gradients to locate the position of local maxima. The number of local maxima or modes is determined automatically by the algorithm: however, the number of modes depends on the width of the density estimation kernel. Park et al. [**90**] suggest using mathematical morphology to cluster and classify pixels in the feature domain. First, a colour histogram is generated and smoothed with a 3-D Gaussian kernel. Next, mathematical morphology, dilation and erosion, is applied to the histogram to remove the outliners and to separate distinct clusters. Carson et al. [**13**] propose using an Expectation-Maximisation (EM) algorithm to perform segmentation based on image features. The distribution function of each cluster is presumed to be Gaussian and the EM algorithm is used to determine the maximum likelihood parameters of a mixture of K Gaussians. This method is repeated for different values of K and the number of clusters is determined by finding the best fit of the estimated parameters to the data.

2.1.2. Primitive Level

This level involves the intermediate level of organisation with edges or curves as input.

Hérault and Horaud [**47**] attack the figure-ground discrimination problem from a combinatorial optimisation perspective. They define the problem as separating a salient curve from noise and make explicit the definition of shape (or figure) based on cocircularity, smoothness, proximity, and contrast in terms of mathematical formulas. Simulated annealing is used for solving the combinatorial optimisation problem.

2.1.3. Structural Level

At this level, lines and regions are organised into a variety of 2-D shapes.

Mohan and Nevatia [**78**] use perceptual organisation for scene segmentation and description. This segmentation system generates hierarchies of features that correspond to structural elements such as boundaries and surfaces of objects. Based on Gestalt principles, edges are grouped to form curves. Contiguous curves are grouped to form contours while symmetric curves are grouped to form symmetries. Next, symmetries will become ribbons if closure is detected. An exhaustive search is used to find relationships between different features. Before each search, invalid or conflicting hypotheses of any joins or groups are removed using geometric constraints: cocurvilinearity, continuity, proximity, and co-termination. Promising results are demonstrated

on real images with a small number of objects. However, because of the inefficient search method, the complexity can grow exponentially for more complex scenes.

To overcome the computational complexity of many hierarchical approaches, Sarkar and Boyer [**107**] propose a voting method and graph-theoretic structure to represent the data organisation. They recognise that the bottleneck of the system is the compatibility test among all pairs of tokens. By building a histogram of the token's feature similar to the Hough transform, the compatibility test then becomes a bounded search through the parameter space.

Both methods proposed by Sarkar and Mohan utilise only edges as input to the system. On the other hand, Schlüter and Posch [**108**] proposed combining both contour and region information for perceptual grouping. In this method, edges are first grouped recursively to form 2-D closures (closed regions). At the same time, region segmentation is performed and then the resulting region map is matched to the closest edge group. Additional boundaries are generated if some regions cannot be matched to any edge group.

2.1.4. Conclusions

Perceptual grouping is a basic and effortless capability of the human visual system. However, as reviewed in this section, this grouping task is deviously not simple but a very complicated process that encompasses several levels of abstraction. Although a lot of research have been done on this topic, there is still no general theory that can explain most of the known visual grouping phenomena, such as figure-ground discrimination and object detection.

2.2. Visual Attention System in Humans

In order to replicate human visual performance, we have to analyse and understand how the system works within our brains. Even though most of the human brain's functional mechanisms and its underlying neural circuitry are still unknown, a basic idea about the visual system can be acquired from psychophysical and neurophysiological experiments conducted in the past. Based on these findings, a biologically motivated model of attention can be devised.

2.2.1. Structure of the Human Visual System

Visual information enters the nervous system in the retina, travels through the lateral geniculate nucleus (LGN), and then enters the cerebral cortex at the back of the head in an area named V1 (also known as the "striate cortex"). From this starting point, information branches off and travels forward into the many specialised visual areas that are located in the posterior half of the brain (called "extrastriate" visual areas). As the information travels forward from the striate cortex into the extrastriate cortex, the features coded by single neurons change from simple bars and edges to more complex attributes of object identity.

2.2.1.1. The Retina

Two types of photosensitive cells, rods and cones, exist in the retina. They have different sensitivities and adaptation mechanisms to different wavelengths. Cones are associated with colour vision whereas rods are associated with vision at low light levels. Three different types of cones (red "*actually yellow*", green, and blue cones) are found in the human retina while a fourth type of cone, the double cone, is found in non-primate visual systems. These cones appear to be distributed more or less randomly in the retina, but there are many fewer cones for blue than for green or red. The relative numbers of red, green, and blue cones are found to be in the ratio of 40 to 20 to 1 [18].

An interesting characteristic of the retina is the non-uniform distribution of the photoreceptors. The density of these receptors is much higher at the centre of the retina, called the fovea, than in the surrounding region. The density of the receptors decreases with the distance from the centre. This foveated-sampling scheme provides significant data reduction at the expense of having to physically move the fovea to the point of interest.

2.2.1.2. The LGN

The LGN represents an intermediate relay stage between the retina and the visual cortex. The LGN, organised in six layers, is an important switching device used to segregate the parvocellar (P) and magnocellar (M) channels and to align the input from the two eyes. The M layers are concerned primarily with non-colour vision processing (e.g., motion of objects and spatial reasoning) while the P layers are very important for colour vision processing (e.g., object recognition). Three of the layers

receive input from the ipsilateral eye and the other three from the contralateral eye. The distinctions between P and M cells are still maintained in the cortex.

2.2.1.3. V1

V1 is layered like the LGN. There are three types of cells or neurons in the V1: simple, complex, and hypercomplex. Simple cells are characterised by receptive fields with excitatory and inhibitory fields, and whose profile can be modelled by Gabor functions [**55**]. Complex cells show orientation selectivity in much the same way as simple cells but they do not have distinct excitatory and inhibitory zones (not phase sensitive). Finally, hypercomplex cells, also called end-stopped cells, are very sensitive to line endings, curvature, and angles. With these cells, several perceptual properties can be detected such as selectivity in orientation, size, position, colour, direction, and depth. The responses of all V1 neurons can be thought of as retinotopic feature maps characterising the visual stimulus captured by the retina.

After V1, both the pathway and functions become more complex. The presence of crossover and feedback make it very difficult to analyse and interpret the actual layout of the neural circuitry.

2.2.1.4. Discussion

One of the reasons for the existence of attention is the need to shift the high-resolution fovea onto the most important parts of a scene, providing a detailed description of the object of interest. The low-level features extracted and encoded in the human visual system include colour (red, green, and blue), texture, position, motion and depth.

2.2.2. Psychophysical Aspects of the Human Visual Attention System

Many of the mechanisms of human visual attention have been discovered through psychophysical experiments. In these experiments, human performance is evaluated during some specific, visuomotor task. Most psychophysical investigations involved with attention are actually concerned with covert attention, and its facilitation effects on visual tasks.

Two basic models of human visual attention are the zoom-lens model and the spotlight model. The first model was initially proposed by Jonides [**56**] and then further developed by Eriksen and his associates [**31**] [**32**]. They propose that attention

is analogous to a zoom-lens system. At a low-power setting, attentional resources are evenly distributed across the visual field. If the discrimination task is difficult, or when a pre-cue had been previously flashed, the attentional system zooms in to that area and allocates a disproportionate share of the processing resources to it. However, not all attentional resources would be employed in the pre-cued area. The remaining resources are shared among other locations. The second model was first introduced by Neisser [**81**] and then modified by Julesz [**58**] and Treisman [**123**] [**124**] [**125**] [**126**] [**127**] [**129**]. This paradigm proposes that attention involves two distinct stages, preattentive and attentive stages. In the first stage, processing is performed in parallel over the whole field, whereas in the second stage, a sequential analysis of some parts of the image occurs. The spotlight metaphor is proposed for the attentive stage since it would only affect a limited area of the visual field. Even though the debate about this second model is still open [**139**], it is by far the most accepted paradigm of visual attention.

2.2.2.1. Top-down and Bottom-up Control

The two basic mechanisms that control visual attention can be described as goal-driven (top-down), and stimulus-driven (bottom up) processes. This distinction is not new. For example, William James (1890) [**54**] characterises this distinction in terms of "active" and "passive" modes of attention. Attention is said to be goal-driven when the attention is controlled by the observer's deliberate strategies and intentions. In contrast, attention is said to be stimulus-driven when it is controlled by some salient attributes of the image that are not necessarily relevant to the observer's perceptual goals.

2.2.2.2. What features catch the eye?

The most important question about the visual attention system is what features can catch the eye's attention or which feature attracts the most fixations. For the passive bottom-up mode of attention, it is necessary to identify a set of basic features used in preattentive processing and determine whether attention depends on the feature itself, the feature contrast, or both. It is also important to find out whether these features have equivalent effects in drawing attention. Many experiments have been conducted to analyse different stimulus properties. In general, targets having distinct features are perceptually salient and stand out from a background pattern.

For the first question, Wolfe [**146**] has an extensive review on defining a basic feature set for visual search. The presumption is made that if a stimulus supports both efficient search and effortless segmentation, then it is safe to include it in the basic set. He states that there is a reasonable consensus about a small number of basic features and more debate over several other candidates. Some of the basic features consistent with the experimental results are:

- *colour*. [**136**] Much research has led to the conclusion that colour is one of the best ways to make a stimulus "pop-out" from its surroundings. For simple patterns, colour difference alone is sufficient for efficient visual search and effortless texture segmentation.
- *orientation*. [**35**] Orientation is also well-accepted as a basic feature in visual search. However, a difference of 15 degree or more is needed to support efficient visual search.
- *curvature*. [**128**] It has been found that curved lines can be found among straight distracters using parallel processing. This implies that the time required for detecting the curved lines does not differ significantly with the number of targets. However, the search is less efficient if the target is straight and the distracters are curved.
- *size*. Treisman and Gormican [**128**] conclude that it is easier to find big objects among small ones than small among big. However, for a given size of distracters, finding a bigger target is no easier than a smaller one. In addition, the slope of the reaction time against the number of targets is very steep, implying that size is not a good basic feature for visual search: except for a simple case in which a big circle is surrounded by much smaller ones.
- *motion*. [**74**] It is apparent that it will be very easy to find a moving stimulus among stationary distracters.
- *shape*. Wolfe states that shape is probably the most problematical basic feature because there is no widely agreed layout of "shape space". Some candidates for the *axes* of this space are line termination [**57**], closure [**27**], and face [**33**].

For the second question about the significance of a certain feature and its contrast in drawing attention, Northdurft [**82**] has performed a series of experiments designed to investigate the role of features versus feature contrast in preattentive vision. His study shows that features, in general, are not found to play an important role in these tasks and performance was instead related to feature contrast. Only in the case of

colour does performance also depend on the hue feature. Theeuwes' [**121**] experiment also shows the attention-grabbing abilities of colour. Recent results presented by Mannan et al. [**70**] also suggest that initial fixation placements are not controlled by perceptual features alone. In this study, eye movements were measured while viewers examined grey-scale photographs of real-world scenes. They also attempted to specify the visual features that determined initial fixation placement [**71**]. They analysed local regions of their scenes for seven spatial features: luminance maxima, luminance minima, image contrast, maxima of local positive physiological contrast, minima of local negative physiological contrast, edge density, and high spatial frequency. From their analysis, only edge density predicted fixation position to any reliable degree and even this feature produced only a relatively weak effect. Thus, the nature of the visual features that control fixation placement in scenes is still unclear.

For the last question, whether or not features have equivalent effects in drawing attention, the intuitive answer would be no. Based on experiments in which subjects search for singletons (a singleton is a single target among homogeneous distracters and differs from those distracters by a single basic feature), Muller and Found [**79**] argue that the contribution of any specific feature to the overall salience of any object is controlled by a weight that can change from task to task and, indeed, from trial to trial. They find that the reaction time for trial *N* is contingent upon the relationship between target identity on trial *N* and *N-1*. That is, people are faster to find a colour singleton on trial *N* if a colour singleton is found on trial *N-1*. While experimental results support the uneven weightings of different features in drawing attention, how these weightings are distributed or how they are altered quantitatively, has yet to be explained.

Most of the early psychological experiments were conducted with simple images having a dark background and simple objects such as bars, circles, squares, and letters. For these images, it is very easy to distinguish the background from the objects. These experiments are useful in isolating the effects of different features, but not for showing their inter-relationships. The attention-grabbing ability of different features on complex real images may differ from these simple ones. To understand how eye movement is controlled in more realistic visual-cognitive tasks, reading and scene viewing have been studied. A common assumption in these studies is that the fixation point of the eye is the focus of attention at a given time. Buswell [**11**] finds that the fixation positions are highly regular and related to information in the pictures.

For example, viewers tend to concentrate their fixations on the people rather than on background regions when examining Sunday Afternoon on La Grande Jatte by Georges Seurat. Henderson et. al. [**46**] also have found that first pass gaze duration and second pass gaze duration are longer for semantically informative than uninformative objects, providing evidence for relatively early, peripherally-based scene analysis. To determine whether attention is related to semantic informativeness (the meaning of the region) beyond visual informativeness (the presence of discontinuity in texture, colour, luminance, and depth), Henderson et al. [**44, 45**], conducted a series of experiments with the semantic informativeness defined as the degree to which an object was predictable within the scene. An unpredictable object will have high semantic informativeness and vice versa. They do not find any tendency by the viewer to immediately fixate their attention on semantically informative objects. De Garaf et al. [**24**] also found no evidence that semantically inconsistent objects were fixated earlier than consistent objects. However, they observe that viewers tend to look back more often to semantically informative than to uninformative scene regions. These results suggest that the attention is first driven by a bottom-up process before a more organised top-down process is engaged to analyse the scene in more detail.

2.2.2.3. Are objects available preattentively?

A recent debate in the literature concerns whether covert attention is directed to unsegmented regions of space, or to segmented perceptual groups that are likely to constitute coherent objects. As our actions must ultimately be directed toward individual objects, some theorists have proposed that it would be efficient for covert attention to operate on segmented objects rather than on unstructured regions of space [**4**] [**29**] [**81**]. The space-based and object-based models of attention are often presented as mutually exclusive alternatives [**4**]. However, many hybrid views are possible. For instance, covert attention may operate within a spatial medium (as argued by Tsal and Lavie [**130**]), but grouping processes may act to modulate the spatial extent of the attended region (Lavie and Driver [**62**]). Lavie et al. [**62**] examined the relation between segmentation and spatial attention by examining patients having disorders (extinction, neglect, and Balint's syndrome) after brain damage. He found that the effects of these brain-damage-related syndromes can be reduced if the two concurrent events formed a good perceptual group such as dumbbell shape instead of two circles. Based on this evidence, he argues that spatial attention is directed

within a segmented representation of the visual scene, with at least some of this segmentation taking place preattentively. Rensink et al. [**102**] also show that objects have some preattentive existence by demonstrating that preattentive processes are sensitive to occlusion. Wolfe [**145**] has conducted a series of experiments that make a similar point. These results support the idea that objects can exist preattentively.

2.2.3. Conclusions

In this section, recent and past discoveries and knowledge about the human visual system are presented. From this review, it can be seen that there is no general agreement on major issues of the visual attention system, such as a model of attention, selection of a basic feature set, and the spatial medium of the attention process. Nevertheless, there is both physical and psychological evidence showing the existence and importance of a small set of basic features, which include colour, texture, position, and motion, within the human visual attention system. In additions, object-based attention systems have also been proposed both as an alternative or as an complement to the space-based model of attention.

2.3. Visual Attention Systems in Machines

Recent advances in computer technology are astonishing and have made a real-time machine vision system feasible. However, despite enormous progress in recent years, machine vision systems still have a long way to go before approaching the level of human performance. The main reason for this is the lack of effective and efficient algorithms for many general computer vision processes, such as image segmentation and object recognition. One remedy to this problem is information selection or data reduction so as to reduce computational time and to suppress irrelevant data and noise. Starting from the mid-80's, specific efforts have been made towards more effective models of attention. Since that time, more than ten models has been proposed [**75, 50, 109, 101, 36, 21, 104, 131, 19**]. Most of these models, however, have been tested only on simulated data. In reality, we seldom see any objects with perfectly uniform colour and texture. Even for artificial objects, the surface property may be affected by shadows, highlights, and non-uniform lighting. For the model to be practical, it should be able to tolerate a certain amount of noise and be applicable to a

wide range of environments. Its performance should also degrade gracefully in case of failure.

The attention models proposed by Koch [**50**] and Milanese [**76**] are very similar and are based on an architecture previously proposed by Koch and Ullman [**59**]. This architecture is related to Treisman's feature integration theory [**128**]. Visual input is first decomposed into a set of feature maps. Colours, intensity, and orientations are used in both models while edge magnitude and curvature are also used in Milanese's model. These maps are then transformed into conspicuity maps representing the "conspicuity" of locations. Integrating all the conspicuity maps forms a final saliency map. The final stage of these two models is not the same because their intended applications are different. Koch's model is used for simulating the scan path so that a winner-take-all selection scheme and inhibition of return are used as the final stage. On the other hand, since the purpose of Milanese's model is for locating and recognising objects, the saliency map is further processed to provide both the position and region information which are fed into another higher-level process for object recognition.

Sela and Levine [**109**] model interest points as the loci of centres of co-circular edges. Experimental results on real images show that centres of symmetry correlate well with human fixation points. Reisfeld et al. [**101**] and Gesú et al [**36**] also use symmetry in predicting fixation centres.

In time-varying imagery, Conception and Wechsler [**21**] proposed an attention scheme based on edge maps, motion cues, and past history. In their algorithm, the saliency map is used to guide the coarse to fine classification of objects so that the amount of information to be processed later is reduced tremendously. Their main contribution is the integration of active and selective attention with learning and memory in a hierarchical framework. Rybak et al. [**104**] described an attention model for explaining invariant object recognition in humans. In their model, attention is used to guide visual perception and recognition. However, the attention mechanism is a top-down process instead of bottom-up.

Apart from general visual attention systems, Tsotsos et al. [**131**] proved that in visual search, if explicit targets are given in advance, the time complexity will be a linear proportion of the image size. On the other hand, if no explicit target is provided, the task is NP-complete. Thus, they propose that the human brain may not be solving this general problem and it is necessary to have attentional selection to

guide the search process. A model of primate visual attention is also presented that is both biologically plausible and computationally feasible. A top-down hierarchy of winner-take-all processes is embedded within the visual processing pyramid. However, they also state that a balance between data-driven and knowledge-driven processes must be achieved.

Osberger and Maeder [86] present a method for determining the perceptual importance of different regions instead of point locations in an image. They selected five factors that have been found to influence visual attention in assessing the overall importance of each region. These factors are: contrast, size, shape, location, and foreground-background. The final saliency measure is obtained by the summation of the square of each factor.

2.3.1. Conclusions

Most of the attention models proposed for machine vision are spaced-based where perceptual saliency is determined by local feature contrast, such as Koch's model and Milanese's model. On the other hand, object-based attention models also are receiving increasing amounts of attention. For these models, object properties, such as symmetry, region size, shape, and intensity contrast are considered. It is not clearly understood which approach is more efficient or effective in modelling human attention. However, since most computer vision tasks are finally focused on individual objects, and not much research have been done on this topic, it is worthwhile and fruitful to investigate object-based attention in greater detail.

CHAPTER 3

Perceptual Saliency Measure

This Chapter explores how object-based visual attention can be modelled in a machine vision system. Those factors which have been identified by Osberger and Maeder [86] will be presented along with several new measures influenced by psychophysical evidence. Methods for combining these factors will also be discussed in this Chapter.

3.1. Perceptual Saliency Factors

In most cases, a perceptually salient region will correspond to a perceptually meaningful or interesting object. However, in some situations, a perceptually salient region may not be related to any logical objects. In scene viewing, Henderson and Hollingworth [45] find that initial fixation placement does not seem to depend on the semantic informativeness of regions. In these experiments, semantic informativeness is defined as how unlikely the scene region is expected from the context. However, people tend to look back more often to semantically informative objects. Hence, if visual attention is defined as the point of fixation, there exist at least two definitions for visual attention. The first definition is what kinds of regions can attract fixations instantaneously within the first two seconds of viewing. The second one is which regions viewers will look back to more often. These revisited regions are what the viewers are interested in and seek to know more about. This overt attention often involves a high-level top-down process with the goal set by the viewer. Objects that people usually look for include human faces, animals, automobiles, and aeroplanes. Usually, people are less interested in objects that often form the background, such as the sky, floor, and wall. As a result, whenever human judgement is used in assessing

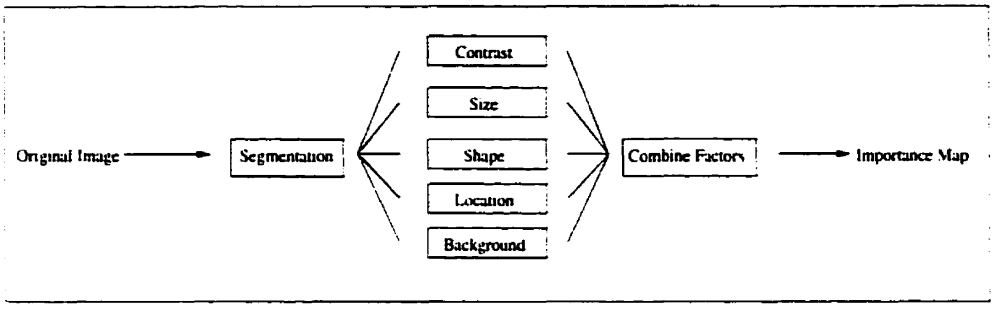


FIGURE 3.1. Block diagram for Osberger and Maeder's Importance Map calculation

an attention model's performance, these two distinctions have to be stated clearly. In this thesis, our attention will be focused primarily on the low-level, bottom-up process.

3.1.1. Osberger and Maeder's model

The purpose of this model is to automatically determine the perceptual importance of different regions in an image. The block diagram for Osberger and Maeder's importance map calculation is shown in Figure 3.1. In [**86**], eight low level features and four higher level factors are identified which have been found to influence human visual attention. These low level features are intensity contrast, size, shape, colour, motion, brightness, orientation, and line endings. Higher level factors are location, foreground/background, people, and context. These features are similar to those identified by Wolfe [**146**] as described in Chapter 2. Of these features, only five factors are selected by them for modelling visual attention. The mathematical definition for these five factors are stated below. In order to be able to compare these factors directly, they are scaled to fit in the range [0,1].

- *Contrast of region.* Regions having high contrast with their surroundings are found to be visually salient. Hence, the contrast importance $I_{contrast}$ is defined as the difference in the mean grey level of the region R_i and its surrounding regions $R_{i-neighbours}$.

$$I_{contrast}(R_i) = \overline{gl}(R_i) - \overline{gl}(R_{i-neighbours}) \tag{3.1}$$

  where $\overline{gl}(R_i)$ is the mean grey level of region R_i, and $\overline{gl}(R_{i-neighbour})$ is the mean grey level of all neighbouring regions of R_i.

- *Size of region.* All else being equal, larger regions are more likely to attract visual attention than smaller ones. In other words, larger regions are easier to detect than smaller ones. However, this effect levels off after a certain threshold. The size importance is defined as:

$$I_{size}(R_i) = min(\frac{A(R_i)}{A_{max}}, 1.0) \tag{3.2}$$

where $A(R_i)$ is the area of region R_i, and A_{max} is a constant used to prevent excessive weighting being given to very large regions. They set this constant to 1% of the total image area.

- *Shape of region.* Elongated objects have been found to attract more attention than rounder blobs of the same area and contrast. Importance due to region shape is defined as:

$$I_{shape}(R_i) = \frac{bp(R_i)^{sp}}{A(R_i)} \tag{3.3}$$

where $bp(R_i)$ is the number of pixels in the region R_i which border with other regions, and sp is a constant. They found a value of 1.75 for sp suitable for discriminating long, thin regions from rounder ones.

- *Location of region.* Experiments have shown that viewers are directed at the centre 25% of a scene while viewing television [**30**]. Thus, importance due to location of a region is defined as:

$$I_{location}(R_i) = \frac{centre(R_i)}{A(R_i)} \tag{3.4}$$

where $center(R_i)$ is the number of pixels in region R_i which are also in the center 25% of the image.

- *Foreground / Background.* Osberger et al. assume that a region connected to the border of the image will have a higher probability of being at the background. This assumption is valid if the main objects are not located along the border of the scene or there are one or two major backgrounds that contain most of the image borders. This measure is defined as:

$$I_{bg}(R_i) = 1.0 - max(\frac{borderpix(R_i)}{0.5 * totalborderpix}, 1.0) \tag{3.5}$$

where $borderpix(R_i)$ is the number of pixels in region R_i which also belong to the border of the image, and $totalborderpix$ is the total number of image

border pixels. Based on this definition, regions with a high number of image border pixels will be classified as belonging to the background and will have a low *foreground/background* importance.

3.1.2. Discussion

The five factors chosen by Osberger and Maeder are useful for modelling human visual attention in simple situations with strong "pop-out" effects. As described in Chapter 2, the most widely agreed assumption that has been used in many psychological experiments is that an object or target is salient and pops-out from the background if its visual features differ from other objects. This idea is proposed by Triesman in her Feature Integration Theory [**127**]. Contrast or difference in visual features can facilitate visual search and thus is visually salient. Contrast can be defined not only by intensity, but also by other low-level features such as orientation and colour. However, contrast alone is not enough for explaining the "pop-out" effect of objects having distinct features among other distracters. Contrast can only be used to explain the *relative* perceptual saliency of *isolated* objects: not for objects adjacent to each other. This is not hard to understand, as shown in Figure 3.2.

Contrast is usually defined as the distance in the feature space. In case 1, intensity contrast for region A and region B is 70 and 50, respectively, and thus region A is perceptually more salient. This prediction is consistent with human judgement. In case 2, however, the contrast for region A and region B is the same, with a value of 70. The problem with this image is the lack of a common reference frame for interpretation. One interpretation of this image can be a very large bright square having a rectangular hole in the middle. Another interpretation can be a dark bar in a uniform white background. Although these two cases are very simple and probably would not occur in reality, they show the necessity for a good measure of figure-ground discrimination. In case 3, if someone is asked to decide whether region A or region C can attract more attention, the answer would be A. From the contrast calculation, the value of saliency of region A is 30 while that of region C is 35, $[(95-30)+(100-95)]*0.5$. Hence, the prediction based on contrast alone could be wrong for regions adjacent to high contrast regions.

In assessing the relative depth information of different regions, Osberger et al. use the percentage of image border as an indication of background. This means salient objects are presumed to occupy none or a very small portion of the border.

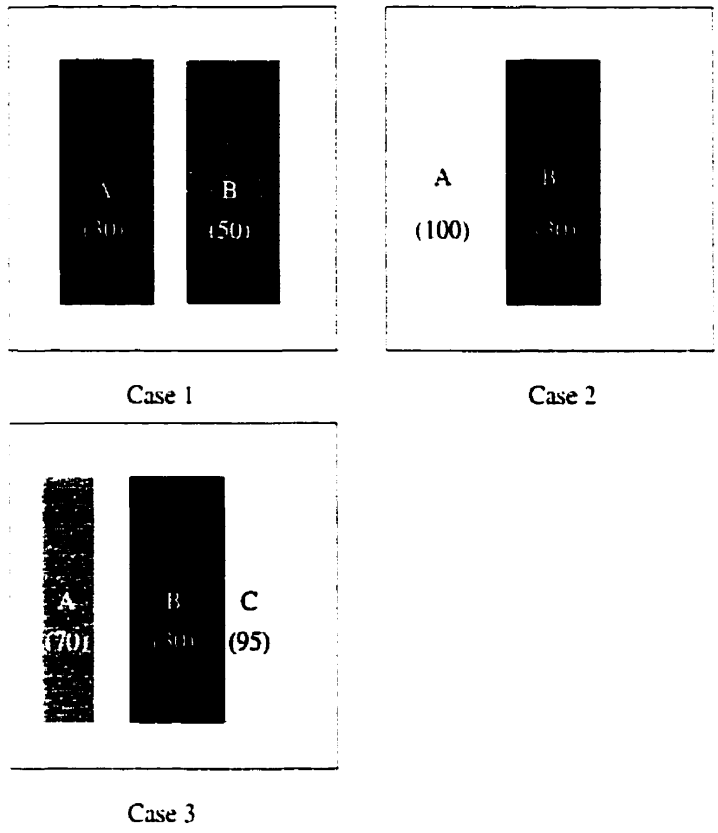


FIGURE 3.2. Situations where a contrast measure succeeds and where it fails. The number within the brackets indicates the region's intensity. For all three cases, the background's intensity is equal to 100.

Such an assumption is valid for most photographs since the most important objects are placed roughly in the centre of the image when the picture is taken. It is not valid, however, if this placement rule is not followed when the image is taken, such as pictures taken from a camera mounted on a mobile robot, or if a background region is separated into two isolated regions by an occluder. Some of these isolated regions may not even be close to the image boundary and thus will be assigned a very high value for foreground/background measure. In Figure 3.3a, region B is obviously in the foreground while regions A,C, and D belong to the background. Since region C does not touch the image border, it will be given a very high value, 1.0, for foreground/background importance. This problem can be solved by grouping regions A and C by similarity and continuation. However, this grouping must be done carefully to avoid grouping two seemingly distinct objects, such as region B and E.

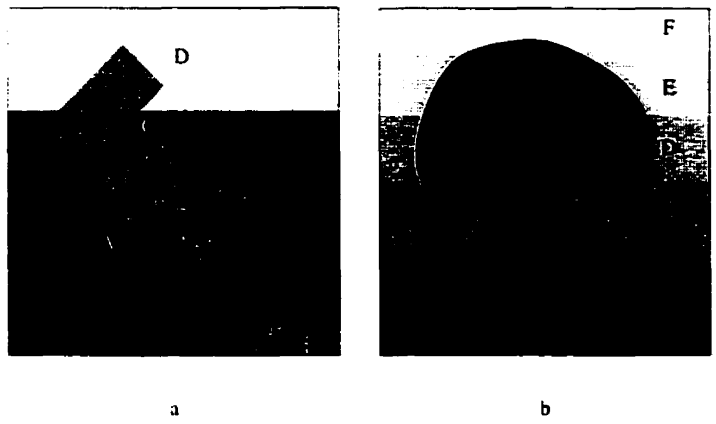


FIGURE 3.3. Situations where foreground/background measure fails

Another problem associated with this foreground/background method is illustrated in Figure 3.3b. In this figure, region A should be the main object with the rest of the regions belonging to the background. However, after counting the number of pixels in each region which also belong to the image border, region A will be assigned a lower foreground/background value (0.6) than those assigned to regions B, C, D, and E (0.9).

The problem of determining depth information from a single image is also explored by Rosenberg [**103**]. He uses occlusion cues to calculate the relative depth of each object. Six cases of occlusion are identified and used in a relaxation algorithm to infer the relative depth graph of the objects. The problem with this method is the requirement of a highly accurate image segmentation and the occurrence of occlusion. Moreover, the number of conflicts which have to be solved may grow exponentially for more complex scenes. Other monocular depth cues include relative size, linear perspective, texture gradient, relative height, and atmospheric perspective [**143**]. Although these depth cues are widely accepted and well-studied, depth perception still poses a big problem in practice since these cues usually involve a high-level understanding of the scene and therefore tend to work only in very restricted environments. For pictures taken by humans with some purpose in mind, the method proposed by Osberger et al. is applicable and is easy to compute without any prior knowledge of the scene.

The shape importance cue, as described in Chapter 2, is very controversial. For simple cases consisting of only circles and long thin rectangles, there is a very high probability that human fixations are more likely to fall on the rectangles than on the

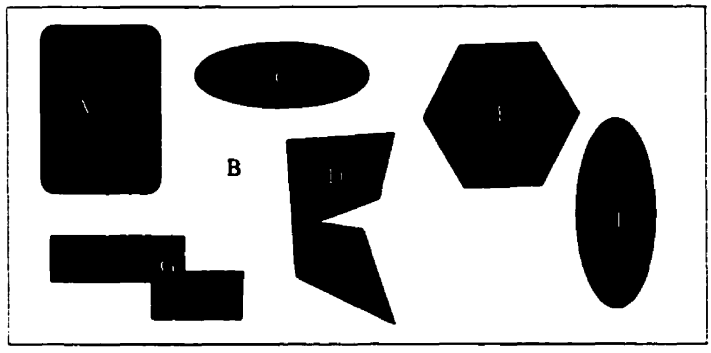


FIGURE 3.4. Situations where shape measure fails

circles. For more complex shapes, such as the example shown in Figure 3.4, however, the evidence is less clear. Which shape attracts most of our fixations? Is the "Z" shaped region G more salient than the irregular shaped region D? How about the hexagon E? The shape importance measure proposed by Osberger et al. favours elongated regions over rounder ones. The shape importance values of these regions are shown in Table 3.1. For this image, the most salient region predicted by the shape importance measure is the background B. This region is certainly not circular and has many long and narrow parts. Hence, it has the highest perceptual saliency value for shape! The major problem for any shape definition is the presence of "hole", such as the background. Do we consider its shape as the outline of its outermost boundary? Or do we also consider the inner boundaries such as the shape of a donut? In other words, do we treat the enclosed regions as textures or not? Apparently, there is no simple answer to these questions. If the application is restricted to certain environments and the most important objects are well-identified and known beforehand, one can make some useful conclusions about the shape saliency of regions. Otherwise, the usage of shape in modelling the human attention system should be approached cautiously if not eliminated altogether because this feature is not well-defined and its effect on attracting human visual attention is not well-understood in general.

3.1.3. New and Modified Importance Factors

Based on the discussion on Osberger and Maeder's method, some of their computational methods are modified and new factors are proposed.

- *Contrast in colour and texture.* The contrast importance $I_{contrast}$ will be redefined as the Euclidean distance in the mean colour and texture of the region

Region	Shape importance value
A	0.24
B (background)	1.00
C	0.40
D	0.55
E	0.31
F	0.35
G	0.45

TABLE 3.1. Shape importance values for Figure 3.4

R_i and its surrounding regions $R_{i-neighbours}$ as follows:

$$I_{contrast}(R_i) = \frac{1}{edgepix(R_i)} * \sum_{R_j \in neighbours\ of\ R_i} (|\ \overline{feat(R_i)} - \overline{feat(R_j)}\ |) \cdot border(R_i, R_j) \quad (3.6)$$

where $edgepix(R_i)$ is the perimeter of region R_i in pixels, and $\overline{feat(R_i)}$ is the mean colour and texture of region R_i[1]. $border(R_i, R_j)$ is the length (number of pixels) of the common border of region R_i and R_j.

- *Hue.* Since colour alone can grab human attention, especially *red* [**121**], it can be used in modelling visual attention. No matter how bright or how dark the object is, as long as its perceived surface hue is red (not black or white), it will be perceptually salient. However, no strong evidence has emerged concerning the attention-grabbing ability of hues other than red. In the case of face recognition, the hue of skin colour can be used to indicate its importance. Hence, the hue importance is defined as the distance from the reference hue as below:

$$I_{hue}(R_i) = e^{-\left(\frac{1-\cos(hue(R_i)-hue(Reference))}{sd}\right)^2} * tanh(sf * sat(R_i)) \quad (3.7)$$

where $hue(R_i)$ is the hue of the mean colour of region R_i in radians, and *Reference* is the preferred hue that is known to attract attention. *sd* is a constant used to control the threshold on the difference in hue between R_i and $reference$, and $sat(R_i)$ is the saturation of the mean colour of region R_i. A value of 0.1 for *sd* is found to be suitable for discriminating red from other hues. The second term is included to represent the uncertainty of hue at

[1] A discussion of how these are computed can be found in Chapter 4.

different saturation levels. At low saturation value, the colour appearance is grey and thus the value of hue is meaningless. Hence, a monotonic increasing function ($tanh(x)$) which levels off after a certain threshold is used for the hue uncertainty function. sf is another constant used to control the saturation level of the uncertainty function. For the CIE $L^*a^*b^*$ color space, a value of 0.0017 for sf is suitable.

- *Saturation.* In general, people are more interested in colourful regions with vivid colour. Colour saturation is considered by Braun [8] as a perceptual saliency factor. Simply, the importance of saturation is just the saturation level of the mean colour of the region .

$$I_{saturation} = sat(R_i) \tag{3.8}$$

- *Location.* The equation proposed by Osberger et al. has a sharp cut-off between the centre 25% of the image and the surrounding region. A more general form of this function is defined below:

$$I_{location}(R_i) = \frac{\sum_{pixel(x,y)\in R_i} f_{loc}(x,y)}{A(R_i)} \tag{3.9}$$

where $f_{loc}(x,y)$ can be any function relevant to the importance of location. In particular, the following function is used:

$$f_{loc}(x,y) = t(\frac{x}{w} - 0.5) * t(\frac{y}{h} - 0.5)$$

where

$$t(v) = \begin{cases} 1 & \text{if } abs(v) < 0.25 \\ 2 - 4v & \text{otherwise} \end{cases}$$

w and h is the width and height of the image.

- *New foreground/background measure.* In order to solve the problem associated with Osberger and Maeder's method discussed in the previous section, their method is modified. First, global region properties are used to group regions together if there is a high probability that these regions come from a single object. In reality, shadows, highlights, uneven lighting, and many other sources of noise are very common and unavoidable. Thus, it is better to perform the similarity testing adaptively so that the merge restrictions are tighter when the noise level is low and looser when the noise level is high. One possible

approach is to impose a restriction that only regions which form a single connected cluster in the feature space will be considered to be "similar". With this approach, the usage of an absolute threshold can be avoided. Since two separate objects can also have similar features that form a single cluster in feature space, as shown in Figure 3.3, another measure of "occlusion" must be used to estimate how likely the two regions belong to a single object and are separated by an occluder. In real scenes, we often observe that if a large background is separated by objects in the foreground, a large portion of the background would still be connected to the border with several much smaller isolated regions. Hence, a more conservative condition on the ratio of regions can be applied to further reduces the error probability of merging two different regions. A high probability for "occlusion" will be assigned only if the ratio of a region is much smaller than the total area of all the regions that are "similar" to this region. To solve the second problem associated with Osberger's method where the main objects occupy a large portion of the image border, the *foreground/background* measure can be defined as the ratio between the number of border pixels and edge lengths. The final foreground/background measure is defined as follows:

$$I_{bg}(R_i) = min(\frac{borderpixel(R_i)}{boundarypixel(R_i)}, 1 - \xi \cdot (1 - \frac{\sum_{condition(R_j,R_i)=1} borderpixel(R_j)}{\sum_{condition(R_j,R_i)=1} boundarypixel(R_j)})) \quad (3.10)$$

where

$$\xi = 1 - 2 * min(0.5, \frac{A(R_i)}{\sum_{condition(R_j,R_i)=1} A(R_j)})$$

$$condition(R_i, R_j) = \begin{cases} 1 & \text{if } R_i \text{ and } R_j \text{ form a single cluster} \\ & \text{and } borderpixel(R_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$borderpixel(R_j)$ is the number of pixels in region R_j which also belong to the border of the image, and the $boundarypixel(R_j)$ is the number of pixels in the boundary of region R_j. The function ξ is the probability of these regions being occluded by other objects.

3.2. Methods for Combining the Importance Factors

After obtaining the importance values for each factor, they have to be combined to give an overall ranking for each region. A simple summation method proposed by Osberger and Maeder [**86**] and four more complex combination strategies proposed by Itti and Koch [**49**] will be discussed in this section.

3.2.1. Osberger and Maeder's Method

Osberger and Maeder [**86**] choose to treat each factor as being of equal importance since it is difficult to determine exactly how much more important one factor is than another. They observe that very few regions would respond strongly for all factors and those regions identified by humans as salient usually have a very high ranking in only some factors. Hence, each factor is squared and then summed together to produce the final importance value as follows:

$$IM(R_i) = \sum_{k=1}^{n} (I_k(R_i))^2 \tag{3.11}$$

3.2.2. Itti and Koch's Method

Itti and Koch [**49**] have conducted an experiment to compare four feature combination strategies for saliency-based visual attention systems. The four strategies they considered are: (1) simple normalised summation, (2) linear combination with learned weights, (3) global non-linear normalisation followed by summation, and (4) local non-linear competition between salient locations. In their visual attention system, visual saliency is defined as the magnitude of spatial discontinuities in colour, intensity, and orientations at different scales. A large number of feature maps (a total of 32) is generated and combined by one of the four methods. They also observe that salient objects appear strongly in only a few maps and may be masked by noise or less salient objects. Experimental results show that the simple normalisation method consistently yields poor performance while the "trained" method yields the best performance. However, different learned weights are used for different image classes. The other two methods yield intermediate performance. Since the last two methods (3

and 4) do not require any learning procedures or any specific models. they are more generic and are applicable to a broader range of situations.

3.2.3. Discussion

As discussed in section 1.2. the foreground/background measure is more important than the contrast and shape measures in region-based attention. Hence. equal weights should not be used. In Itti and Koch's experiments. the "trained" method consistently yielded the best performance with a two-fold improvement when compared to the other methods. Since the parameters are allowed to vary for different test images. this method cannot be used in a general vision system. However. it would be useful to analyse the performance of a "trained" method with only one set of parameters for all test images. The other two methods proposed by Itti and Koch are more generic. however. the spatial normalisation functions used in these methods cannot be extended directly to a region-based feature map. Thus. these two methods will not be considered. As a result. the integration method that was used in this research is the weighted summation of all importance factors. with the weights obtained by experimentation from a large collection of test images. If no specific weights fon any factor can be found to improve the overall performance with confidence. one can either use equal weights for all importance factors or classify the test images into different categories and then find the optimum weights for each group.

CHAPTER 4

Feature Selection

The perceptual saliency functions described in Chapter 3 require the image to be pre-segmented into coherent, non-overlapping regions, that resemble the original physical objects in the scene. However, before an image can be segmented, it must be transformed into a set of feature maps that allow similarity and surface continuity to be defined. The most commonly used features for image segmentation are colour [20], texture [85], and position [13]. These features are intuitive to humans in discriminating and separating different objects. We usually use colour and texture when describing the visual properties of an object such as brown and curly hair, a smooth and shiny surface, etc. Position is also an important cue in discriminating objects since if two regions are far apart in the spatial domain, they have a lower probability of belonging to the same object. Biologically, special neurons in the human visual system are capable of detecting all of these features at an early stage. Spatial information about the objects can be easily included in the feature vector by including the x,y-coordinates of each pixel. However, utilising this extra information can have negative side-effects such as breaking up a large uniform region [13].

For colour and texture, many feature spaces and computational methods have been proposed in the literature. Hence, selection criteria must be adopted to choose a particular representation scheme for these features. Since the objective of the segmentation stage is to have the image segmented as if it were performed by a human, the feature space should also be perceptually uniform. That means the perceived difference of any two samples separated by a fixed distance in the feature space should be constant.

After the extraction of these features, they must be combined to form a single feature vector. During this integration process, decisions have to be made on how the features are to be combined and what to do if these features contradict each other. For example, how similar is the perceptual difference of a unit distance in colour space compared to a unit distance in texture space? In addition, since texture refers to the spatial distribution of colour, the colour of the pixels within a texture region will not be the same or even similar, unless it is a uniform "non-texture" region.

In the following sections, a brief review of colour and texture is presented, as well as methods for resolving conflicts that arise from the feature integration process.

4.1. Colour

The human visual system uses three different kinds of cones, each with different spectral sensitivity, to sense the colourful world (see Figure 4.1). These cones have peak responses at wavelengths of 580, 540, and 440 nm, respectively. With these three receptors, we can distinguish coloured lights with different wavelengths and intensities. Since the power spectrum of light in the visible frequency range is encoded by three channels only, this encoding is a many-to-one mapping and the original power spectrum cannot be recovered completely by the human visual system. However, this provides a useful additive property of the appearance of light. A mixture of two lights at different wavelengths can produce a colour that appears different from the two original light sources. As a result, the whole visible colour spectrum can be produced by mixing three or more primary colours at different proportions. As three channels are used in the human visual system, trichromacy has been adopted in computer vision for representing colour quantitatively. However, the wavelengths of the three primary colours defined in the CIE (Commission Internationale de l'Éclairage) standard are 700, 546.1, and 435.8 nm instead of the peak responses of human cones in order to match the light emitted by artificial light sources.

Based on this standard, all image capture and display devices are designed with these three primary colours, subject to small variations depending on the actual materials used. The "raw" format of any colour image is the RGB format specifying the relative intensity of the three primaries. Any colour is represented by a point $C(r, g, b)$ in a colour cube, as shown in Figure 4.2. The origin of the RGB colour space is the "colour" black and the full brightness of all three primaries together appears as

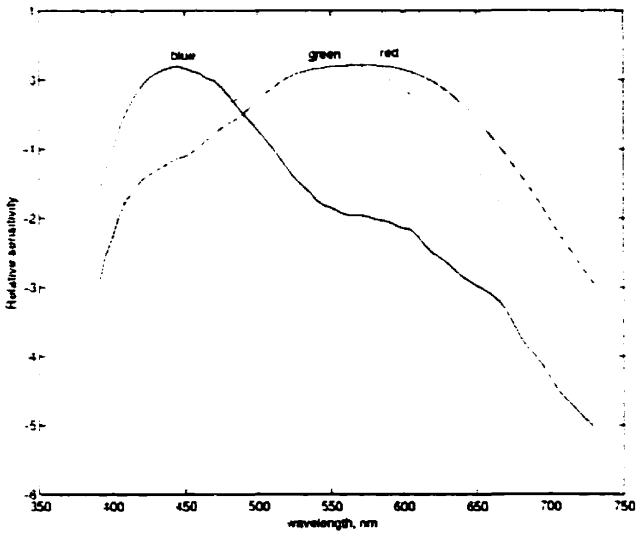


FIGURE 4.1. Spectral sensitivity of cones from Vos, Estévez, and Walraven [137]

white. Three corners of the colour cube located on the major axes correspond to the three primary colours: red, green, and blue. The remaining three corners correspond to the secondary colours: yellow, cyan, and magenta.

In the computer, each of these axes is encoded with 8-bits, ranging form 0 to 255. Initially, the *RGB* colour space is linearly related to the intensity. However, because of the nonlinear relationship between the input signal and the resulting brightness of most display systems, such as the cathode ray tube(*CRT*), the input signal to a display device must be modified to eliminate this nonlinear property. This compensation method is called gamma-correction. For a typical monitor, the electro-optical radiation transfer function is often expressed by a mathematical power function:

$$I = A * V^{gamma} \tag{4.1}$$

where I is the brightness of the pixel, A is the maximum luminance of the *CRT* and V is the applied voltage in the range of 0 and 1. For a conventional *CRT*, gamma is around 2.2. For convenience, images or photographs, especially those posted on the internet, which are intended to be viewed primarily from a PC, are already gamma corrected during the encoding process so that no extra correction is needed when displaying them. The resulting colour space is called nonlinear *RGB* space or *sRGB* space [118].

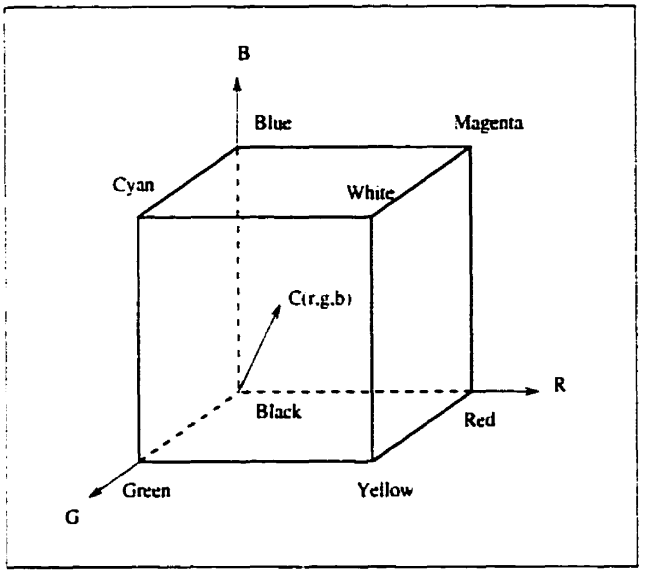


FIGURE 4.2. Colour cube

4.1.1. Colour Spaces

Due to the logarithmic relationship between the perceived brightness by humans and the actual intensity, the linear *RGB* space is perceptually nonlinear. Moreover, this colour system is not intuitive since people are more accustomed to the three basic attributes of colour: hue, saturation, and brightness. To correct these problems, new colour spaces and transformations of the *RGB* colour spaces have been proposed [**48, 147, 96**]. Some colour spaces are simply linear transformations of the RGB space: CIE 1931 *XYZ* and *Yxy*, CIE 1960 *YUV*, and CIE 1976 *YU'V'*. Colour spaces generated by nonlinear transformation include: YC_BC_R(JPEG and MPEG digital standard), *PhotoYCC*(Kodak PhotoCD system), *HSI*(Hue, saturation, and intensity), CIE 1976 ($L^*a^*b^*$), and CIE 1976 ($L^*u^*v^*$). Some colour spaces are obtained by collections of colour samples in the form of patches of paint, swatches of cloth, pads of papers, or printings of inks. Such systems are referred to as *colour order systems* and include the Munsell system, *DIN* system, Coloroid system (designed for use by architects), and *OSA* (Optical Society of America) system. No mathematical transformations have been proposed yet for these colour order systems except the Munsell system [**77**].

4.1.1.1. CIE 1931 *XYZ* and *Yxy*

The CIE 1931 XYZ system is defined such that all visible colours can be defined using only positive values [14]. Transformation from RGB to XYZ is defined as:

$$\begin{aligned} X &= 0.490 * R + 0.310 * G + 0.200 * B \\ Y &= 0.177 * R + 0.812 * G + 0.011 * B \\ Z &= 0.000 * R + 0.010 * G + 0.990 * B \end{aligned} \tag{4.2}$$

where both the RGB and XYZ values range from 0 to 1.

CIE also defines a normalisation process to compute the chromaticity coordinates to facilitate the representation of colour in the absence of brightness:

$$x = \frac{X}{X + Y + Z} \tag{4.3}$$

$$y = \frac{Y}{X + Y + Z} \tag{4.4}$$

4.1.1.2. CIE 1960 *Yuv* and CIE 1976 *Yu'v'*

Both Yuv and $Yu'v'$ are designed to produce a uniform chromaticity scale diagram in which a colour difference of unit magnitude is equally noticeable for all colours. However, the logarithmic response of the human eye on brightness is not modelled. The Yuv and $Yu'v'$ are obtained by the following equations, and Y is unchanged from the CIE XYZ system.

$$\begin{aligned} u &= \frac{4x}{12y - 2x + 3} \\ v &= \frac{6y}{12Y - 2x + 3} \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} u' &= u = \frac{4x}{12y - 2x + 3} \\ v' &= 1.5v = \frac{9y}{12Y - 2x + 3} \end{aligned}$$

4.1.1.3. YC_BC_R Colour Space

The YC_BC_R colour space is used in the JPEG and MPEG digital image format. The three channels are luminosity(Y), blue chrominance(C_B) and red chrominance(C_R). The separation of luminance from chrominance allows image-compression techniques

to take advantage of the eye's lesser need for resolution of colour than of brightness. RGB values are converted to YC_BC_R values in two steps. First, a nonlinear transformation is applied to the signal. The resulting values are converted to YC_BC_R through a linear transformation.

$$\begin{aligned} R' &= R^{0.45} \\ G' &= G^{0.45} \\ B' &= B^{0.45} \end{aligned} \tag{4.6}$$

$$\begin{aligned} Y_C &= 0.2990 * R' + 0.5870 * G' + 0.1140 * B' \\ C_B &= -0.1687 * R' - 0.3313 * G' + 0.5000 * B' \\ C_R &= 0.5000 * R' - 0.4187 * G' - 0.0813 * B' \end{aligned} \tag{4.7}$$

4.1.1.4. *PhotoYCC* Colour Space

The Kodak *PhotoYCC* colour space is designed for encoding images with the PhotoCD system and is similar to the YC_BC_R colour space. The only difference is that a different transformation matrix is used in the second step. The goal of the *PhotoYCC* colour-encoding scheme is to provide a definition that enables the consistent representation of digital colour images from negatives, slides, or other high-quality input and allows rapid, efficient conversion for video display. The nonlinearity of this colour space is based on the nonlinear property of video displays instead of the logarithmic sensitivity of the human eye.

For $R, G, B > 0.018$

$$\begin{aligned} R' &= 1.099 * R^{0.45} - 0.099 \\ G' &= 1.099 * G^{0.45} - 0.099 \\ B' &= 1.099 * B^{0.45} - 0.099 \end{aligned} \tag{4.8}$$

For $R, G, B \leq 0.018$

$$\begin{aligned} R' &= 4.5 * R \\ G' &= 4.5 * G \\ B' &= 4.5 * B \end{aligned} \tag{4.9}$$

$$\begin{aligned} Y &= 0.299 * R' + 0.587 * G + 0.114 * B' \\ C1 &= -0.299 * R' - 0.587 * G' + 0.886 * B' \\ C2 &= 0.701 * R' - 0.587 * G' - 0.114 * B' \end{aligned} \tag{4.10}$$

4.1.1.5. *HSV* (hue, saturation, and value) Colour Space

Different versions of *HSV* colour space have been proposed in the literature [**122**] [**43**]. The most commonly used *HSV* colour space is the cylindrical space where maximum saturation does not depend on the intensity value [**114**]. The problem with this space is the high sensitivity to noise for very dark colours. Alternative colour spaces are generated with different relationships between the intensity and maximum saturation, such as linear [**37**] and quadratic [**144**]. Despite modifications to the shape of this colour space, all *HSV* colour spaces make no reference to the perception of light by the human vision system. The transformation from *RGB* to *HSV* proposed by Travis [**122**] is given below:

$$V = max(R, G, B) \tag{4.11}$$

$$S = \frac{V - min(R, G, B)}{V} \tag{4.12}$$

$$\text{let } R' = \frac{V - R}{V - min(R, G, B)}, \; G' = \frac{V - G}{V - min(R, G, B)}, \text{ and } B' = \frac{V - B}{V - min(R, G, B)}$$

$$H = \begin{cases} 5 + B' & \text{if } R = max(R, G, B) \text{ and } G = min(R, G, B) \\ 1 - G' & \text{if } R = max(R, G, B) \text{ and } G \neq min(R, G, B) \\ 1 + R' & \text{if } G = max(R, G, B) \text{ and } B = min(R, G, B) \\ 3 - B' & \text{if } G = max(R, G, B) \text{ and } B \neq min(R, G, B) \\ 3 + G' & \text{if } B = max(R, G, B) \text{ and } R = min(R, G, B) \\ 5 - R' & \text{otherwise} \end{cases} \tag{4.13}$$

4.1.1.6. CIE 1976 $L^*a^*b^*$ and CIE 1976 $L^*u^*v^*$

Both CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ color spaces are intended to be uniform colour spaces. The colour differences in chromaticity and luminance are both taken into account in the minimisation process of the variation of perceptual differences of unit vectors. The nonlinear transformation for L^* is designed to mimic the logarithmic

response of the human eye. The CIE $L^*u^*v^*$ colour space is based on the CIE 1976 $Yu'v'$ while CIE $L^*a^*b^*$ is based directly on CIE XYZ. The equation for the parameter L^* is the same for both spaces:

$$L^* = \begin{cases} 116\left(\frac{Y}{Y_n}\right)^{1/3} - 16 & \text{if } \frac{Y}{Y_n} > 0.008856 \\ 903.3\left(\frac{Y}{Y_n}\right) & \text{otherwise} \end{cases} \tag{4.14}$$

$$\begin{aligned} u^* &= 13 * L^*(u' - u'_n) \\ v^* &= 13 * L^*(v' - v'_n) \end{aligned} \tag{4.15}$$

$$\begin{aligned} a^* &= 500 * (fn(\frac{X}{X_n}) - fn(\frac{Y}{Y_n})) \\ b^* &= 200 * (fn(\frac{Y}{Y_n}) - fn(\frac{Z}{Z_n})) \end{aligned} \tag{4.16}$$

where

$$fn(t) = \begin{cases} t^{1/3} & \text{if } t > 0.008856 \\ 7.787 * t + \frac{16}{116} & \text{otherwise} \end{cases} \tag{4.17}$$

Y_n, X_n, and Z_n define the appropriately chosen reference white and u'_n and v'_n are the values obtained from the equation for $Yu'v'$ using this reference white point.

4.1.1.7. The Munsell System

The Munsell system is one of the most widely used colour order systems, originated by the artist A.H. Munsell in 1905. An important feature of the Munsell system is that the colours are arranged so that, the perceptual difference between any two neighbouring samples is as close to constant as possible. Miyahara and Yoshida [77] proposed a transformation, called the Mathematical Transformation to Munsell (MTM), based on the CIE 1976 $L^*a^*b^*$. However, this is just an approximation to the Munsell system. There does not exist a simple and exact mapping from RGB or XYZ to the Munsell coordinate.

4.1.2. Conclusions

All linear transformations of the *RGB* space do not agree with the logarithmic brightness sensitivity of human eyes. Among the nonlinear transformations, it is not clear which colour space has the highest perceptual uniformity and how much more uniform one colour space is when compared to another colour space. Nevertheless, since the CIE $L^*u^*v^*$ and CIE $L^*a^*b^*$ colour spaces both have been tested extensively using psychophysical experiments [**117**] and are widely accepted as perceptually uniform spaces, either one of these two colour systems can be used in representing the surface colour of objects. In particular, the CIE $L^*a^*b^*$ is selected for this project.

4.2. Texture

Texture is an important attribute in describing the surface properties of objects. Images of real objects often exhibit certain particular patterns of colour. These patterns can be the result of physical surface properties, such as irregular surface orientation, or they could be the result of reflectance differences, such as differences in material and colour. This perception of texture, while very obvious and effortless for humans, is very difficult to define formally and precisely. A large number of features have been identified by researchers and have proven to play an important role in texture identification. These features include coarseness, contrast, directionality, line-likeness, regularity, roughness, uniformity, density, linearity, direction, frequency, phase, and complexity [**120**][**1**][**63**]. These features are not independent and are correlated with each other, such as directionally and line-likeness. Because of the high dimensionality of the texture space, there is no single method of texture representation which can model adequately all aspects of texture [**133**]. Most texture research has been conducted on the Brodatz texture collection, samples of which are illustrated in Figure 4.3.

Although there is no generally agreed definition of texture, several basic assumptions are commonly used in texture analysis. First, textures are homogeneous patterns or spatial arrangements of pixels. Many papers on texture have considered only grey-scale images, although colour texture has become a focus of recent research [**89**][**51**]. Secondly, unlike colour, texture is a region property instead of a point property. As a result, its definition must involve pixels in a spatial neighbourhood. The decision on selecting a suitable size for this neighbourhood depends on the texture type and

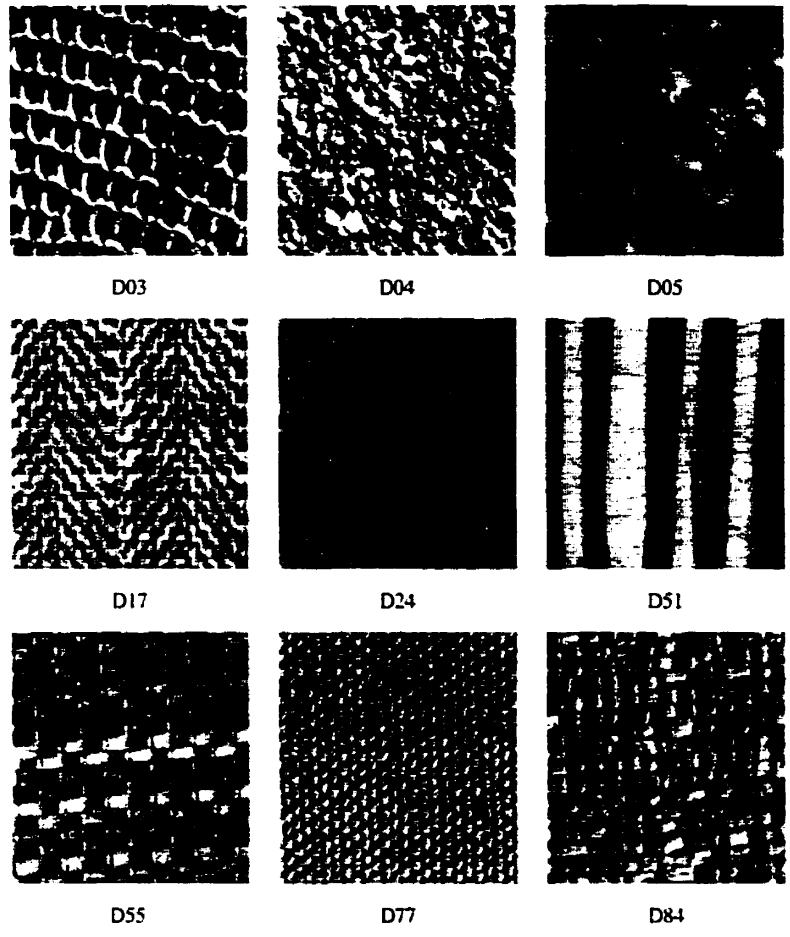


FIGURE 4.3. Texture samples from the Brodatz collection

the trade-off between noise-suppression and edge-localisation. With a larger spatial support, a more robust estimation of the texture can be obtained. At the same time, utilising a bigger neighbourhood reduces the spatial resolution of the texture by smoothing out the edges. The last assumption on texture is its multi-scale properties. For example, a coarse view of a tree shows the leaves and branches while a closer look at the tree reveals the fine details of the bark and the veins of the leaves. Unfortunately, it is unclear where this transition (when the leaves are perceived as objects by themselves) occurs in texture segmentation.

4.2.1. Related Work on Texture

A substantial amount of work has been done on the problem of texture analysis, classification, segmentation, and synthesis. A large number of surveys have already

been published [142] [40] [138] [135] [28] [100] [84] [133] [98] on texture analysis alone.

In [133], Tuceryan and Jain categorise existing texture models into four major classes: statistical, geometrical, model-based, and signal processing method. Statistical methods extract texture features from the spatial distribution of grey values, such as co-occurrence matrices [41]. Under the category of geometrical methods, texture is defined as a composition of "texture elements" or primitives. Voronoi tessellation features proposed by Turceryan and Jain [132] is one example of this category. In model-based methods, textures are presumed to possess certain structures and these structures can be described locally. Based on these assumptions, Markov random fields (MRFs) [88] and fractal geometry are commonly used for modelling images. These methods can be used not only for describing texture, but also to synthesize it. In signal processing methods, the texture features are obtained from a set of filtered images. Studies in psychophysiology have suggested that the visual system decomposes the image formed on the retina into filtered images of various frequencies and orientations [12]. The study conducted by De Valois et al. [25] on the brain of the macaque monkey concluded that simple cells in the visual cortex of the monkey are tuned to narrow ranges of frequency and orientation. Moreover, the receptive fields of simple cells can be modelled closely by Gabor functions. These studies have led to the use of multi-channel analysis for texture representation. As a result, Gabor and wavelet models, in particular, are widely used for texture analysis.

Very few quantitative comparisons between different texture feature representation schemes have been presented. Most studies have used mosaic images for benchmarking. These test images are generated by randomly selecting two or more texture samples from the Brodatz's collection and then combining them side-by-side to form a texture mosaic. Despite the small number of comparative studies, experimental results do not agree with each other [98][16]. Co-occurrence features give the best performance in the studies of Strand and Taxt [119] and Ohanian and Dubes [83], while Laws [63] and Pietikainen et.al. [94] had the opposite conclusions. Recently, Randen and Husoy [98] compared a large number of filtering approaches including the Gabor filter, different versions of the wavelet, and two classical non-filtering approaches, co-occurrence and auto-regressive features. This study shows that the performance of various filtering approaches vary for different textures. No single approach performs consistently well for all test images, and thus, no single approach may be selected as

the clear winner. However, if only the overall performance is examined, the 16-tap FIR quadrature mirror filter bank achieves the best overall results. To obtain the performance on real scene images instead of synthetic images, Chang, Bowyer, and Sivaguranath [**16**] compare grey-level co-occurrence, Laws texture energy and Gabor filters on 35 real images. Their results show that the performance of these three texture algorithms is much higher when tested on mosaic images than on real scenes. For example, 85% classification rate for Gabor filters on mosaic image and 71% on real images. In this study, Gabor filters offer the best performance.

The assumptions and objective for segmenting real scene images differ from that of segmenting mosaic images. For a real scene, it is preferable to have the image segregated into several non-overlapping regions depending on their perceptual similarity, since the size of the objects may vary from 5-pixels wide to half the size of the whole image. On the other hand, the objective of segmenting mosaic images is to segregate different texture patches regardless of their visual similarity. Thus, it is desirable to test not only a texture algorithm's discrimination power, but also how close the distance measure is to the perceived difference.

In an attempt to reduce the dimensionality of the texture space, Rao and Lohse [**99**] have conducted a psychophysical experiment to identify the high level features that are most relevant to the attentive perception of textures. To achieve this, they had 20 subjects perform an unsupervised classification of 30 pictures from Brodatz's album. Both hierarchical clustering analysis and multidimensional scaling analysis are used to identify and verify the dimensionality of the experimental data. This analysis shows that 95.5% of the variability in the classification data is preserved in a three-dimensional space. Rao and Lohse interpret these axes as repetition, orientation, and complexity. Although the sample size of 30 may not be large enough to give a complete picture of the texture space, the result of this study still indicates that many texture features are highly correlated and as few as three dimensions may be enough to represent a wide variety of textures.

4.2.2. Related Work on Unsupervised Segmentation of Natural Images

Many new image segmentation algorithms proposed in the last few years utilise both colour and texture to segment images. Most of these algorithms have been tested on a large set of real images to show their robustness and performance. Carson et.

al. [13] use joint colour, texture, and position as feature vectors. Instead of using classical methods for representing texture, they introduce a novel method to estimate the scale parameter of the underlying texture. At each pixel location, the average magnitude and direction of edge vectors within a local neighbourhood at several scales are computed. The process of estimating the "actual" texture scale is based on the changes in the magnitude and direction of the local edge vectors across scales. This method is similar to a soft version of local spatial frequency estimation. Three texture features, polarity, anisotropy, and scale, are extracted. Williams and Alder [144] use a mask for feature extraction. The mask consists of k*k blocks and each block is n pixels wide. Within each block, the average intensity, standard deviation of intensity, and average colour are computed. Within this framework, texture is implicitly extracted by the standard deviation of intensity within each block and the spatial distribution of colour within the mask. Liu and Picard [65] have investigated the Wold random field model for modelling texture. The Wold model decomposes the image into three mutually orthogonal components which can be described as periodicity, directionality, and randomness. These three properties correspond to the three most important perceptual dimensions identified by Rao and Lohse, [99].

4.2.3. Texture Representation

As discussed in the review papers, not a single representation scheme can be identified as the clear winner that can perform consistently well on all test images. Hence, it is not clear how to select a particular scheme for general image segmentation. However, since the segmentation results are usually judged by a human, it would be desirable to have the texture representation scheme that most closely resembles the human visual system. In particular, Gabor filters have proved to model sufficiently the psychophysical data obtained in texture discrimination experiments [22] [55]. Moreover, Gabor filters have some desirable optimality properties. They attain maximum joint resolution in the space and frequency domains [23]. This property is highly valuable in balancing the conflicting objectives of accurate estimation of texture features in the frequency domain and good spatial localisation. Hence, Gabor filters are selected to represent texture. Transformation on this texture space to simulate the orientation invariance and perceptual uniformity will also be discussed in the following sections.

4.2.4. Gabor Filter Bank

A 2-D Gabor function can be defined as a complex sinusoid modulated by a 2-D Gaussian function in the spatial domain. Thus, Gabor functions are complex-valued functions in $\Re^2$. However, some techniques use real-valued, even-symmetric Gabor filters only [**53**]. A family of 2-D Gabor function $g(x, y)$ and its Fourier transform $G(u, v)$ are characterised by the following formulas [**69**]:

$$g(x, y, \theta, \sigma, f) = \frac{1}{2\pi\sigma_x\sigma_y} exp^{\left(\frac{1}{2}\left[\frac{x_\theta^2}{\sigma_x^2}+\frac{y_\theta^2}{\sigma_y^2}\right]+2\pi j f x_\theta\right)} \tag{4.18}$$

$$G(u, v, \theta, f) = exp^{\left(\frac{1}{2}\left[\frac{(u_\theta-f)^2}{\sigma_u^2}+\frac{v_\theta^2}{\sigma_v^2}\right]\right)} + exp^{\left(\frac{1}{2}\left[\frac{(u_\theta-f)^2}{\sigma_u^2}+\frac{v_\theta^2}{\sigma_v^2}\right]\right)} \tag{4.19}$$

$$x_\theta = xcos(\theta) + ysin(\theta)$$

$$y_\theta = -xsin(\theta) + ycos(\theta)$$

$$u_\theta = ucos(\theta) + vsin(\theta)$$

$$v_\theta = -usin(\theta) + vcos(\theta)$$

where $\sigma_u = 1/2\pi\sigma_x$ and $\sigma_v = 1/2\pi\sigma_y$. θ is the orientation of the Gabor kernel, σ_x and σ_y control the width of the Gaussian envelope and f is the frequency of the sinusoidal waveform. The frequency and orientation selective properties of a Gabor filter are more explicit in the frequency domain as shown in equation(4.19). Figure 4.4 shows the real and imaginary parts of a Gabor filter with $\theta = 0$, a wavelength of 5.3 pixels, and unity aspect ratio ($\sigma_x = \sigma_y$). The frequency response of the filter is also shown on the same figure.

4.2.4.1. Parameter selection

Due to the fact that Gabor wavelets are not orthogonal, some information in the filtered images is redundant and some of the original data may be lost. Hence, the design objective is to utilise the smallest number of Gabor filters to cover approximately the whole feature space. This objective can be achieved by having the half-peak magnitude of the filter responses in the frequency domain touch each other. As in [**53**] [**69**], the half-peak radial frequency bandwidth, B_r, and orientation bandwidth, B_θ

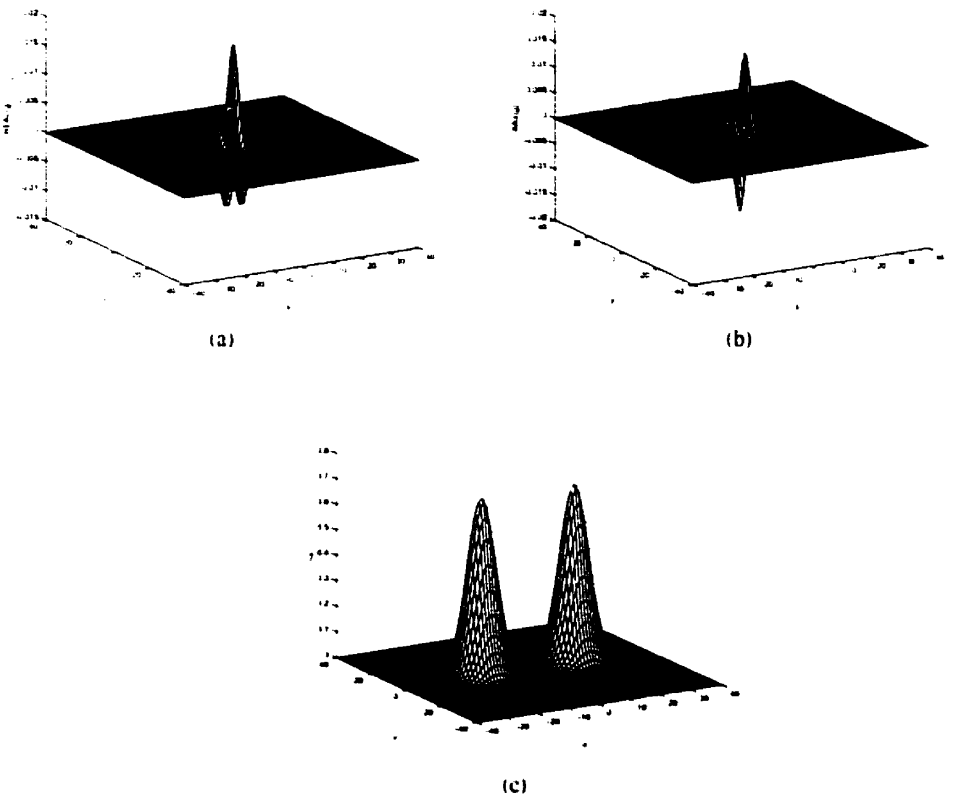


FIGURE 4.4. (a) Real and (b) imaginary components of a Gabor filter with a wavelength $(1/f)$ of 5.3 pixels and unity aspect ratio. (c) Frequency response of this filter.

are given by

$$B_r = log_2\left(\frac{f+\sigma_u\sqrt{2ln2}}{f-\sigma_u\sqrt{2ln2}}\right) \tag{4.20}$$

$$B_\theta = 2\ tan^{-1}\left(\frac{\sigma_v\sqrt{2ln2}}{f}\right) \tag{4.21}$$

where B_r is in octaves and B_θ is in degrees. If the frequency of two consecutive scales are f_1 and f_2. the required bandwidth. B_r is then given by $log_2(f_1/f_2)$. Once the highest radial frequency (f_0) and the scaling factor of the kernels (f_0/f_1) are fixed. the width of the Gaussian function (σ_x and σ_y) can be obtained from equations(4.20 and 4.21).

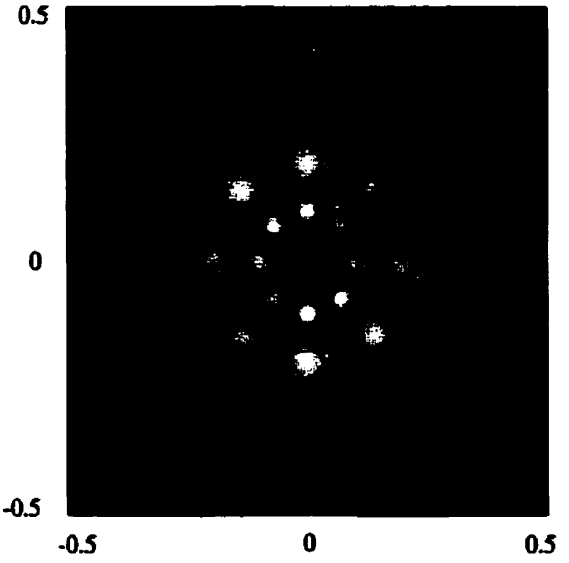


FIGURE 4.5. The frequency response of a dyadic bank of Gabor filters with 3 scales and 4 orientations.

When implementing a Gabor filter bank, it is necessary to choose the number of scales (wavelengths) and orientations. This determines the total number of channels in the filter bank. Randen and Husøy [98] found that the performance of texture classification increases with the number of features. The overall best texture feature representation in their study also has the highest feature dimensionality of 40. On the contrary, Smith [113] discovered that the algorithm with 3 scales and 4 orientations gave the best overall accuracy on 10 texture classification problems. He found that utilising a higher number of scales and orientations could have negative effects on performance. He called this observation the peaking phenomenon. The frequency response of the bank of 12 Gabor filters at 3 scales and 4 orientations is shown in Figure 4.5. This filter bank covers most of the frequency plane except for the low frequency range at the centre. For natural images, low frequency filters will pick up the structure of objects rather than the objects' texture. Hence, it is preferable to exclude the extremely low frequency filters.

One of the major issues in filter design relates to the efficiency of filter implementation. In the general form of the Gabor function, it is not a separable filter. This means a single convolution of a Gabor function and an image, with a size of $K \times K$ and $N \times N$ respectively, requires N^2K^2 multiplications and additions. One way to

reduce this computational workload is by reducing the redundancy of the Gabor decomposition using a pyramidal approach [**38**]. Because of the frequency selective property of the Gabor filter, the band-passed image can be sub-sampled without any loss of information. Hence, efficient methods, such as Burt's HDC method [**10**], can be used to down-sample the image before the convolution. However, this approach also limits the choice of subband decomposition to dyadic (octave band) decomposition. It should also be noted that the filtered images are smaller than the original image due to the sub-sampling. In order to generate a feature map at the highest resolution, up-sampling and interpolation is required.

An alternative solution to this problem is proposed in [**51**]. The Gabor function is decomposed into 2 separable functions. The requirement for this decomposition is to use a circular shaped rather than an elliptical shaped Gaussian function. Replacing both σ_x and σ_y by a single variable σ, the Gabor function in equation 4.18 can be expressed as a separable function as follows:

$$g(x, y, \theta, \sigma, f) = \frac{1}{\sqrt{2\pi}\sigma} exp^{\left(\frac{x^2}{2\sigma^2} - 2\pi j f x cos(\theta)\right)} * \frac{1}{\sqrt{2\pi}\sigma} exp^{\left(\frac{y^2}{2\sigma^2} - 2\pi j f y sin(\theta)\right)} \tag{4.22}$$

This filter is more efficient to implement than the direct implementation since convolving an $K \times K$ filter with an $N \times N$ image takes only $2K.N^2$ computations. Besides, unlike the pyramidal approach, there is no constraint on the scaling factor of the Gabor filter banks and no up-sampling is required as the filtered outputs already have the same dimensions as the original image.

4.2.5. Generation of Texture Feature Set

An overview of the generation of a texture feature set is shown in Figure 4.6. First, a set of Gabor filters is applied to the input image, generating n texture channels. These filter responses are then subjected to a series of linear and nonlinear transformations and smoothing to form the final texture feature maps.

4.2.5.1. Linear Transformation on Texture Space

For natural scene images, it is desirable that the texture features are invariant to rotation and scaling. For example, the stripes of the zebra in Figure 4.7a are at different orientations and scales. In order to have the zebra segmented out as a single region, the texture features must be insensitive to changes in orientation and scale.

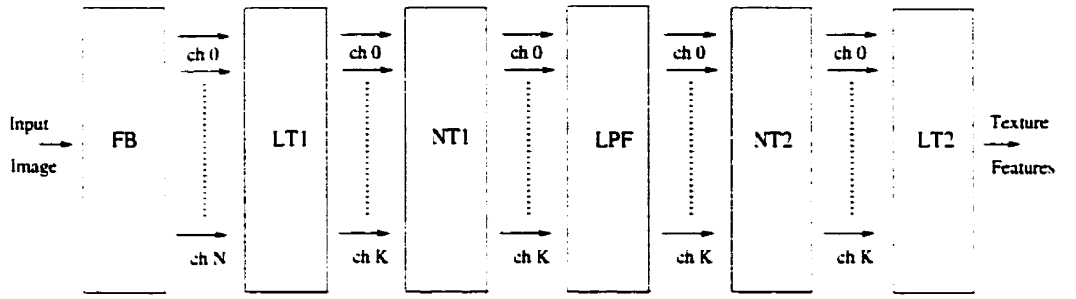


FIGURE 4.6. Block diagram of the generation of texture features. The filter bank (FB) generates N texture channels. The first linear transformation (LT1) approximates the orientation-invariance transformation, resulting in K channels where $K \leq N$. The next nonlinear transformation (NT1) and low-pass filter (LPF) produce a local energy estimation of the filter output. The second nonlinear transformation (NT2) is included to compensate for the effect of NT1 and the final linear transformation (LT2) improves the perceptual uniformity of the texture space.

One way to remove the orientation selectivity of the Gabor filters is by summing the filter responses of different orientations at each scale [**114**]. The resulting filter acts like a band-passed filter which can be modelled by Difference-of-Gaussian (DOG) filters. The magnitude of the Gabor outputs of the zebra image are shown in Figure 4.7. This test image explicitly shows the discriminative power of the Gabor filters on scale and orientation. The horizontal stripes are completely separated from the vertical and diagonal ones. However, the texture features of the zebra's body form several well-separated clusters. From the combined channels, (f) and (k), the shape of the zebra becomes more prominent and complete. It should be stated that combining channels of different orientations will lower the discrimination power since classification between two texture regions can no longer be based on the distribution of energy across different orientations. That means two textures are not distinguishable if their total amount of energy within each frequency channels is the same, regardless of their directionality (eg. mono-direction or bi-directions). Fortunately, this situation seldom happens in natural scenes.

4.2.5.2. Local Energy Measure

It is a common practice to use the local energies as the texture features, rather than directly using the output of the filters. This approach is understandable since the filter output of a sinusoidal signal will still be a sinusoid, see Figure 4.7, (b)-(f)

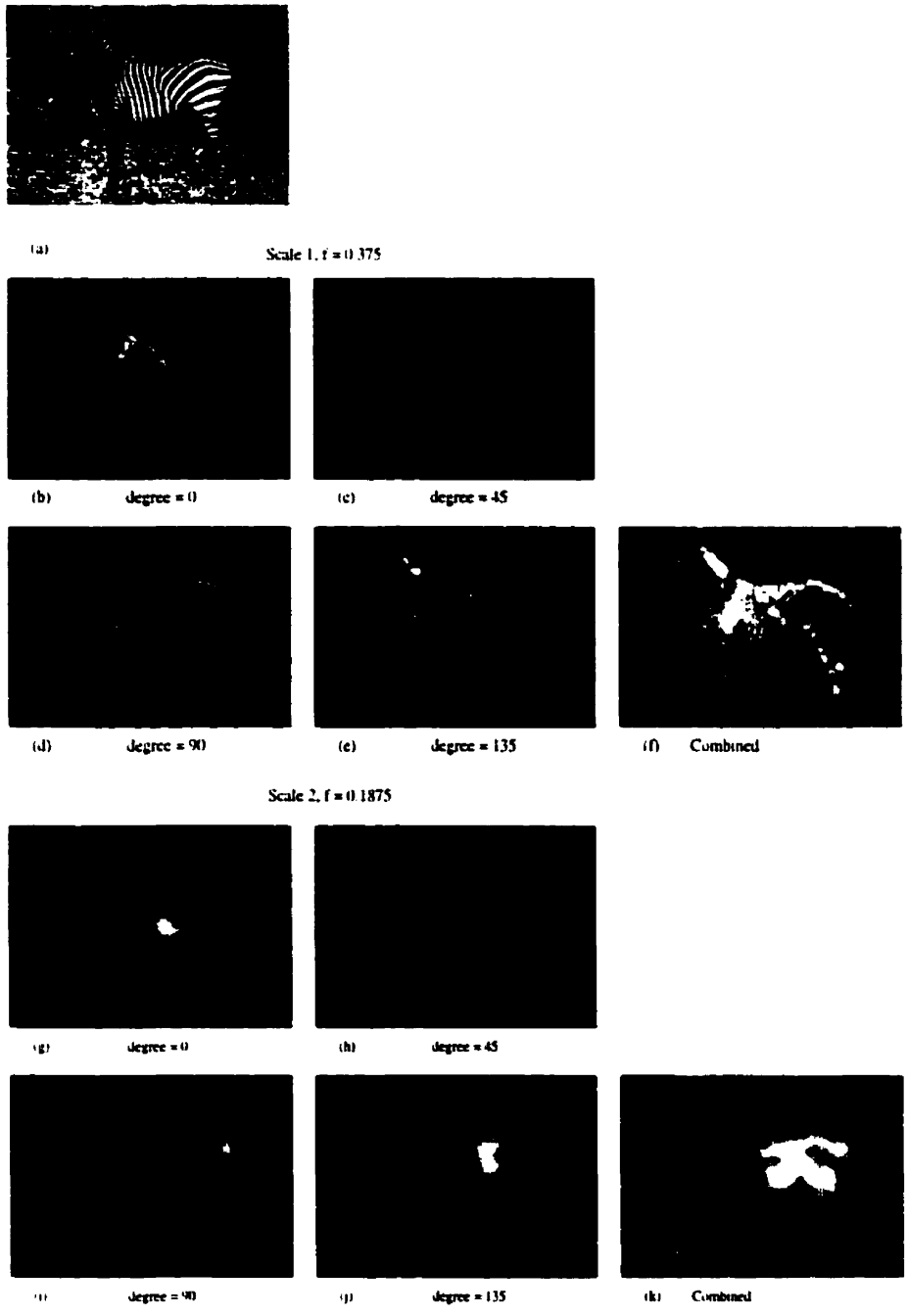


FIGURE 4.7. (a)A zebra image. Magnitude of different texture channels at 2 scales and 4 orientations :(b)-(e) capture the high frequency components of the image and (f) is the summation of (b)-(e). (g)-(j) capture the low frequency components and (k) is the summation of (g)-(j).

in particular. Hence, a local energy function, such as a Gaussian, rectangular, or circular function, is used to estimate the energy in a small local region. Among these functions, the Gaussian kernel clearly outperforms the other two functions because of its smooth transition from the centre to the boundary without any discontinuities. To achieve high edge localisation, a small neighbourhood is preferred. On the other hand, to achieve accurate energy estimation, a large local neighbourhood is required. As a compromise, the size of the filter will be set to a function of the radial frequency of the Gabor filter. A Gaussian smoothing function, $\sigma_s = 1/(2\sqrt{2}f)$ is used by Randen and Husoy [**98**] and $\sigma_s = 0.5/f$ is suggested by Jain and Farrokhnai [**53**].

In order to increase the feature distance between different textures while reducing the variance within each texture region, a nonlinear function is commonly applied before the smoothing. Commonly used nonlinearities are magnitude $|x|$, squaring $(x)^2$, and rectified sigmoid $|tanh(\alpha \cdot x)|$. To provide a feature value that is in the same units as the input signal, a second nonlinear function is applied. This function is an inverse of the first nonlinear function to counterbalance its effect. Different characteristics of these nonlinearities can be obtained by testing them on a test signal. Because of the band-limited property of Gabor filters, the filter output will contain a set of sinusoidal signals within the frequency bandwidth of the filter. The strength of these signals are usually not the same depending on their central frequencies and amplitudes. Hence, a test signal is created to simulate three different textured regions, for simplicity. These regions are two sine waves which differ in magnitude and a no-response region. Salt and pepper noise is added to the signal to simulate the randomness and uncertainty in real images. This test signal and the resulting local energies are shown in Figure 4.8. The saturation parameter, α, of the sigmoid function is set to 0.25 as suggested by Jain and Farrokhnia [**53**]. A larger value for this parameter will cause the signal to saturate more rapidly, causing the sine wave to become more similar to a square wave. From Figure 4.8b, comparing the fluctuations in the second region and the differences between the mean energy of the three regions, we can see that the sigmoid function produces the smallest intra-texture variation while squaring achieves the highest inter-texture separation. From experimentation, we have found that it is more important to have a larger inter-class distance than a lower intra-class variance. As a result, squaring will be used in the subsequent experiments.

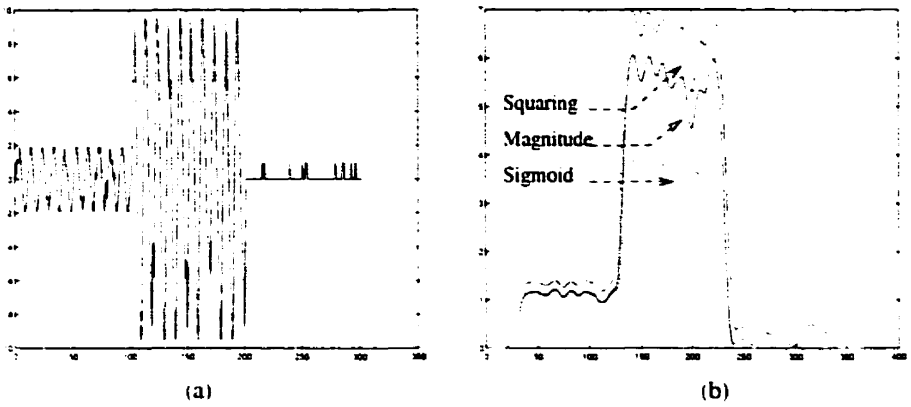


FIGURE 4.8. (a) Test signal, two sine waves with different magnitude and a no response region, with salt and pepper noise. (b)Local energy estimated by three different nonlinear functions: magnitude, squaring, and rectified sigmoid. $\alpha = 0.25$.

4.2.5.3. Perceptual Uniformity of Texture Space

Unlike the colour space, there is no generally agreed perceptually uniform texture space. However, it is still desirable to have a texture space that at least does not violate any obvious perceptual properties of texture. For example, a texture with a dominant direction at a high spatial frequency will be perceptually closer to a texture with a similar surface pattern at a lower spatial frequency than to a smooth non-texture region. If the orientation-invariance transform is performed, the number of texture channels will be reduced from 12 to 3, one channel per scale. The resulting texture space can be easily visualised in 3-D as shown in Figure 4.9a, where g_1, g_2, and g_3 correspond to the responses of the low, medium and high spatial frequency components. If one calculates the Euclidean distance between the three vertices, v_1, v_2,and v_3, of the triangle in Figure 4.9a, and the distance between these three points to the origin, it is clear that the distance is $\sqrt{2}$ between v_1, v_2, and v_3, and 1 between any of these points to the origin. This means that these three points are closer to the origin than to each other. The visual meaning of these four points is: v_1 has a unit amount of energy at low frequency, while v_2 and v_3 have the same amount of energy at medium and high spatial frequencies respectively. Obviously, the origin corresponds to a non-textured region. Although it is not clear how similar these four texture features are quantitatively, it would never be the case that a texture region like v_3 or v_2 is closer to a smooth region than a region like v_1 which contains a similar amount of energy. Hence, the objective of this transformation is to rectify this problem so

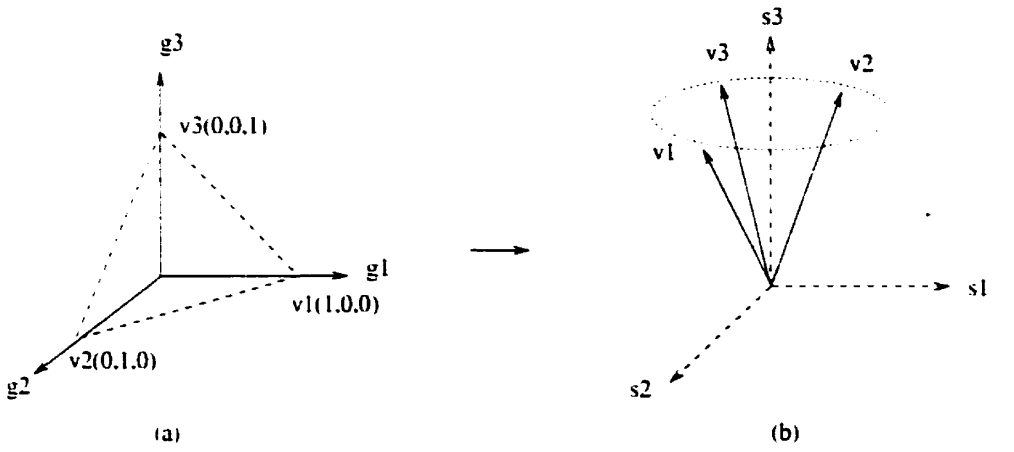


FIGURE 4.9. A transformation of the texture space is proposed to improve the perceptual uniformity. This transformation normalises the distance between the origin and the three vertices v1, v2, and v3 and the distance between these three vertices.

that the distance between any of these four points is the same. One linear transform that achieves this objective is as follows:

$$\begin{aligned} s_3 &= \frac{g_1 + g_2 + g_3}{\sqrt{3}} \\ s_2 &= \beta \times (g_1 - 0.5 \times (g_2 + g_3)) \\ s_1 &= \beta \times \sqrt{0.75}(g_2 - g_3) \end{aligned} \tag{4.23}$$

where β is a weighting factor that controls the relative importance of scale differences in the new horizontal plane, s_1, s_2, versus the differences in total amount of energy, s_3. This transformation is a combination of rotation and scaling (see Figure 4.9b). After the transformation, the three vectors become:

$$\begin{aligned} v_1(s_1, s_2, s_3) &= (\frac{1}{\sqrt{3}}, \beta, 0) \\ v_2(s_1, s_2, s_3) &= (\frac{1}{\sqrt{3}}, -0.5\beta, \sqrt{0.75}\beta) \\ v_3(s_1, s_2, s_3) &= (\frac{1}{\sqrt{3}}, -0.5\beta, -\sqrt{0.75}\beta) \end{aligned} \tag{4.24}$$

To determine the value of the parameter β, one can set the distance between v_1 and the origin and the distance between v_1 and v_2 in the new feature space to be the same. After simple manipulation, the value of β is found to be $\sqrt{1/3}$. To compress further the distance between v_1, v_2, and v_3, a smaller value for β can be used.

4.3. Feature Integration

After extracting features for colour, texture, and position, they must be combined to form a single feature vector. There are several issues that need to be addressed before this can be achieved. The first issue is the dependency of the three sets of features. In fact, the colour and texture of any region are highly correlated. A non-zero vector in texture space implies that the surface colour in the local neighbourhood is not uniform but varying, either randomly or in a regular pattern. Hence, a uniform textured region will not be uniform in colour space. In order to have a textured region remain intact after segmentation, the colour and texture features of the pixels within this region must form a well-separated single cluster. This can be done by replacing the colour with the average computed from a local region. The size of this local region should be proportional to the scale of the texture. The straightforward way to estimate the texture scale is to locate the scale which contains the largest amount of energy. However, this method limits the resolution of scale to the number of frequency bands used for texture extraction. To increase this resolution without increasing the number of filters, interpolation can be used. Let e_1, e_2, and e_3 be the amount of energy at three scales and λ_1, λ_2, and λ_3 be the wavelengths of the corresponding texture channel. Then, the scale, s, can be estimated using the following formula:

$$s = \frac{e_1\lambda_1 + e_2\lambda_2 + e_3\lambda_3}{e_1 + e_2 + e_3} \times min(1.0, \frac{e_1 + e_2 + e_3}{st}) \tag{4.25}$$

where the first term is the estimate of s, and the second term is the confidence of this estimate. When the magnitudes of e_1, e_2,and e_3 are small, such as in a uniform region, the scale of the texture is meaningless. Hence, the sum of e_1, e_2, and e_3 can be used as a measure of the confidence of the estimation. The constant, st, controls the saturation of this measure. The estimated scale of the image in Figure 4.7 is shown in Figure 4.10.

The second issue in feature integration concerns the dynamic range of each feature and their relative importance in perceptual grouping. Depending on the feature extraction method, the dynamic range can vary dramatically. For example, if *RGB* colour space is used for representing colour, the dynamic range of each colour channel is 0 to 255. However, if the *Lab* colour space is used instead, the dynamic range is 0 to 100 for L, -500 to 500 for a, and -200 to 200 for b. As a result, the features

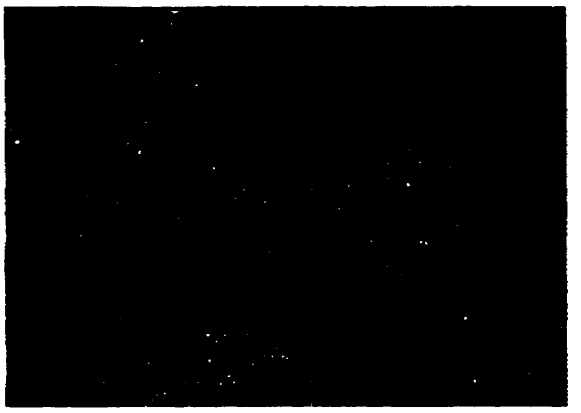


FIGURE 4.10. Estimated texture scale of the image in figure 4.7. Brighter regions indicate larger scales.

must be normalised so that different features (colour, texture, and position) all have the same variance and can be compared directly. It would also be desirable to scale the dynamic range of each feature so that the perceived difference of two regions which differ by one unit in any dimension of feature space would be the same. Hence, each feature is scaled by a weighting factor, which represents both normalisation and scaling, before the integration. Since no perceptual theory exists regarding how to select these parameters, these weights will be determined empirically. The final feature vector is formed as follows:

$$f(x, y) = [w_c c_1, w_c c_2, w_c c_3, w_t t_1, w_t t_2, \ldots, w_t t_k, w_p p_1, w_p p_2] \tag{4.26}$$

where w_c, w_t, and w_p are the weights for colour, texture, and position, respectively, and (c_1, c_2, c_3), $(t_1, t_2, \ldots t_k)$, and (p_1, p_2) are the coordinates of colour, texture and position respectively.

CHAPTER 5

Image Segmentation

Segmentation is a process of partitioning a digital image into disjoint connected sets of pixels, each of which corresponds to an object or region in the spatial domain. The division of an image into regions is based on criteria such as similarity and proximity, such that each region is *homogeneous* and no union of any two regions is *homogeneous* with respect to the same criteria. Image segmentation is a very critical component of an image processing system because errors at this stage influence feature extraction, classification, and interpretation. Therefore, image segmentation has long been an active research topic in image processing since the early 70's [**9**]. Despite a vast amount of research, the performance of even the most state-of-the-art techniques are still less than satisfactory and cannot be regarded as general purpose. In this chapter, a brief review of existing techniques on image segmentation is given. Issues concerning the implementation of the selected segmentation method will also be discussed.

5.1. Review of Image Segmentation Techniques

In general, image segmentation techniques can be classified into four major classes: clustering-based, edge-based, region-based, and hybrid methods. Clustering-based methods refer to groupings that are done in measurement or feature space, while edge-based and region-based methods refer to groupings that are done in the spatial domain of the image. The main difference between an edge-based and a region-based method lies in the different segmentation criteria. In a edge-based method, the segmentation process is based on spatial discontinuity. On the other hand, in a

region-based method. it is based on spatial similarity among pixels. Hence. region-based methods are the logical dual to the edge-based methods. The last category. hybrid methods. are combinations of one or more of the first three methods which take advantage of their strengths and minimise their weaknesses.

5.1.1. Clustering-based Methods

Clustering is a type of classification imposed on a finite set of objects or datum points. Each object is classified to one of the cluster labels depending on its relationship to other objects. This relationship can be represented by a proximity matrix or distances between objects in a d-dimensional space. A brief review of approaches that have been applied to image segmentation is given below. For more detailed descriptions. readers are referred to [**52**].

5.1.1.1. K-means

This "classical" method is probably the best-known and most widely-used for clustering data. If the clusters are well separated. a minimum-distance classifier can be used to separate them. In this method. the means of k clusters are estimated by a recursive labelling and updating procedure. First. an initial guess of the number of clusters and their means must be provided as input to the classifier. One popular method for obtaining the means of the k clusters is by randomly selecting k samples from the data set as an initial guess. Next. a minimum distance classifier is used to classify the objects into one of the k clusters. After the labelling. the means of the clusters are replaced by the centroids of the new resulting clusters. This process is repeated until no changes are made to any object in a given cycle. The method is very simple and works well for large and well-separated data sets. Unfortunately. this method also has a number of disadvantages. First. the number of clusters must be known in advance. which itself is a very difficult problem. This algorithm may also not converge to the real cluster centre if the clusters are unbalanced or elongated clusters are involved and the result produced depends on the initial values of the means. Recently. modifications to this method have been proposed to improve its robustness and efficiency. such as fuzzy k-means and sequential k-means [**93**].

5.1.1.2. Density Estimation

Another popular approach to clustering is to estimate the underlying density of the datum points and to allocate each point to one of the identified populations. If the

form and number of underlying population densities can be determined in advance. parametric density estimation methods can be used. Otherwise. non-parametric density estimation methods should be used instead.

One commonly used density model for parametric density estimation methods is the Gaussian density function and the underlying densities are assumed to be a mixture of g Gaussian densities [13]. If this assumption holds. and a rough estimation of the number of clusters or classes is available. then the parameters of the population densities can be estimated from the data by maximising the likelihood of the parameters. A number of techniques. such as the Expectation-Maximisation algorithm. can be used to obtain the optimum solution. The major drawback of this method is the assumption about population densities which limits its application. For natural scenes. this Gaussian assumption does not seem to hold for most situations.

Without any assumptions about the distribution of datum points. non-parametric methods are based solely on the notion that clusters are regions of feature space having high density and separated by regions of low data density. The probability density estimate at a point x is determined by a weighted summation of datum points falling within a small region around x. Clusters are then identified by locating local density-maxima. Since there is no need to specify in advance the shape and number of the clusters (determined from the number of local maxima). this approach is more general and can be used to identify any unknown or irregular shaped clusters.

5.1.1.3. Pairwise Data Clustering

Sometimes the characteristics of a data set cannot be represented in a metric space. Instead. they are characterised indirectly by pairwise comparisons as in a proximity matrix or graph. Advantages of pairwise comparisons over distance in metric space include the support of higher level similarity that violates the triangular inequality [105] [5]. However. techniques for finding the optimum partition or merging among the datum points based on the more general similarity matrix are. in general. less efficient and require more memory storage [110] [97]. For example. the proximity matrix of a small image of size 128x128 has $n^2 = 268.000.000$ entries.

5.1.2. Edge-based Methods

Segmentation can be obtained by detecting the boundaries of various regions. This task is usually accomplished by locating points of abrupt change in local features.

such as intensity, colour, or surface texture. A large variety of edge-detection methods are available in the literature, such as the Sobel, Prewitt, Roberts, and Canny edge operators. However, since the edges are often broken, edge linking is required to ensure that the boundaries form closed contours. Because of the small spatial support of the edge detector, the edges are very close to the actual boundaries. However, due to the same fact, this operator is very susceptible to noise and false edges can appear in highly textured regions. Ma and Manjunath [**68**] have proposed a novel boundary detection scheme, which they called "edge flow", to facilitate the integration of different image attributes for edge detection.

5.1.3. Region-based Methods

Region-based methods are the logical dual to the edge-based methods. Instead of locating changes in surface properties, region-based methods detect the homogeneous regions directly, usually by iterative split and merge phases. Unlike the edge-based methods, a measure of region homogeneity must be defined in advance. In general, available approaches for the task can be divided into two groups, region growing and split-and-merge. In a region growing approach, a number of uniform regions (seeds) are given a priori and the surrounding pixels are merged to one of these seeds (region growing) if the uniformity criteria are satisfied. For split and merge methods, non-uniform regions are broken down into smaller areas until all the resulting regions are classified as "uniform" based on the uniformity criteria. Next, neighbouring regions are compared and merged if they are close enough in feature space. In all cases, the quality of the segmentation output is directly related to the uniformity criteria, and hence the selection of a good uniformity measure is vital for success. Recently, Deng et al. [**26**] introduced a new measure for homogeneity, called the *J measure*, which measures the uniformity of colour distribution in a local region. By doing this, colour-texture patterns are incorporated into the homogeneity measure and thus no explicit texture feature extraction is needed. In general, region-based methods are more robust than edge-based methods because segmentations are based on much larger local neighbourhoods. However, according to uncertainty theory, this approach also has poor boundary localisation.

5.1.4. Hybrid Methods

Each approach mentioned in the previous sections has both advantages and drawbacks. Hence, it is desirable to combine some of the existing methods, making use of each approach's advantages. Because of the duality property of edge-based and region-based methods, these methods are commonly combined [**15**] [**92**] [**6**]. Zhu and Yuille [**150**] have proposed a method called "region competition" to unify existing techniques such as snake/balloon models, region growing, and Bayesian/MDL (minimum description length) within a statistical framework. Nazif and Levine [**80**] have proposed a rule-based approach which systematically organises and applies a large number of different heuristics for low level image segmentation.

5.1.5. Conclusions

Each method has its own advantages and disadvantages. Edge-based methods achieve good localisation but are sensitive to noise. On the other hand, region-based methods are more robust but at the expense of poorer edge localisation. Although hybrid methods produce the best segmentation results, these approaches are, in general, more complex and computationally expensive. Also, since the objective of this thesis is focused on real scenes, it is preferable to select a method which imposes a minimum number of assumptions on the image formation and the form of the underlying populations. Among the methods mentioned above, non-parametric density estimation satisfies the minimum assumptions requirement. It also provides feature density information that is needed for the ensuing attention process. Thus, this method is used for segmenting real scenes in this work.

5.2. Non-parametric Density Estimation for Image Clustering

The method described here follows the works in [**91**] and [**20**]. Non-parametric clustering starts with the estimation of the density. Let $\{X_i\}_{i=1,\ldots,n}$ be a set of n datum points in the d-dimensional space. Then the multivariate density estimation at a point $\mathbf{x}$ is defined as:

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{5.1}$$

where h is the radius of the density estimation kernel and $K(x)$ is the density estimation kernel.

The optimum kernel yielding minimum mean integrated square error (MISE) is the Epanech-Nikov kernels[**112**]:

$$K_E(x) = \begin{cases} \frac{(d+2)(1-x^T x)}{2c_d} & if \quad x^T x < 1 \\ 0 & otherwise. \end{cases} \tag{5.2}$$

where c_d is the volume of the hypersphere. Other types of kernels, such as linear and Gaussian are also frequently used.

5.2.1. Clustering Algorithm

The steps for the clustering are described below:

- *Generate a random sub-sample of the datum points.* To speed up the computation, a set of m points $X_1....X_m$ is randomly selected from the data. Moreover, in order to reduce outliers and "invalid" datum points, pixels lying on the regions of abrupt changes in spatial domain are excluded from the sample set.
- *Estimate the local density of each point in the sample set and then apply the gradient-ascent or hill-climbing method to locate the local maxima.* For each sample point X_i, equation(5.1) is used to estimate the density at X_i. k nearest neighbours of each data point are also determined. The gradient ascent method is used to associate each data point to a nearby density maximum by moving along the point of highest density among the k nearest neighbours.
- *Merge nearby cluster centres.* Any pair of cluster centres whose distance is less than a threshold will be merged. If no significant valley exists between any two cluster centres, these clusters will also be merged.
- *Re-classifying the samples.* Each sample point is relabelled to the cluster defined by a majority of its k nearest neighbours. Fewer nearest neighbours can be used if small clusters are expected.
- *Hierarchical clustering.* After the cluster centres are found, they are merged together hierarchically. The criterion for this merging process is the inter-cluster distance. However, this criterion can produce undesirable results, such as merging two well-separated but close clusters before other well-connected clusters that have centres further apart in feature space. To avoid this problem, Pauwels and Frederix [**91**] have taken a different approach. First, the choice

of h (the width of the density estimation kernel) and k (the number of nearest neighbours) are set to result in an over-segmentation of the feature space. Then, the clusters are merged based on the ratio of densities at the saddle-point and the neighbouring cluster centres, thereby producing an ordered tree of clustering. They defined the saddle-point as the point of maximal density among the boundary points which have neighbours in both clusters. Depending on the size of k, the estimation of the saddle-point can deviate from the actual boundary by the distance to the k^{th} nearest neighbour. To reduce this error, the boundary points can be further limited to points having at least 30% of neighbours in both clusters. The reason provided by the authors for using density instead of distance in the merging process is to avoid the unwelcome chaining-effect of hierarchical clustering. However, if distance information is ignored completely, the merging process will be vulnerable to error and noise in the density estimation, especially for small clusters. Hence, it is better to merge the clusters based on both density and distance. To make these two measures directly comparable, the distance is normalised by the average inter-cluster distance. Preference can be given to indicate the relative importance of density and distance. From experimentation, the best clustering results are achieved when the relative weights between density and distance are in the ratio of 10:1.

- *Selecting the optimum number of clusters.* At the last stage, the number of clusters is determined from indices of cluster-validity or an absolute threshold. This topic will be discussed in the next section.

5.2.2. Cluster Validity Indices and Stopping Criteria

Determining the number of clusters present in an image is a very difficult problem. This arises from the unclear definition of what is a *good segmentation*. For artificial images, it is easy to produce a definition since the ground truth of the image formation is known a priori. However, for natural images, obtaining the ground truth is not at all an easy task or may even be impossible. As discussed in Chapter 2, any image can be interpreted at different levels of abstraction and it may not be clear which level of abstraction is optimal for a given image. As a result, many image segmentation techniques rely on specific heuristics based on the application area, and the definition of a *good segmentation* is hard-coded into the program. Although heuristics are

widely used in a variety of fields, it is desirable to have a mathematical definition of a *good segmentation* so that it can be analysed systematically. In [**52**], a large number of indices of cluster validity are reviewed, such as the Davies-Bouldin index (DB) and the modified Hubert Γ index (MH). The problem with these indices is that they all are based on the assumption of Gaussian-shaped and well separated clusters. To overcome this problem, Pauwels and Frederix [**91**] have proposed a new non-parametric measure for cluster validity which does not exhibit any shape preference. To compare the performance and validity of different indices in image segmentation, three different methods are considered and analysed experimentally: a simple threshold-based index, the MH index, and the Pauwels and Frederix's non-parametric measures. The reason for selecting these methods is because they represent three major classes of cluster validity indices, from simple threshold methods to more complex indices both with and without any specific assumptions on the distribution of the data set. In the following, a brief review on these methods is provided and the analytical results will be presented in Chapter 6.

5.2.2.1. Threshold-based Index

Thresholds are very commonly used as stopping criteria because of their simplicity (no additional computations is required). However, in general, they require fine-tuning to optimise performance. This can be an advantage if it is easy to tune this parameter, or a disadvantage otherwise. Since hierarchical clustering is based on the density and distance between the clusters, a threshold on this measure can be used as a stopping criterion. Thus, clusters are merged if the following condition is satisfied:

$$density(i,j) + \rho \cdot distance(i,j) > \tau \tag{5.3}$$

where $density(i,j)$ is the ratio of the density at the saddle-point between cluster i and cluster j and the density at the cluster centres, and $distance(i,j)$ is the distance between these two clusters. ρ is a constant indicating the relative importance of density to distance and τ is the pre-defined threshold. From experimentation, we have found that the relative importance of density and distance is about 10:1 and thus a value of 0.1 is used for ρ.

5.2.2.2. Modified Hubert Γ Index

This index is proposed by Dubes [**52**] and is based on the assumption that estimates of the cluster centres are close to the "true" position of the clusters in the

pattern space and deviations from the centres are due to errors and distortions. Hence, there is an implicit assumption of ball-shaped clusters. For a given clustering, the MH index is defined as follows:

Let $L(i)$ be the label function.

$$L(i) = k, \quad \text{if pattern i is in the } k^{th} \text{ cluster}$$

and $d_{j,k}$ is the Euclidean distance between cluster centres j and k. Define

$$Y(i,j) = d_{L(i),L(j)}$$

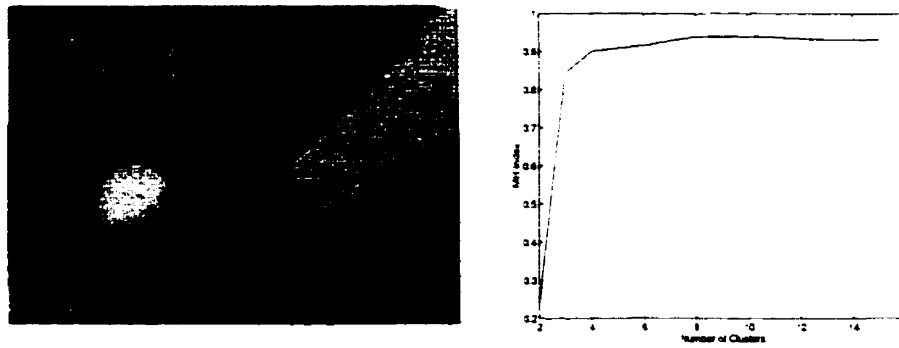
The modified MH index in then given by:

$$MH = \left\{ \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [X(i,j) - m_x] \times [Y(i,j) - m_y] \right\} / s_x s_y \tag{5.4}$$

where $X(i,j)$ is the Euclidean distance between pattern i and j, n is the total number of patterns, $M = n(n-1)/2$, and

$$\begin{aligned} m_x &= \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X(i,j) \\ m_y &= \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Y(i,j) \\ s_x^2 &= \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X^2(i,j) - m_x^2 \\ s_y^2 &= \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Y^2(i,j) - m_y^2 \end{aligned}$$

This index measures the degree of linear correspondence between the entries of X and Y. The matrix X is the same for all clusterings but the matrix Y varies depending on the corresponding cluster centres. For strong and well-separated clusters, the cluster centre associated with each data point should not deviate significantly from the true centre as long as the clustering is over-segmented. However, when the merging process exceeds the optimum level and tries to merge two well-separated clusters, the cluster centres will then start to deviate from the real centres and the similarity between the proximity matrices X and Y will begin to decrease. As a result, the optimum number of clusters is defined as the "knee" point of the MH function where



(a) (b)

FIGURE 5.1. (a) A simple image that contains roughly 5 different colours and (b) the MH index for this image.

sudden change occurs. As can be seen from the definition of this index, the *MH* index is computationally intensive ($O(n^2)$). Figure 5.1 provides an example of this index for a simple image.

5.2.2.3. Non-Parametric Cluster-Validity Indices

Pauwels and Frederix [91] introduced two non-parametric measures that quantify the notion of "good clusters" as a relatively well-connected region of high data-density. The first index, called the NN-norm, measures the average isolation of each cluster. This measure is based on the notion that similar patterns (close in feature space) should be assigned to the same cluster. This index is defined as:

$$NN-norm: \quad N_k = \frac{1}{n}\sum_{i=1}^{n} v_k(x_i) \tag{5.5}$$

where $v_k(x_i)$ is the fraction of the k nearest neighbours of feature x_i that have the same label as x_i. This index favours well-connected regions to be assigned the same cluster label. However, it cannot distinguish whether two well-separated clusters should be merged or not.

The second index, the C-norm, is proposed to compensate for the deficiencies of the first. This index is designed to give a high response when a given cluster is well-connected and a low response when a cluster contains two or more well-isolated regions. To achieve this, the average connectivity of any two points in the same cluster is measured based on the density at their midpoints: A high density midpoint implies good connectivity and vice versa for low density midpoints. This method is good for Gaussian-shaped clusters. For clusters whose shaped is curved, however, the

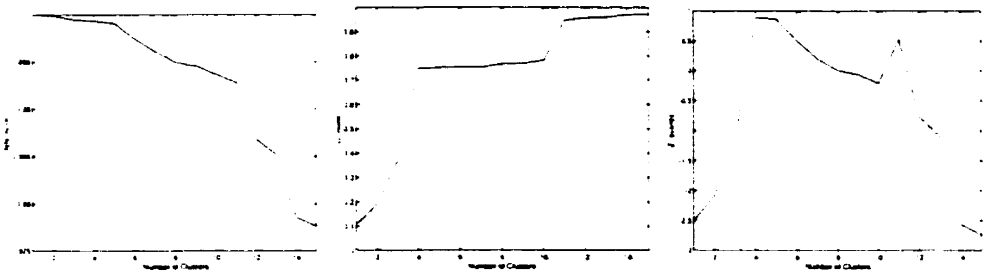


FIGURE 5.2. NN-norm (left), C-norm (center), and Z-score (right) for the image in Figure 5.1

mid-point of two randomly selected points can lie on the void between the arc. To rectify this problem, the midpoint is shifted towards the high density region until the local maximum is reached. During this shifting process, the same distances between the midpoint to the two test points must be maintained to avoid ending up at either one of the test points. This index is defined as:

$$C - norm: \quad C_k = \frac{1}{k}\sum_{i=1}^{k} f(t_i) \tag{5.6}$$

where K is the number of randomly chosen pairs of test points and t_i is the mid-point after the shifting process. $f(t_i)$ is the data density at the point t_i.

To select a single clustering, these two cluster-validity indices must be combined to give a single measure. Pauwels and Frederix propose first computing the Z-scores of the C-norm and NN-norm to make the indices directly comparable: the two resulting Z-scores are summed to give the final score, Z. The clustering having the maximum Z-score is selected as the optimum segmentation for the given image. The equations for computing the Z-scores and the final Z score is defined as follows:

$$Z(x_i) = \frac{x_i - median(x)}{MAD(x)} \tag{5.7}$$

$$Z_k = Z(C_k) + Z(N_k) \tag{5.8}$$

where MAD stands for median absolute deviation. Typical curves for the NN-norm, C-norm and the Z-scores are shown in Figure 5.2 for The disadvantage of using the median and the MAD to normalise the cluster measures is that their values depend on the range of valid clusters, or number of observations.

5.2.3. Post-processing

After the segmentation, mathematical morphology, dilation and erosion, are utilised to remove small and thin regions that usually correspond to noise. Next, three conservative region merging processes are applied to the segmentation result. First, regions that are smaller than 0.5% of the whole image are merged to their 4-connected or 8-connected neighbours If more than one neighbour is found, the one closest in feature space is selected. When position is also included in the feature vector, large regions may be split into two or more regions. Hence, a second step of the region merging is used to merge similar regions based on colour and/or texture only. In some images, the regions' surface features are not uniform but change smoothly (for instance, from light to dark, such as the sky). Hence, another merging process is carried out to merge regions whose contrast along their common borders are below a pre-defined threshold.

CHAPTER 6

Evaluation and Test Results

This chapter examines the performance of the object-based attention algorithm. There are basically three areas to be analysed. First, several important parameters that could not be determined using logical and theoretical arguments are evaluated experimentally. The second part will compare and analyse different methods for selecting the best number of clusters. The last part of this chapter will discuss the performances of different saliency factors in predicting the perceptual saliency of regions in a scene. The image database used in the experiments was chosen from the Corel image collection[1]. (See Appendix B for all of the images in the experimental database).

6.1. Determining Parameter Values

The parameters that need to be determined experimentally are the weights on the colour, texture, and position features in the feature extraction, and the sample size, kernel width, and number of nearest neighbours in the process of image segmentation.

6.1.1. Weights for Colour, Texture, and Position

As stated in Chapter 4, the purpose of imposing weighting factors on colour, texture, and position features is to normalise the dynamic range of different features and to improve the perceptual uniformity of the combined feature space. It would be preferable to evaluate the perceptual differences among these features through psychophysical experiments. However, this is beyond the scope of this thesis and no

[1] http://www.corel.com./products/clipartandphotos/photo/photolib.htm

appropriate literature is available on the topic. Alternatively, these weights can be determined by finding a parameter set that produces the best overall segmentation results.

Before applying any modification to the weights, the features are obtained as follows:

- Colour features are obtained by converting the *RGB* values of each pixel into $L^*a^*b^*$ space, with L ranging from 0 to 100, a^* ranging from -500 to 500 and b^* ranging from -200 to 200.
- Texture features are formed by applying a set of band-pass filters on the intensity, L. Next, the set of transformations described in Chapter 4 is applied.
- Position features are the x, y coordinates of the pixels normalised to the range of 0 to 1 by a scaling factor. To preserve the aspect ratio, the same scaling factor is used for both x and y coordinates. If the original x, y coordinates ranges from 0 to width and height, respectively, $1/max(width, height)$ can be used as the scaling factor.

6.1.1.1. Optimisation process for finding the weighting factors

Although a number of measures have been proposed for estimating the quality of a particular segmentation [**149**][**7**], they are not very accurate or effective when compared to human performance. In order to avoid extensive psychological experimentation and still have a subjective justification for the segmentation results, the following process was used for selecting the best parameter set to gives the best overall results:

- From a preliminary examination of the image segmentations, we found that, for a large portion of the database, the segmentation results did not vary significantly with different weighting factors. Only on a small subset of the database could we observe significant improvement by modifying the weights. Hence, in order to reduce the complexity of the optimisation process, only a small subset (about 50) of the database was employed, including all of the images that preferred a different parameter set from the majority of images in the complete database.
- Segmentation results using different weighting factors were obtained and judged by human observers. In particular, the judgement were based on the following criteria: 1. Grouping should be consistent with the visual appearance. No

visually distinct regions should be merged and vice versa for visually similar regions. 2. More emphasis should be placed on the major objects in the image rather than the background. 3. The overall quality of the segmentation results for a given parameter set were obtained by counting the number of images judged acceptable based on the first two criteria.

6.1.1.2. Results and Discussion

From extensive experimentation on a wide variety of test images, we have found that the weighting factors for colour, texture, and position should be approximately equal to 1, 1, and 10, respectively, to achieve the best results. It was observed that the inclusion of position in the feature vector has both advantages and disadvantages. The major advantage is that the proximity of pixels is also considered in the grouping process. On the other hand, this can be a disadvantage since the position information may cause an occluded object to form two or more clusters in the feature space. Fortunately, this problem can be solved easily by merging regions having similar colour and texture. For normal scene images where different objects form distinct clusters in feature space, the segmentation results do not differ significantly whether position is included or not. However, if two or more objects in a scene have similar surface properties, a much better result is produced if position is incorporated into the feature vectors. Generally, including position into the feature vector improves the separability of different regions and produces more compact and smooth regions, thus, yielding a better segmentation result. Figure 6.1 & 6.2 show the final segmentations of 30 randomly selected images.

6.1.2. Parameters Used in Image Clustering

There are three parameters in the clustering algorithm outlined in Chapter 5 that need to be set. The first one is the sampling rate, s. From the whole image, m pixels are randomly selected and used in the subsequent density estimation and clustering process, where $m = sN$ and N is the total number of pixels. The last two parameters to be determined are the width of the density estimation kernel, h, and the number of nearest neighbours, k, that are used in the gradient-ascent process.

6.1.2.1. Sample Size

For an image of size 180x120, the total computation time of the clustering algorithm and the time needed for density estimation at different sampling rates are

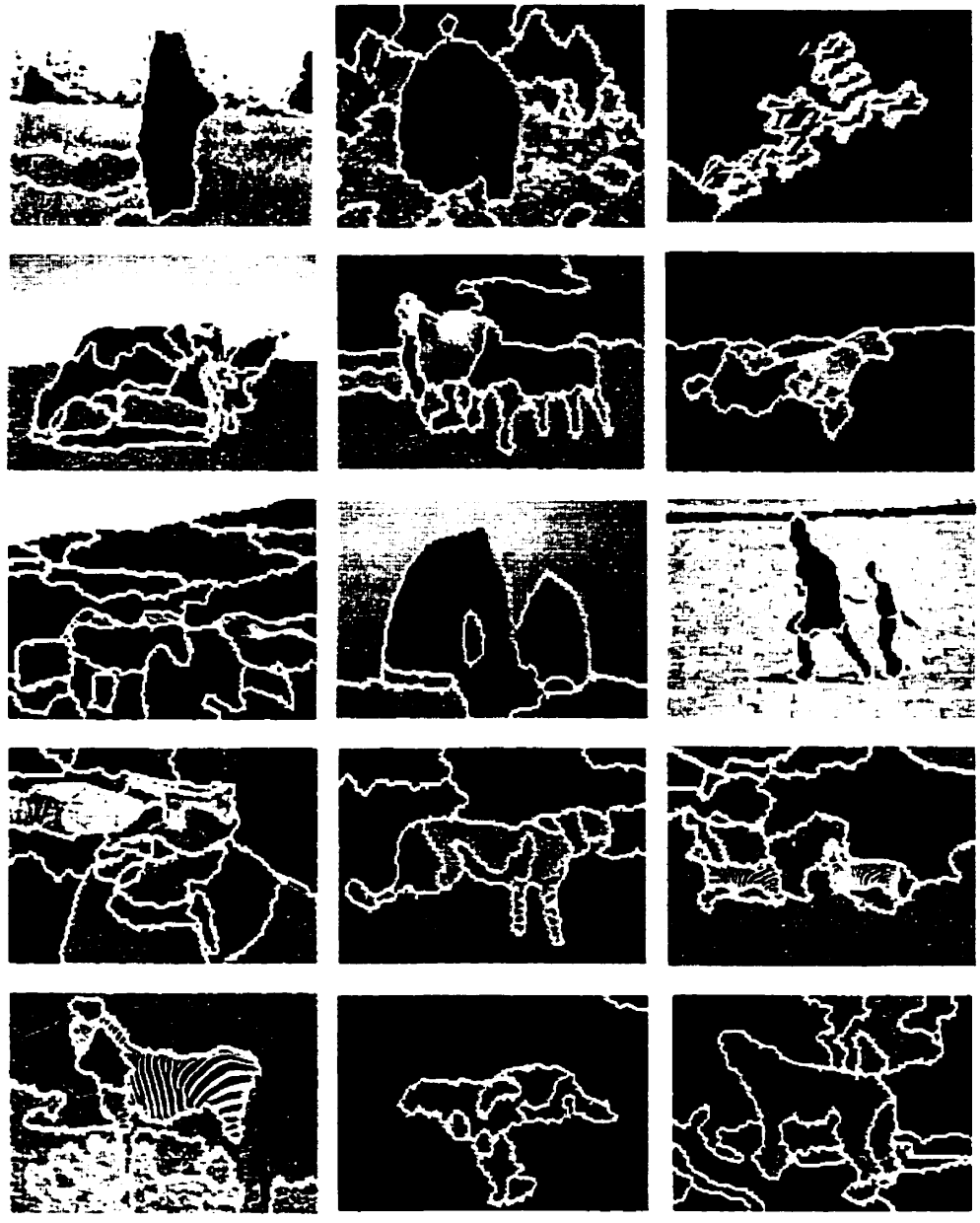


FIGURE 6.1. Part A. Segmentation of 30 randomly selected images. Boundaries are shown in gray. See figure 6.2 for the other 15 images.

shown in Figure 6.3. Clearly, the bottleneck of the clustering algorithm is the density-estimation process. By examining the density-estimation equation on page 63, we can see that this operation has a computation complexity of $O(n^2)$. This process takes 2.5 minutes on a 300 MHz Pentium II PC if 100% sampling is used, but only 35 seconds if half the data set are considered. Hence, it is desirable to analyse how much

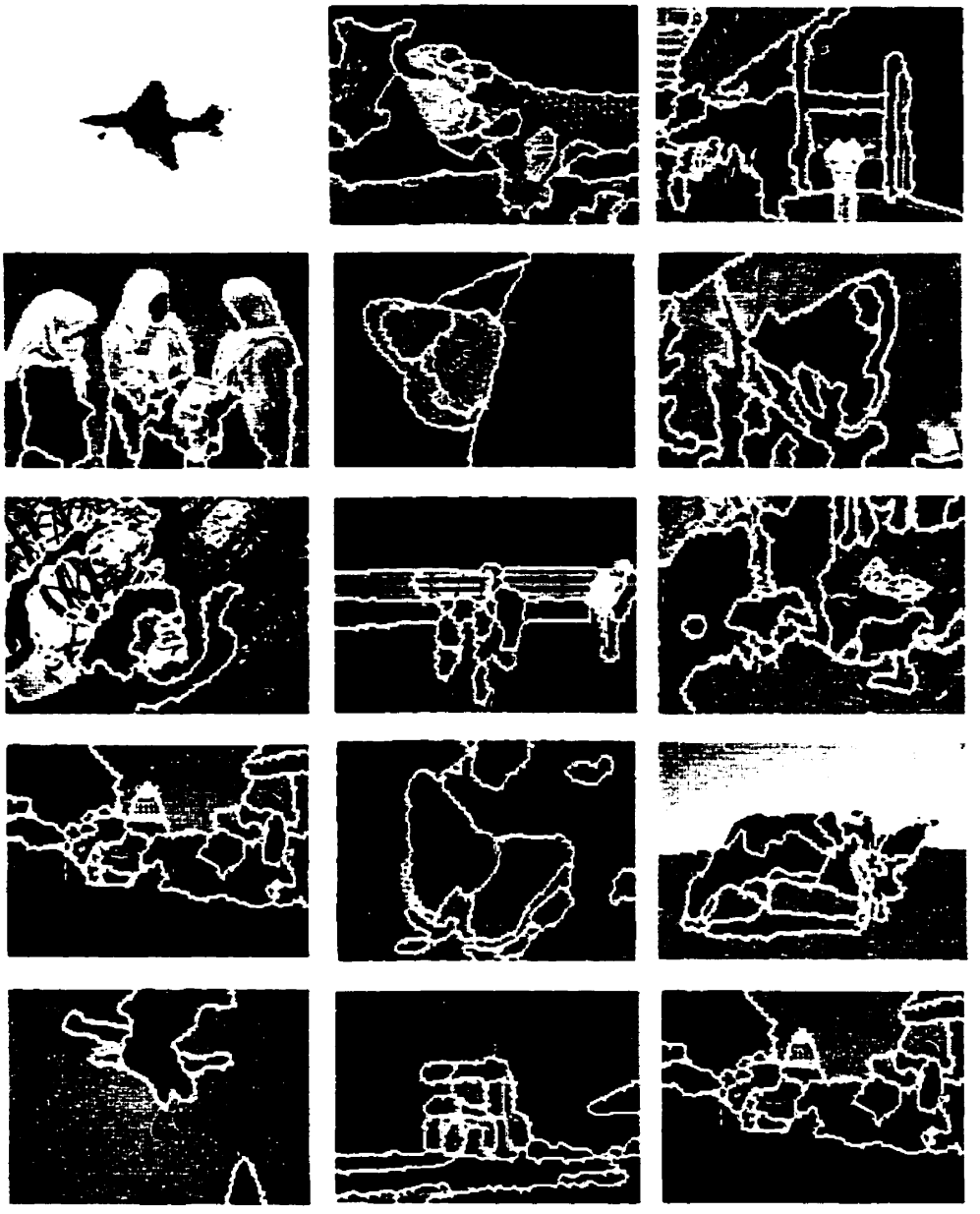


FIGURE 6.2. Part B. Segmentation of 30 randomly selected images. Boundaries are shown in grey. See figure 6.1 for the other 15 images.

segmentation error is introduced when the data set are sub-sampled. From an examination of the segmentation results of a wide variety of test images, there seems to be a general trend that the outputs are very similar for any sampling rate between 40% and 100%. Below this range, small objects begins to disappear and the boundaries start to deviate from their actual location. As a result, 40% of the whole image is

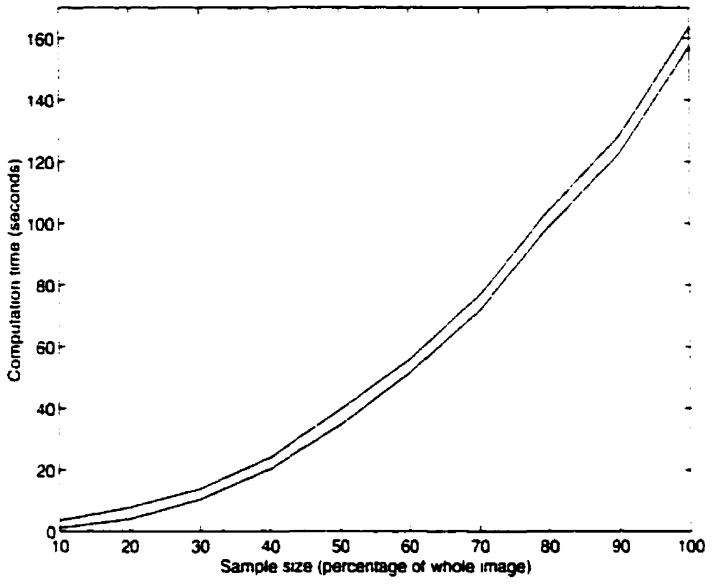


FIGURE 6.3. Computation time of the whole clustering algorithm (upper curve) and the time spent on the density estimation process (lower curve) at different sampling rates on a 300 MHz Pentium II PC

used to estimate the underlying feature distribution. The segmentation results for test images with sampling rates ranging from 10% to 100% are shown in Figure 6.4.

6.1.2.2. Kernel Width and Number of Nearest Neighbours

In determining the values for these two parameters, Pauwels and Frederix [**91**] have stated that the specific value of these two parameters is not critical as long as small values, with respect to the range of the data, are used. However, we have observed that the segmentation results are directly related to the specific values of these two parameters. The parameters can be interpreted as smoothing factors on the density of the data in feature space. A larger value for h and k will cause more clusters to merge, thus yielding fewer regions in the image domain. To avoid merging small regions, a smaller value for these parameters is preferred. However, if we wish to reduce the effects of noise and outliners, a larger value for h is preferred. For the images used in this experiment, we found that k equal to 0.4 percent of the total number of data points produced the best results without over-smoothing the density. Based on this kernel width, the number of nearest neighbours is selected as follows:

$$h = \sqrt{\sum_{i=1}^{N} dist(i,k)^2} \tag{6.1}$$

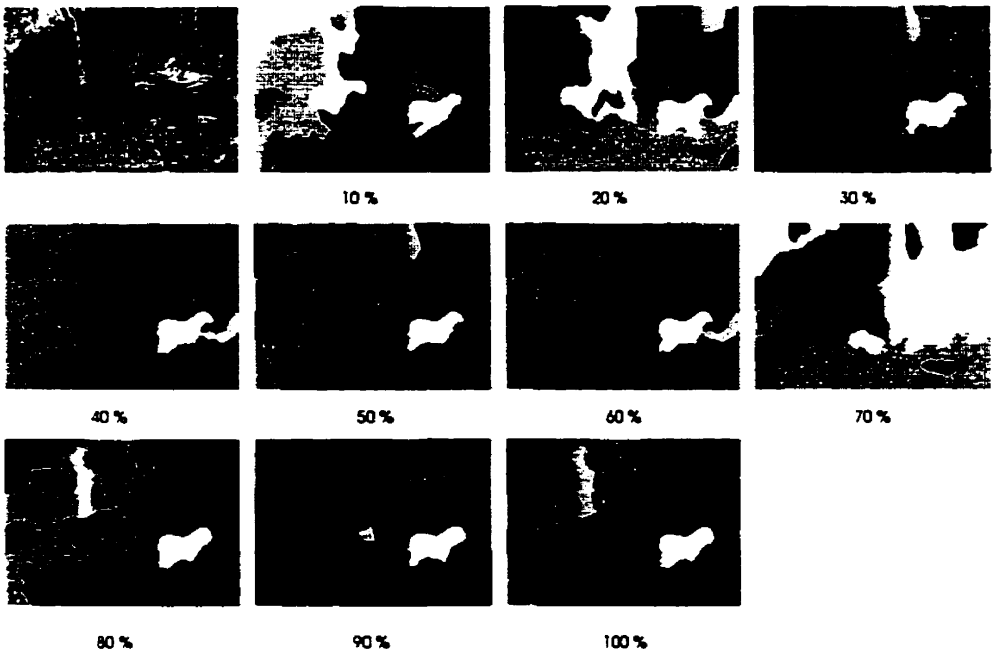


FIGURE 6.4. Segmentation results of a test image at different sampling rates

where *dist(i,k)* is the distance of the k^{th} nearest neighbour of point i in the feature space.

6.2. Cluster Measures

Although it is important to develop better techniques for feature extraction or grouping criteria, and which have a closer resemblance to the performance of the human visual system, it is equally important to explore new techniques for measuring the validity of different clusterings that usually arise in the many image segmentation techniques. The challenge of working with real scenes is that there may be more than one possible way to segment an image, and they may all result in valid segmentations. Hence, a natural question is what determines the validity of a particular segmentation and whether or not this definition can be formally defined in terms of mathematical formulas. In other words, how can we estimates the true or best number of clusters or regions for a given image? In Chapter 5, three different methods that are designed for measuring the cluster-validity are described: a threshold-based index, modified Hubert Γ index (MH), and Pauwels and Frederix's non-parametric measures (NP). In this section, the performances of these three methods on real scene images will be analysed and compared.

6.2.1. Assumptions Used in Each Method

Before explaining the test methods and results, it is useful to restate the assumptions used in these three methods. In a threshold-based method, an invalid clustering is defined as a violation of a pre-defined threshold (see equation(5.3)). Since, it is desirable to minimise the amount of over-segmentation, the optimum number of clusters is the one that is both valid and has the smallest number of clusters. The last two methods, *MH* and *NP*, are global measures that compute the overall goodness of a segmentation. Both methods are based on the notion that the clusters are well-separated in feature space. Hence, the performance of these methods may not be very reliable for weakly separated clusters. However, Gaussian distributions are assumed in *MH* but not in *NP*. Unlike the threshold-based method, the decision scheme of estimating the best number of clusters depends only on the changes of the indices (as a function of the number of clusters) but not on their specific values. The drawback of this kind of decision scheme is that a sudden transition or a "knee" in a function is often not easy to detect or define precisely. In addition, since it is not effective to search for all possible cases (the maximum number of regions will be the total number of pixels), for any given image, the search must be limited to a specific range. For instance, 1 to 6 clusters is used in [**91**] and [**13**]. Given this restriction, it is important to determine whether the ideal number of clusters lies on the boundaries of the search range or even outside this range.

6.2.2. Test Images and Implementation Issues

To test the robustness and the validity of the assumptions of the three methods, a set of 40 real scene images from the database in Appendix B was carefully selected to capture the variations in object size, contrast, and other properties present in real world scenes. Samples of these test images are shown in Figure 6.6 (See Appendix C for the whole test set and segmentation results). In this test set, we found that the best number of clusters can actually be as large as 15. Hence, the search range is set to $[1, \ldots, 15]$. For the threshold-based method, based on the criteria stated in section 1.1.1, a threshold of 0.5 gives the best overall result. Hence, the threshold (τ) is set to 0.5. For the *MH* index, the optimum number of clusters is defined as the "knee" point of the *MH* function. In actual implementation, the "knee" point is defined as the maximum in the second derivative of the *MH* function. Besides, since

the proximity matrices X and Y of a 180x120 image contains 467 million entries each. only 10% of the pixels are used for computing these matrices. For the NP method. since the definition and procedures for finding the cluster number are clearly defined. no extra assumption is required.

It may not be fair to compare a method that requires "training" to other methods that do not. Thus. if both methods achieve the same level of performance. the one that does not requires any "training" is preferred since it is more general. On the other hand. if a fixed parameter set can be used throughout the experiments. the threshold-based method could perhaps also be classified as an *unsupervised* method.

6.2.3. Test Results and Discussion

The performance of the three methods on the 40 test images can be summarised with reference to 8 images. The final segmentations selected by each method are shown in Figure 6.6. As expected. all methods are capable of selecting the optimum number of clusters when the clusters are well-separated in feature space. such as the aeroplane and the eagle. Although the head and the tail of the eagle are merged with the background in the segmentation selected by the *NP* method. the major objects are still clearly visible and separated. At the other extreme. such as images *C* and *D*. the important objects (the cheetah and the tree branches in *C* and the horses in *D*) are not well-separated from the background. Part or all of these objects are lost in the segmentation selected by the *MH* and *NP* methods. As a result. these methods should not be applied if weakly-separated clusters are expected. The threshold-based method. because the importance of these objects has already been considered in the selection of the threshold parameter τ. these salient objects are well separated in the segmentations selected by this methods.

Apart from the compactness assumption of the clusters. both the *MH* and *NP* methods also implicitly assume the existence of one and only one answer to the number of clusters. In addition. they also assume that the values obtained for the number of clusters is located in the middle of the search range. In reality. where nothing is perfect and noise is unavoidable. these assumptions cannot be guaranted to hold under all situations. From experimentation. we have observed that the *MH* and *NP* indices can have not only one but two or more knee points (see Figure 5.2). When this happens. it is not clear which knee point is the best description of the data distribution. On the other hand. if the "real" number of clusters lies outsides

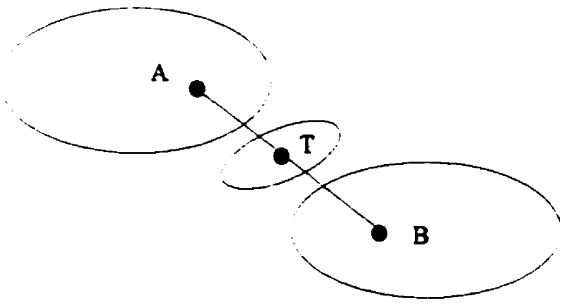


FIGURE 6.5. A situation where the C-norm in *NP* indices gives a wrong result.

the search range, no significant knee point will be found. If these two cases are not handled appropriately, arbitrary results will be returned.

In general, it is better to have an image over-segmented than under-segmented. However, it is not clear how much over-segmentation is acceptable and how this measure could be quantified mathematically.

The non-parametric indices proposed by Pauwels and Frederix[**91**] are supposed to perform equally well as the *MH* index on Gaussian-distributed clusters and perform better on irregularly shaped clusters. On the results of 40 test images, this claim does not seem to hold. In some cases, the segmentations picked by the *MH* index are better than the one selected by the *NP* indices. One possible reason for this observation is that the assumption of Gaussian distributions actually holds for most real images. We also found that the method used for measuring the connectivity in *NP* indices does not always give the true connectivity of a given cluster. A situation where this measure breaks down is illustrated in Figure 6.5. Suppose in a given clustering, all three clusters are merged and assigned the same cluster label and the two *anchor-points* for the C-norm are points A and B. Then the test point T halfway between the two anchor-points will fall on the high-density region. As a result, a high value for connectivity will be reported.

The time needed with 180x120 images to compute the *NP* indices and *MH* index (with a 10% sampling rate) are 22 seconds and 85 seconds on a 300 MHz Pentium II PC. For the threshold-based method, the only computation is equation (5.3). Since the inputs to this equation, density(i,j) and distance(i,j), have already been computed during the hierarchical clustering stage, the computation time for this equation is negligible. Among these methods, the clear winner is the threshold-based method.

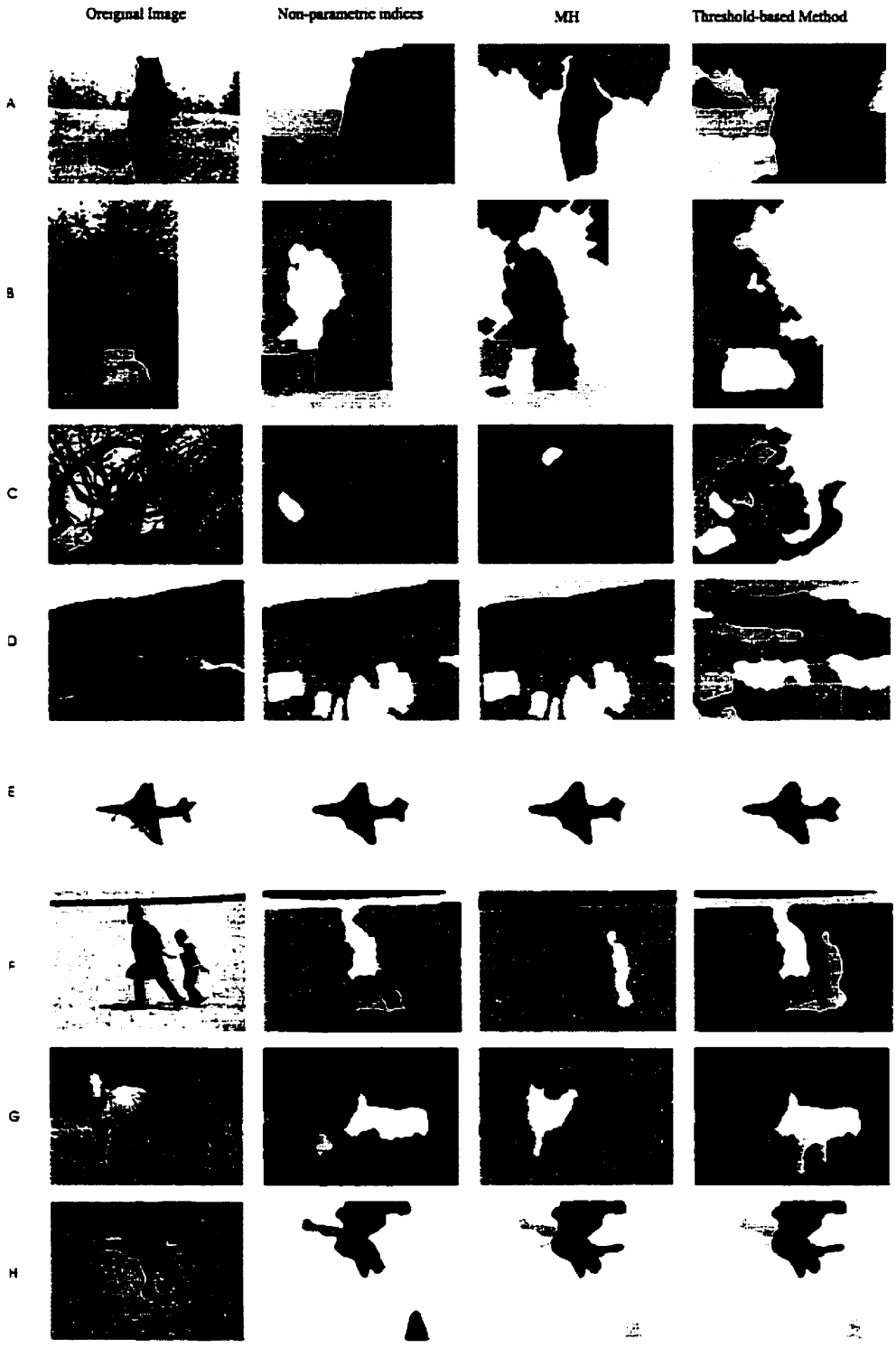


FIGURE 6.6. Samples of the test images and the segmentations selected by different methods: non-parametric indices (2^{nd} column), modified Hubert index (3^{rd} column), and the threshold-based method (4^{th} column).

It performs well on all test images and requires only simple comparisons. A minor drawback is that a suitable threshold must be known a priori.

6.3. Saliency Factors

Before being able to determine the contents of a scene, it is necessary to first focus attention on the most salient parts of an image. This entails an effective model of the human attention system and it is vital to the development of a powerful computer-based vision system. In this section, the region-based attention model described in Chapter 3 is analysed and evaluated.

6.3.1. Determining the Weights of Different Saliency Factors

Seven saliency factors are described in Chapter 3. These factors are: contrast, colour, location, size, foreground/background or depth, saturation, and shape. After considerable experimentation, we found that only the first five factors are useful for predicating the importance of a region. Saturation and shape factors are useful in some situations. However, their rates of failure are much higher than their success rates. As a result, they will not be considered in the subsequent experiments.

The final importance value is defined as a weighted sum of each factor as follow:

$$I.M(R_i) = \sum_{k=1}^{n} w_k \cdot I_k(R_i) \tag{6.2}$$

where w_k is the weight on the k^{th} factor, I_k, of region i.

Since the results will be judged finally by a human, a traditional trial and error method was used to determine the importance of different saliency factors in human visual attention. At present, no extensive psychological experiment has been conducted and the weights of the saliency factors were selected and judged solely by the author. If more time was available, these factors could be obtained more formally and reliably by having a group of subjects rank the relative importance of different regions in a set of test images. After obtaining these statistics, numerical methods or neural networks could be used to find the optimum weights by minimising the overall difference between the expected and estimated importance values.

From experimentation, it is found that the results closest to human performance were obtained with weights of 1.0 for foreground/background, 0.5 for contrast, and 0.3 for colour, location, and size. For the size factor, a saturation value of region

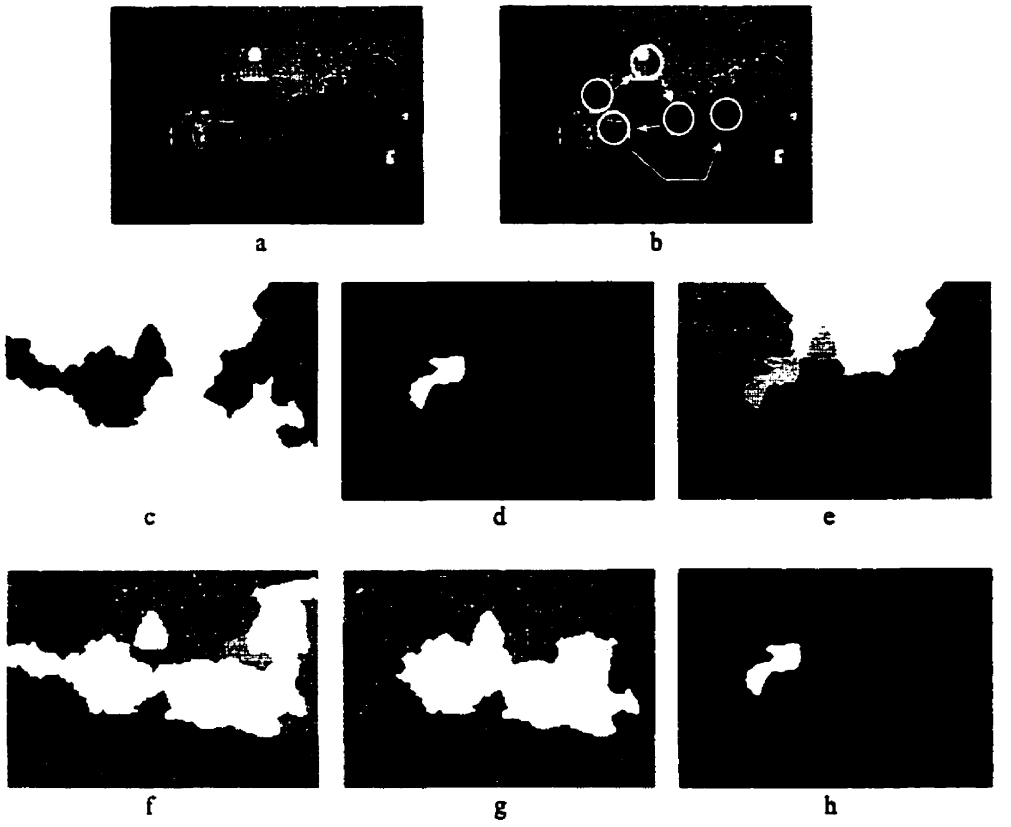


FIGURE 6.7. Importance maps for a sample image, (a). For (c)-(h), brighter regions represent higher importance. (c)size factor. (d)colour factor. (e)contrast factor. (f)foreground/background. (g)location factor. and (h)final importance map produced by weighted summation of (c)-(g). To facilitate the evaluation of the final importance map. the ranking of the top-five most important regions are highlighted in (b). Arrow directions indicate the next most salient regions.

size equal to 5% of the whole total image area is found to be better than 1%. The performance of these five factors and the final important values are illustrated in Figure 6.7. To indicate visually the ranking of these regions. the top-five important regions are highlighted in Figure 6.7b. For these images. the importance values predicated by the model are very consistent with the results obtained from a human. The most important objects. the caleche. horses. and the bright dome roof. are within the top-five regions. Moreover. the scan path generated from the importance map also agrees well with expected human performance.

6.3.2. Discussion

To test the robustness of this model, it was applied to 100 images with a fixed parameter set. Results of 16 images are shown in Figure 6.8. In general, the attention model gives consistently good results for a variety of images. As we can see from the weights comprising of the importance factor, the final importance values are highly biased to the foreground/background factor. Since the test images used follow conventional photographic techniques, the objects of interest are usually placed at the centre of the image. Hence, the probability that these objects touch the image border are much lower than the background. As a result, the foreground/background measure can separate the objects from the background quite accurately. However, if the object touches the border, such as the elephant at the bottom left of Figure 6.8, a false negative error occurs. In this case, the importance factor fails to predict the saliency of the elephant and it ranks the clouds as the most salient region in that picture. For some images, regions among the top-five ranks selected by the attention maps do not really respond to important objects, such as the sky, shadows, and the ground. In order to further refine the results, higher level reasoning and knowledge are required. Nevertheless, for a low-level system, the results are promising and the method is general enough to be used in many computer vision applications including content-based image retrieval.

6.4. Applications

This technique for locating salient "objects" in an image can be extended easily to handle a number of task-specific applications, such as face finding, image compression, machine vision, and CBIR.

6.4.1. Face finding

This problem is of significant interest in the field of computational vision, and has posed numerous practical challenges to date. For face finding, the importance of a face can be encoded into the weights factors of the importance factors. For discriminating face from other objects, skin colour (hue) and shape (roughly circular or elliptical) can be used. The roundness of a region can be obtained by measuring the ratio of area to edge length. In figure 6.9, a test image and its importance map is shown. In this experiment, only two importance factors are used, colour (red) and

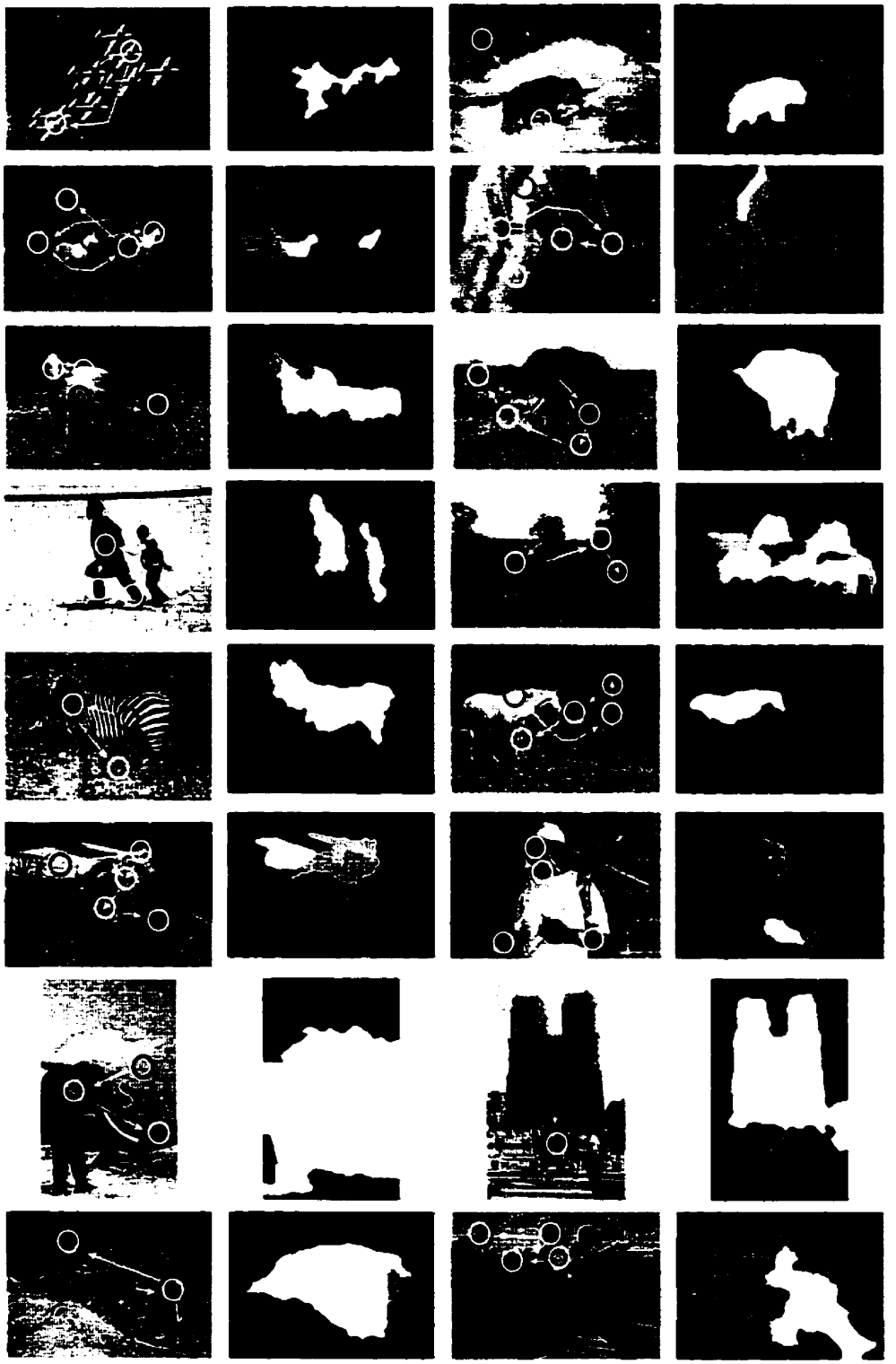
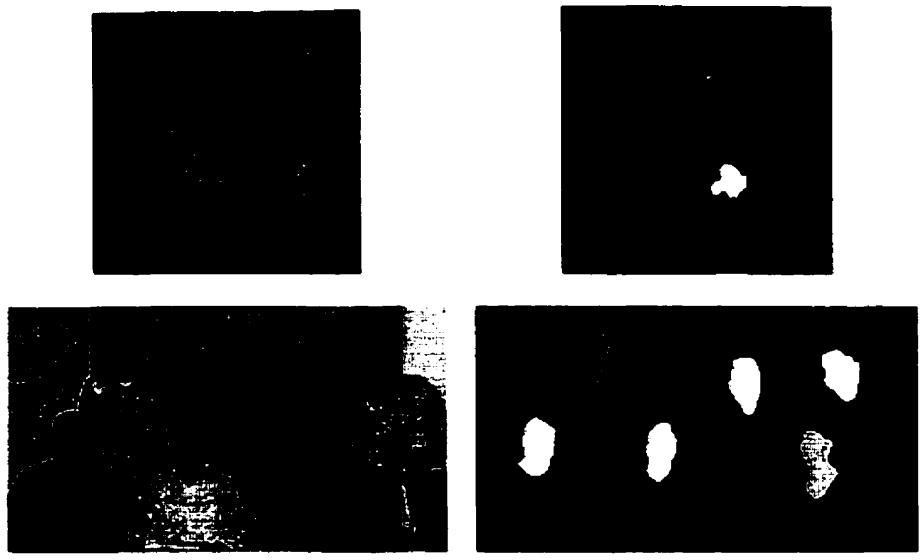


FIGURE 6.8. Importance maps for 16 test images and the most salient regions highlighted in the original image. The most salient region is indicated by a red circle.



(a) (b)

FIGURE 6.9. Face detection. Original imaged (a) and the corresponding importance maps (b). Only color (red) and shape (circular) factors are used in computing the importance map.

shape (elliptical with an aspect ratio of 1:1.5). From the importance map, all the faces are clearly visible in the importance map with very high importance value when compared to other non-face regions. However, this method also detects the arm of the person who is at the far right. Thus, after these candidate regions are identified, more sophisticated algorithms could be applied to further screen out the non-face regions.

6.4.2. Image compression, machine vision, and CBIR

With the availability of an importance map, the major computational resources can be utilised more efficiently and effectively by concentrating on the most salient regions. These resources could be measured by the image compression ratio or the processing time. For CBIR, one of the major goals is to develop a similarity measure that closely resembles the observed visual differences. It generally accepted that global features are not adequate for judging visual differences. Using an importance map, the similarity measure can be based on the salient regions only and hence will not be affected by the background.

CHAPTER 7

Conclusions

In recent years, considerable emphasis has been placed on the development of computer vision systems emulating the performance of a human. Despite the vast difficulties encountered in modelling the human visual system (HVS), the benefits in being able to achieve this have led to continued widespread research in this area. One active research topic is the simulation of the human visual attention system. To function in a real-world environment, an autonomous agent must have an attentional process to locate *objects* in order to build a high-level interpretation of its environment. With this knowledge, the agent can navigate around and perform more complex tasks. Apart from active vision, such an attentional system could be beneficial to other computer vision applications, such as content-based image retrieval (CBIR). This thesis has discussed the implementation issues related to the development of such a system for locating salient *objects* in a scene image.

First, the attention model proposed by Osberger and Maeder [**86**] is analysed. Satisfactory results on real images can be obtained with their original method. However, under certain situations, their method fails to identify some importance regions that are salient to a human. To correct these problems, a number of modifications and several new saliency factors are proposed. From experimentation, we have found that only some of these factors are actually useful for estimating a region's saliency in general. These factors are: contrast, foreground/background, colour, size, and location. Other factors, such as shape and saturation, are applicable only in a number of specific conditions. These factors do not seem to have an equal influence on visual attention. For photographs, where important objects are usually located in the centre of the image, the foreground/background factor is much more important than the

others. The second most important factor is contrast. The rest of the factors have less but similar abilities to attract human attention.

Next, issues related to the implementation of image segmentation and feature selection is discussed. Since the performance of the object-based attention model just described depends largely on the quality of the "object" information, an effective image segmentation technique is required. To mimic the perceptual grouping mechanism in HVS, a number of biologically motivated features for representing the visual property of a region are selected. These features are colour ($L^*a^*b^*$), texture (Gabor), and position. A simple method for estimating the scale of the texture feature is also described.

A number of image segmentation techniques are reviewed with emphasis on their relative strengths and weaknesses. In particular, non-parametric density estimation techniques are best suited to the algorithm used in the attention process since no context-related information is assumed and the regions' information is represented in both spatial and feature domains. In order to have the system fully automatic without any human supervision, a number of clustering validity measure are considered for estimating the *best* number of clusters. These measures are: modified Hubert index [**52**], Pauwels and Frederix's non-parametric measures [**91**], and a threshold-based measure. Surprisingly, the simple threshold-based measure clearly out-performs the other more complex measures for all test images. We believe this contradiction is caused by the incorporation of human preference in the threshold-based measure. Although it is desirable to have an algorithm that is formally defined and does not require any training, it is much more important to have an algorithm that performs correctly as intended. Our experiments indicated that both the modified Hubert index and the Pauwels and Frederix's non-parametric measure did not provide consistent segmentations over a wide range of images.

7.1. Direction of Future Work

The next logical step in the research is the incorporation of high-level, context-dependent grouping and attentional cues. In reality, we seldom find an object that is uniform in colour and texture. In general, most objects, including natural and artificial ones, are composed of several heterogeneous parts. For example, a car has four tires and a chassis. Utilising this higher-level knowledge can help reduce the

over-segmentation inherent in the low-level definition of an *object* as a coherent and homogeneous region. An example of this approach is the body plan of Forsyth and Fleck [**34**].

Another area deserving further attention is the extension of the system to CBIR. In current approaches to CBIR, the similarity measure used treats the whole image as a single region or each sub-regions with equal importance. With a saliency value associated with each region, the comparison between two images can be focused on the salient parts only regardless of the background. This approach is desirable since most image classification methods consider only the few major objects in the scene, such as images containing zebras, cars, or eagles.

APPENDIX A

The Graphical User Interface (GUI)

To facilitate the experimentation with different approaches and methodologies, a graphical user interface (GUI) was created (see Figure A.1). Before any operation can be performed, the user must specify an input image either from the "File Open" dialog or the "Thumbnails" dialog (see Figure A.2). Both dialogs can be accessed from the "File" menu or the toolbar located at the top-left corner of the window. After an image is selected, it will be displayed on the left side of "Main" section of the main window. Then, the image can be analysed by selecting "colour segmentation" from the "Action" menu. This operation takes about 20 seconds for a 180x120 image. After this operation has completed, the best segmentation selected by the cluster validity measure and the corresponding saliency map will be displayed in the first row of the "Results" section. Apart from this information, the segmentations for two to eleven regions from the hierarchical clustering will also be displayed on the last two rows of the same section. Each region in the segmented images is colour coded according to its saliency ranking. The colour scheme used is shown on the right side of the "Main" section.

All major parameters of the feature extraction and image segmentation processes can be modified from the "Test Parameters" dialog (see Figure A.1) by selecting the "Test Parameters" from the "Setting" menu. To change the parameters of the importance map calculation, one can select the "Saliency Parameters" from the same menu to open the "Saliency Parameters" dialog (see Figure A.2).

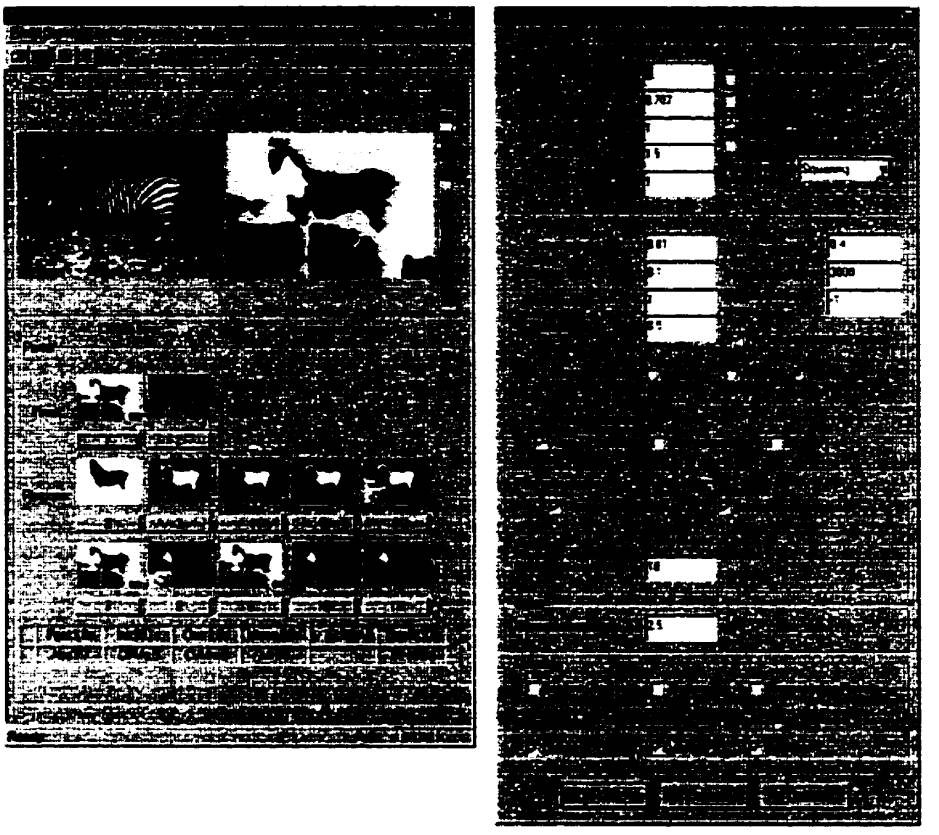


FIGURE A.1. The main window and the test parameter dialog.

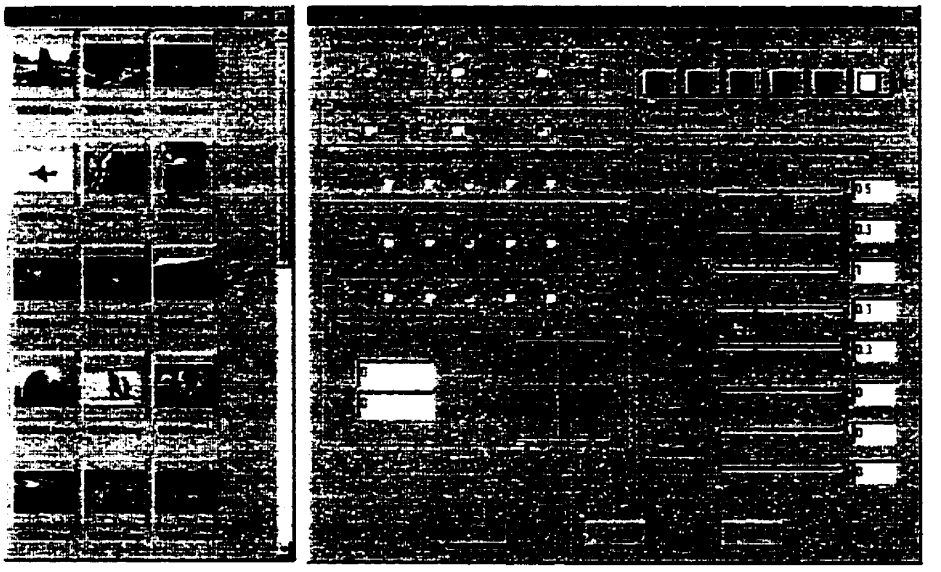


FIGURE A.2. The thumbnail dialog and the saliency parameter dialog.

APPENDIX B

The Image Database

The image database was randomly selected from the Corel image collection[1]. It contains 180 colour images which were used for testing different image segmentation methods and calculating the importance map. Each image has a resolution of 180x120. In order to show the strengths and weaknesses of different approaches, these images were selected from a wide variety of categories including animal, building, insect, people, aeroplane, and scenic pictures. For most of these images, either one or a few salient objects can be easily identified.

FIGURE B.1. The first part of the image database.

[1]http://www.corel.com./products/clipartandphotos/photo/photolib.htm

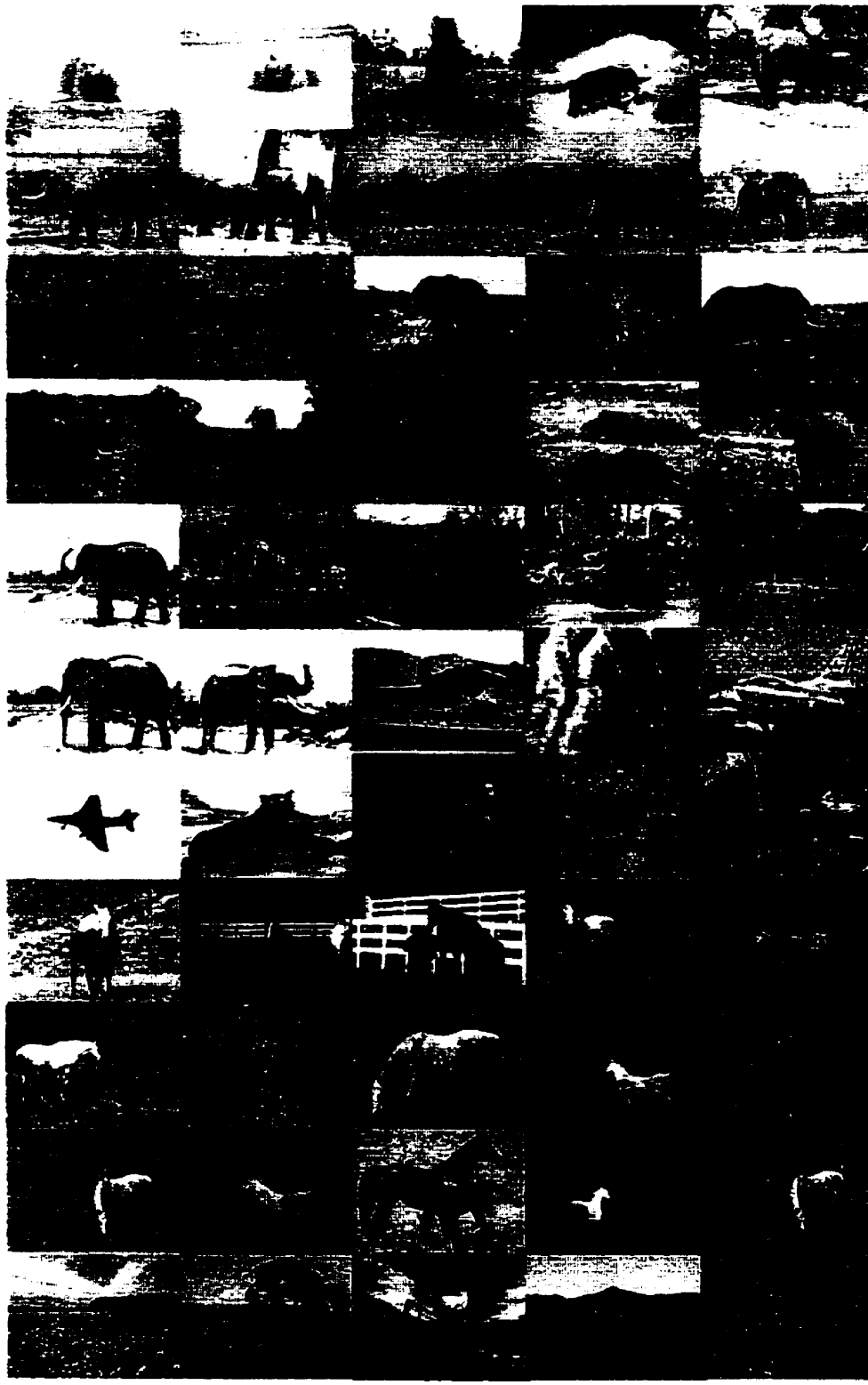


FIGURE B.2. The second part of the image database.

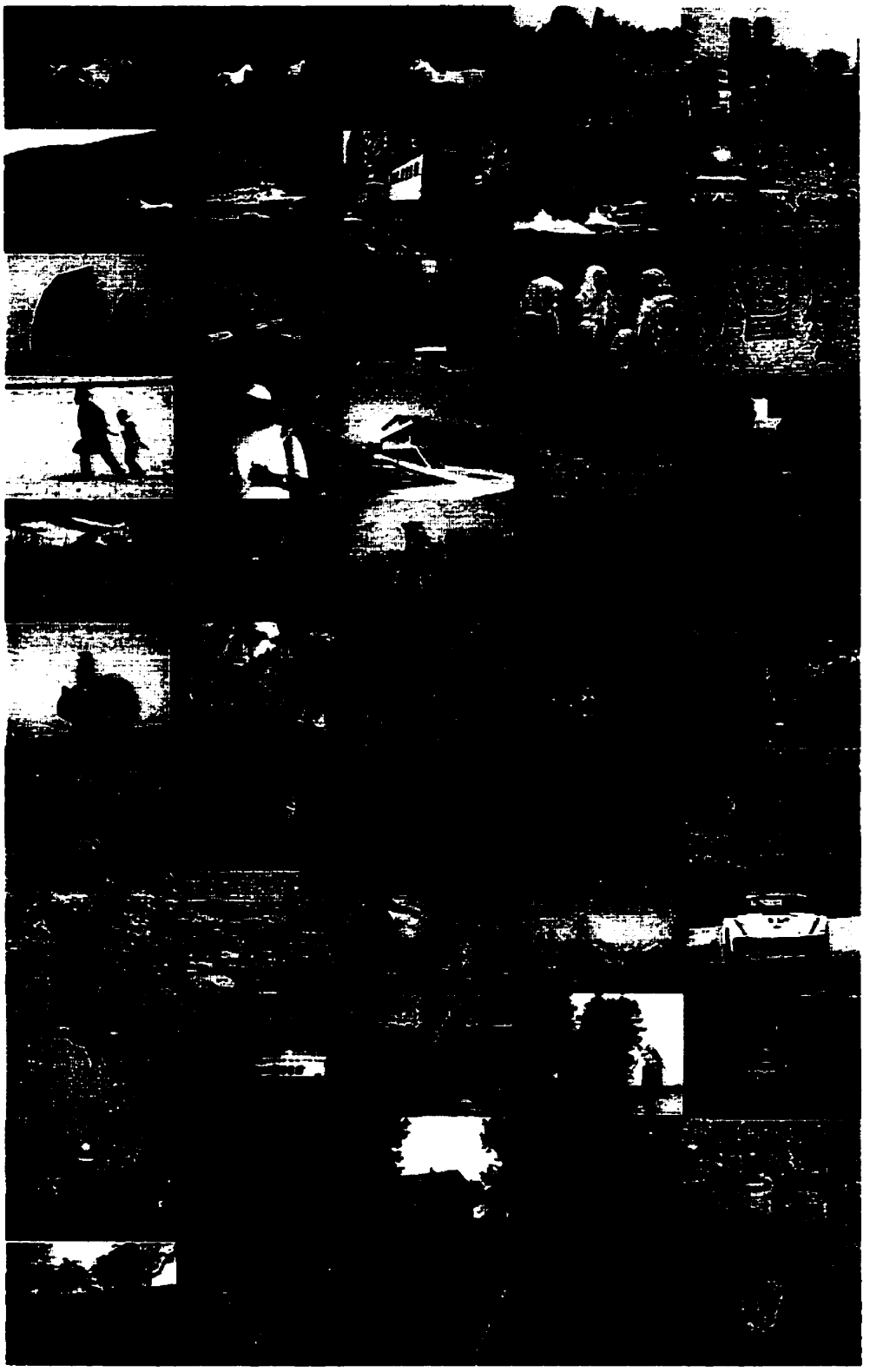


FIGURE B.3. The third part of the image database.

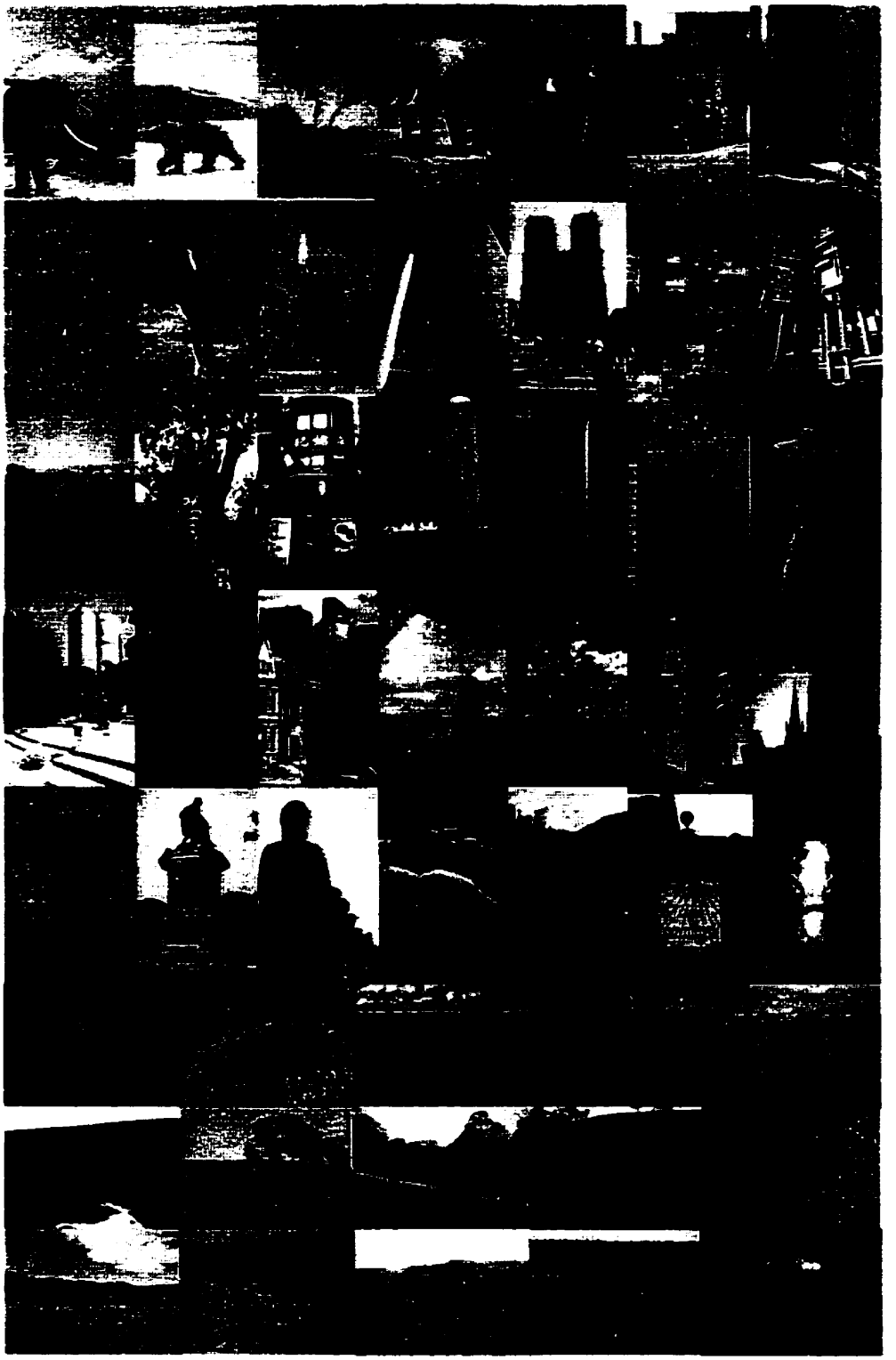


FIGURE B.4. The last part of the image database.

APPENDIX C

The Test Set and Results

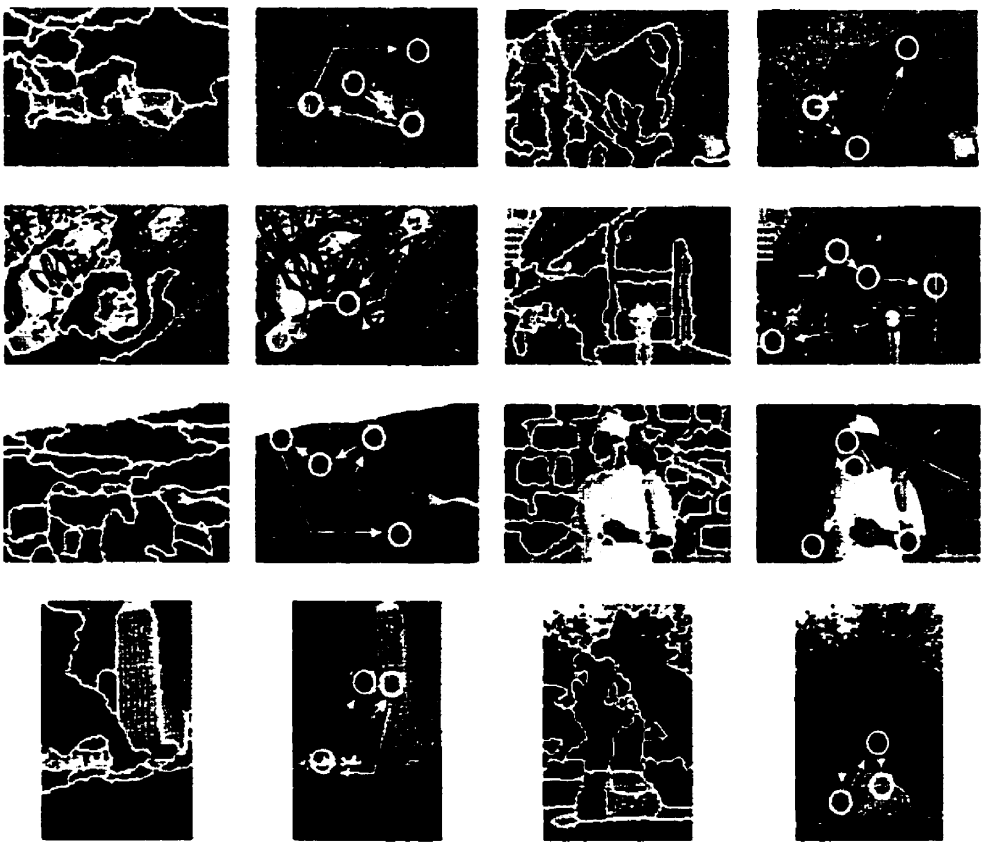


FIGURE C.1. The first part of the test set along with the final segmentation selected by the threshold-based method and the focus of attention (FOA) path. The FOA path is ordered according to decreasing saliency.

FIGURE C.2. The second part of the test set along with the final segmentation selected by the threshold-based method and the focus of attention (FOA) path

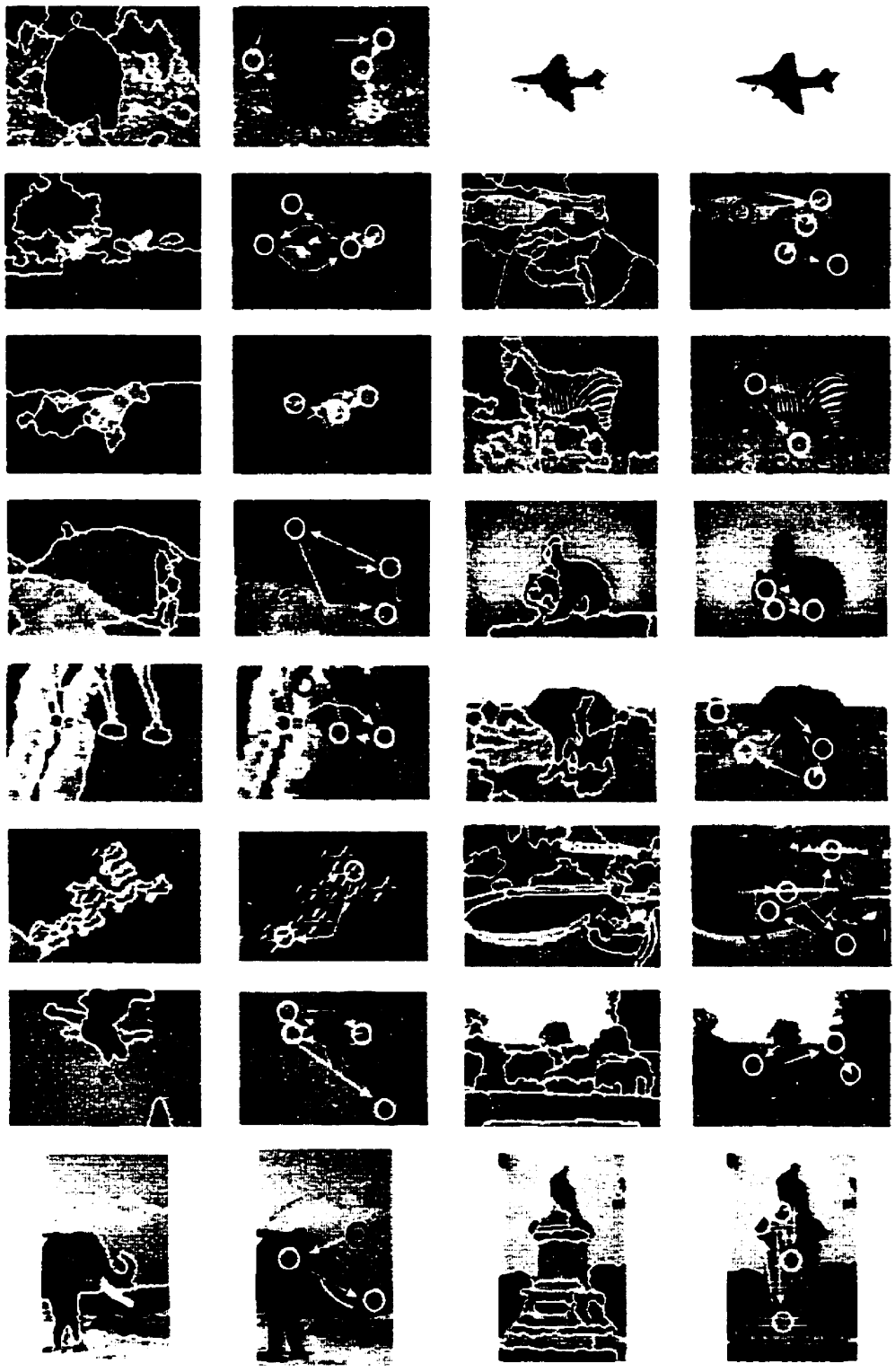


FIGURE C.3. The last part of the test set along with the final segmentation selected by the threshold-based method and the focus of attention (FOA) path

REFERENCES

[1] M. Amadasun and R. King. Texture features corresponding to textural properties. *IEEE Trans. on System Man and Cybernetics*, 19:1264–1274, 1989.

[2] J. Ashley, R. Barber, M.D. Flickner, J.L. Hafner, D. Lee, W. Niblack, and D. Petkovic. Automatic and semiautomatic methods for image annotation and retrieval in query by image content (qbic). *SPIE*, 2420:24–35, 1995.

[3] R. Barsi. Viewer-centered representations in object recognition: A computational approach. In *Handbook of Pattern Recognition and Computer Vision*, pages 925–944. World Scientific, 2 edition, 1999.

[4] G.C. Baylis and J. Driver. Visual attention and objects: Evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance*, 19:451–470, 1993.

[5] S. Belongie and J. Malik. Finding boundaries in natural images: A new method using point descriptors and area completion. In *5th European Conference on Computer Vision*, Freibury, Germany, June 1998.

[6] A. Bhalerao and R. Wilson. Multiresolution image segmentation combining region and boundary information. *CVGIP-Image Understanding*, 59(3):259–366, 1994.

[7] M. Borsott, P. Campadelli, and R. Schettini. Quantitative evaluation of color image segmentation results. *Pattern recognition letters*, 19:741–747, 1998.

[8] J. Braun. Visual search among items of different salience: removal of visual attention mimics a lesion in extrastriate area v4. *J. Neurosci*, 14(2):554–567, 1994.

[9] C.R. Brice and C.L. Fennema. Scene analysis using regions. *AI*, 3:205–226, 1970.

[10] P.J. Burt. Fast filter transform for image processing. *Computer Graphics and Image Processing*, 16:20–51, 1981.

[11] G.T. Buswell. *How people look at pictures*. University of Chicago Press, Chicago.

[12] F.W. Campbell and J.G. Robson. Application of fourier analysis to the visibility of gratings. *J. Physiol.*, 197:551–566, 1968.

[13] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 675–682., Piscataway, NJ, USA, 1998.

[14] Central Bureau of the Commission Internationale de L'Éclairage, Vienna, Austria. *Publication CIE No.15.2*, 2 edition, 1986.

[15] A. Chakraborty. Game theoretic integration for image segmentation. *IEEE Trans. on PAMI*, 21(1), January 1999.

[16] K.I. Chang, K.W. Bowyer, and Sivagurunath Munish. Evaluation of texture segmentation algorithms. *IEEE Computer Vision and Pattern Recognition*, 1:294–299, 1999.

[17] H. Christensen, K. Bowyer, and H. Bunke. *Active Robot Vision*. World Scientific Press, Singapore, 1993.

[18] C.M. Cicerone and J.L. Nerger. The ratio of l cones to m cones in the human parafoveal retina. *Vision Research*, 32(5):879–888, 1992.

[19] C. Colby. The neuroanatomy and neurophysiology of attention. *J. Child Neurol.*, 6:90–118, 1991.

[20] D. Comaniciu and P. Meer. Robust analysis of feature space: Color image segmentation. In *Proceedings of Computer Vision and Pattern Recognition*, 1997.

[21] V. Concepcion and H. Wechsler. Detection and localization of objects in time-varying imagery attention, representation and memory pyramids. *Pattern Recognition*, 29(9):1543 -1557, 1996.

[22] J.G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20:847 - 856, 1980.

[23] J.G. Daugman. Complete discrete 2d gabor transforms by neural networks for image analysis and compression. *IEEE Trans. ASSP*, 36:169–179, 1988.

[24] P. De Garef and G. Christiarens, D. adn d'Ydewalle. Perceptual effects of scene context on object identification. *Psychological Research*, 52:317 -329, 1990.

[25] R.L. De valois, D.G. Albrrecht, and L.G. Thorell. Spatial-frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22:515–599, 1982.

[26] Y Deng, B.S. Manjunath, and H. Shin. Color image segmentation. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1999.

[27] N. Donnelly, G.W. Humphreys, and M.J. Riddoch Parallel computation of primitive shape descriptions. *Journal of Experimental Psychology: Human Perception and Performance*, 17(2):561–570, 1991.

[28] J. du Buf, M. Kardan, and M. Spann. Texture feature performance for image segmentation. *Pattern Recognition*, 23:291 309, 1990.

[29] J. Duncan. Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, 113:501–517, 1984.

[30] G. Elias, G. Sherwin, and J. Wise. Eye movements while viewing NTSC format television. SMPTE Psychophysics Subcommittee White Paper, March 1984.

[31] C.W. Eriksen and J.D. S.T. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, 40(4):225–240, 1986.

[32] C.W. Eriksen and Y. Yeh. Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 11:583–597, 1985.

[33] M. Farah. Is an object an object an object? Cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science*, 1(5):165–169, 1992.

[34] D. Forsyth and M. Fleck. Body plans. In *Proc. IEEE Comp. Soc. Conf. Comp. Vis. and Patt. Rec.*, pages 678–683, 1997.

[35] D.H. Foster and P.A. Ward. Asymemetries in oriented-line detection indicate two orthogonal filters in early vision. *Proceedings of the Royal Society*, 243:83–86, 1991.

[36] V.D. Gesú, C. Valenti, and L Strinati. Local operators to detect regions of interest. *Pattern Recognition Letter*, 18:1077–1081, 1997.

[37] H. Greenspan, S. Belongie, C. Carson, and J. Malik. Recognition of images in large databases using color and texture. *CVPR'97*, 1997.

[38] H. Greenspan, S. Belongie, P. Perona, R. Goodman, S. Rakshit, and C.H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 222–228, 1994.

[39] W.E.L Grimson. The combinatorics of object recognition in cluttered evnironments using constrained search. In *Proc. of the Int. Conf. on Comp. Vis.*, 1988.

[40] R. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67:786–804, 1979.

[41] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image segmentation. *IEEE Trans. on System Man and Cybernetics*, 3:610–621, 1973.

[42] P. Havaldar, G. Medioni, and Stein F. Perceptual grouping for generic recognition. *Int. Journal of Comp. Vis.*, 20(1/2):59–80, 1996.

[43] D. Hearn and M.P. Baker. *Computer graphics*. Prentice Hall, 2 edition, 1986.

[44] J.M. Henderson and A. Hollingworth. Eye movements during scene viewing: An overview. Technical report, Michigan state University, 1997.

[45] J.M. Henderson and A. Hollingworth. High-level scene perception. *Annu. Rev. Psychol.*, 50:243–271, 1999.

[46] J.M. Henderson, P.A. Jr. Weeks, and A. Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25:210–228, 1999.

[47] L. Hérault and R. Horaud. Figure-ground discrimination: a combinatorial optimization approach. *IEEE trans. on PAMI.*, 15(1):899–914, 1993.

[48] R.W.G. Hunt. *Measuring colour*. Ellis Horwood Limited, 2 edition, 1991.

[49] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. *SPIE Human Vision and Electronic Imaging IV*, January 1999.

[50] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20(11):1254–5259, November 1998.

[51] A. Jain and G. Healey. A multiscale representation including opponent color features for texture recognition. *IEEE Trans. of Image Processing*, 7(1):124–128, 1998.

[52] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice Hall, Inc., 1988.

[53] A.K. Jain and F. Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[54] W. James. *The principles of psychology*, volume I. Henry Holt & Co., New York, 1890.

[55] J.B. Jones and Palmer L.A. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1233–1258, 1987.

[56] J. Jonides. Further toward a model of the mind's eye's movement. *Bulletin of the Psychonomic Society*, 21(4):247–450, 1983.

[57] B. Julesz. A brief outline of the texton theory of human vision. *Trends in Neuroscience*, 7:41–45, Feburary 1984.

[58] B. Julesz. Towards an axiomatic theory of preattentive vision. In *Dynamic Aspects of Neocortical Function*, pages 585–612. Wiley, 1984.

[59] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Reidel Publishing, 1987.

[60] M. Lades. Face recognition technology. In C.H. Chen, L.F. Pau, and P.S.P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, pages 667–683. World Scientific, 2 edition, 1999.

[61] P. Lambert and T. Carron. Symbolic fusion of luminance-hue-chroma features for region segmentation. *Pattern Recognition*, 32:1857–1872, 1999.

[62] N. Lavie and J. Driver. On the spatial extent of attention in object-based visual selection. *Perception and Psychophysics*, 58:1238–1251, 1996.

[63] K.I. Law. *Texture image segmentation*. PhD thesis, University of Southern California, 1980.

[64] T. Leung and J. Malik. Detecting, localizing and grouping repeated scene elements from an image. In *4th European Conference on Computer Vision*, Cambridge, England, April 1996.

[65] F. Liu and R.W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(7):722–733, 1996.

[66] D.G. Lowe. *Perceptual organization and visual recognition*. Kluwer academic publishers, 1985.

[67] D.G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[68] W.Y. Ma and B.S. Manjunath. Edge flow: A frame work of boundary detection and image segmentation. Technical Report 97-02, University of California, Santa Barbara, CA, 1997.

[69] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[70] S.K. Mannan, K.H. Ruddock, and D.S. Woodings. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spat. Vis.*, 9:363–386, 1995.

[71] S.K. Mannan, K.H. Ruddock, and D.S. Woodings. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spat. Vis.*, 10:165–188, 1996.

[72] D. Marr. *Vision: A computational investigation into the Human presentation and processing of visual information*, chapter 1, page 270. W.H. Freeman and Company, 1982.

[73] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company, 1982.

[74] P. McLeod, J. Driver, and J. Crisp. Visual search for conjunctions of movement and form in parallel. *Nature*, 332:154–155, 1988.

[75] R. Milanese. *Detecting salient regions in an image: from biological evidence to computer implementation*. PhD thesis, University of Geneva, Switzerland, December 1993.

[76] R. Milanese, H. Wechsler, S. Gil, J.M. Bost, and T. Pun. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. *IEEE*, pages 781–785, 1994.

[77] M. Miyahara and Y. Yoshida. Mathematical transformation of (r,g,b) color data to Munsell (H,V,C) color data. *SPIE Visual Communications and Image Processing*, 1001:650–657, 1988.

[78] R. Mohan and R. Nevatia. Perceptual organization for scene segmentation and description. *IEEE Trans. on PAMI*, 14(6):616–635, June 1992.

[79] H.J. Muller and A. Found. Visual search for conjunctions of motion and form: Display density and asymmetry reversal. *Journal of Experimental Psychology: Human Perception and Performance*, 22(1):122–132, 1996.

[80] A.M. Nazif and M.D. Levine. Low level segmentation: An expert system. *IEEE Trans. Pattern Anal. and Machine Intell.*, 6(5):555–577, 1984.

[81] U. Neisser. *Cognitive psychology*. Appleton, New York, 1967.

[82] H.C. Northdurft. The role of features in preattentive vision: Comparison of orientation, motion, and color cues. *Vision Research*, 33(14):1937–1958, 1993.

[83] P.P. Ohanian and R.C. Dubes. Performance evaluation for four classes of textural features. *Pattern Recognition*, 25(8):819–833, 1992.

[84] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[85] T. Ojale and M. Pietikainen. Unsupervized texture segmentation using feature distributions. *Pattern Recognition*, 32:477–486, 1999.

[86] W. Osberger and A.J. Maeder. Automatic identification of perceptually important regions in an image. In *ICPR'98*, pages 701–704, Brisbane, Australia, August 1998.

[87] N.R. Pal and S.K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[88] D.K. Panjwani and G. Healey. Markov random field models for unsupervised segmentation of textured color images. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 17(10):939–953, 1995.

[89] T.V. Papathomas, R.S. Kashi, and A. Gorea. A human vision based computational model for chromatic texture segregation. *IEEE Trans. on System Man and Cybernetics*, 27(3):428–439, June 1997.

[90] S.H. Park, I.D. Yun, and S.U. Lee. Color image segmentation based on 3-d clustering: Morphological approach. *Pattern Recognition*, 31(8):1061–1076, 1998.

[91] E.J. Pauwels and G. Frederix. Finding salient regions in images. *Journal of Computer Vision and Image Understanding*, 75(1/2):73–85, 1999.

[92] T. Pavlidis and Y.T. Liow. Integrating region growing and edge detection. *IEEE Trans on Pattern Anal. and Machine Intell.*, 12(3):225–233, March 1990.

[93] D.L. Pham and J.L. Prince. An adaptive fuzzy c-means algorithm for image segmentation in the presence of intensity inhomogeneities. *Pattern Recognition Letter*, 20(1):57–68, 1999.

[94] M. Pietikänen, A. Rosenfeld, and L.S. Davis. Experiments with texture classification using averages of local pattern matches. *IEEE Trans. on System Man and Cybernetics*, 13(3):421–426, 1983.

[95] S. Posch and D. Schlüter. Perceptual grouping using markov random fields and cue integration of contour and region information. Technical Report SFB360-TR-98/10, University of bielefeld, 1998.

[96] C.A. Poynton. *A technical introduction to digital video*. Wiley, New York, 1996.

[97] J. Puzicha and J.M. Buhmann. Multiscale annealing for real-time unsupervised texture segmentation. In *Proceedings of the IEEE Int. Conf. on Comp. Vis. 98*, pages 267–273, Piscataway, NJ, USA, 1998.

[98] T. Randen and J. Husoy. Filtering for texture classification: A comparative study. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(4):291–310, 1999.

[99] A.R. Rao and G.L. Lohse. Identifying high level features of texture perception. *CVGIP: Graphical Models and Image Processing*, 55(3):218–233, 1993.

[100] T. Reed and J. du Buf. A review of recent texture segmentation and feature extraction techniques. *CVGIP: Image Understanding*, 57(3):392–372, May 1993.

[101] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: the generalized symmetry transform. *Int. Journal of Comp. Vis.*, 14:119–130, 1995.

[102] R.A. Rensink and J.T. Enns. Pre-emption effects in visual search: evidence for low-level grouping. *Psychological Review*, 102(1):101–130, 1995.

[103] D. Rosenberg. Monocular depth perception for a computer vision system. Master's thesis, McGill University, September 1981.

[104] I.A. Rybak, V.I. Gusakova, A.V. Golovan, L.N. Podladchikova, and N.A. Shevtsova. A model of attention-guided visual perception and recognition. *Vision Res.*, 38:2387–400, August 1998.

[105] S. Santini and R. Jain. Similarity measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[106] S. Sarkar and K.L. Boyer. Perceptual organization in computer vision: A review and a proposal for a classificatory structure. *IEEE Tran. on SMC*, 23(2):382–399, 1993.

[107] S. Sarkar and K.L. Boyer. A computational structure for preattentive perceptual organization: Graphical enumeration and voting methods. *IEEE Trans. on SMC*, 24(2):246–267, February 1994.

[108] D. Schlüter and Posch S. Combining contour and region information for perceptual grouping. In *Proceedings 20. DAGM-Symposium*, pages 393–401. Mustererkennung, 1998.

[109] G. Sela and M.D. Levine. Real-time attention for robotic vision. *Real-time Imaging*, 3:173–194, 1997.

[110] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proc. of the IEEE Conf. on Comp. Vision and Pattern Recognition*, San Juan, Puerto Rico, June 1997.

[111] J. Shi and J. Malik. Self inducing relational distance and its application to image segmentation. In *5th European Conference on Computer Vision*, June 1998.

[112] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, New York, 1986.

[113] G.M. Smith. *Image texture analysis using zero crossings information*. PhD thesis, The University of Queensland, 1998.

[114] J.R. Smith. *Integrated spatial and feature image system: Retrival, analysis, and compression*. PhD thesis, Columbia University, February 1997.

[115] J.R. Smith and C.S. Li. Image classification and querying using composite region templates. *Computer Vision and Image Understanding*, 75(1):165–174, 1999.

[116] M. Spann. Figure/ground separation using stochastic pyramid relinking. *Pattern Recognition*, 24(10):993–1002, 1991.

[117] P.F.M. Stalmeier and M.M. de Weert. Large color differences and selective attention. *J. Opt. Soc. Am. A*, 8(1):237–247, 1991.

[118] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta. A standard default color space for the internet-srgb. http://www.w3.org/Graphics/Color/sRGB.html, 1999.

[119] J. Strand and T. Taxt. Local frequency features for texture classification. *Pattern Recognition*, 27(10):1397–1406, 1994.

[120] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Trans. on System Man and Cybernetics*, 8:460–473, 1978.

[121] J. Theeuwes. Perceptual selectivity for color and form. *Perception and psychophysics*, 51(6):599–606, 1992.

[122] D. Travis. *Effective color displays*. Academic Press, London, 1991.

[123] A.M. Treisman. The role of attention in object perception. In O.J. Braddick and A.C. Sleigh, editors, *Proc. of Royal Society Int. Symp. on Physical and Biological Processing of Images*, pages 316–325, New York, 1983. Springer.

[124] A.M. Treisman. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31:156–177, 1985.

[125] A.M. Treisman. Features and objects: The fourteenth barlett memorial lecture. *Quarterly Journal of Experimental Psychology*, 40a:201–237, 1988.

[126] A.M. Treisman, P. Cavanagh, B. Fisher, V.S. Ramachandran, and R. von der Heydt. Form perception and attention. In *Visual perception: The neurophysiological foundations*. Academic Press, New York, 1990.

[127] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[128] A.M. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95:15–48, 1988.

[129] A.M. Treisman and R. Paterson. Emergent feature, attention and object perception. *Journal of Experimental Psychology: Human Perception and Performance*, 10:12–31, 1984.

[130] Y. Tsal and N. Lavie. Location domination in attending to color and shape. *Journal of Experimental Psychology: Human Perception and Performance*, 19:131–139, 1993.

[131] J.K. Tsotsos, S.M Culhane, Y.K. Wai, N. Lai Yuzhong, Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995.

[132] M. Tuceryan and A.K. Jain. Texture segmentation using voronoi polygons. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12:211–216, 1990.

[133] M. Tuceryan and A.K. Jain. Texture analysis. In C.H. Chen, L.F. Pau, and P.S.P. Wang, editors. *Handbook of pattern recognition*, 2, pages 207–248. World Scientific, 1999.

[134] S. Ullman. *High-level vision: Object recognition and visual cognition*, chapter 8, pages 234–235. The MIT Press, 1998.

[135] L. van Gool, P. Dewael, and A. Osterlinck. Texture analysis anno 1983. *Computer Vision Graphics and Image Processing*, 29:336–357, 1985.

[136] K.F. Van Orden. Redundant use of luminance and flashing with shape and color as highlighting codes in symbolic displays. *Human Factors*, 35(2):141–160, 1993.

[137] J.J. Vos, O. Estévez, and P.L. Walraven. Improved color fundamentals offer a new view on photometric additivity. *Vision Research*, 30:936–943, 1990.

[138] H. Wechsler. Texture analysis - a survey. *Signal Processing*, 2:271–282, 1980.

[139] E. Weichselgartner and G. Sperling. Dynamics of automatic and controlled visual attention. *Science*, pages 778–780, 1987.

[140] M. Wertheimer. Experimentelle studien üiber des sehen von bewegung. *Zeits. f. Psychol.*, 61:161–265, 1912.

[141] M. Wertheimer. *Principles of perceptual organization*. Princeton, N.J., 1958.

[142] J. Weszka, C. Dyer, and A. Rosenfeld. A comparative study of texture measures for terrain classification. *IEEE Trans. on Systems Man and Cybernetics*, 6:267–285, 1976.

[143] C.D. Wickens. *Engineering Psychology and Human Performance*. HarperCollins Publishers Inc., New York, 2 edition, 1992.

[144] P.S. Williams and M.D. Alder. Segmentation of natural images using hierarchical and syntactic methods. In *Second International Workshop on Statistical Techniques in Pattern Recognition*, August 1998.

[145] J.M. Wolfe. Extending guided search: Why guided search needs a preattentive "item map". In *Converging operations in the study of visual selective attention*, pages 247–270. American Psychological Association, Washingon, DC, 1996.

[146] J.M. Wolfe. *Attention*, chapter 1. Psychology Press, 1998.

[147] G. Wyszecki and W.S. Stiles. *Color science: Concepts and methods, quantitative data and formulae*. A Wiley-Interscience Publication, 2 edition, 1982.

[148] C.T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. on Comp.*, 20(1):68–86, 1971.

[149] Y.J. Zhang. A survey of evaluation methods for image segmentation. *Pattern Recognition*, 29(8), 1335-1349 1996.

[150] S.C. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(9):884–900, 1996.

[151] S.W. Zucker. Toward a low-level description of dot clusters: labeling edge, interior, and noise points. *IEEE Proceedings*, pages 213–233, 1979.