COLOUR IMAGE DISPLAY:

A COMPUTATIONAL FRAMEWORK BASED ON A UNIFORM COLOUR SPACE

Philip Keith Robertson

April 1985

A thesis submitted for the degree of Doctor of Philosophy
at the Australian National University

Declaration

I hereby declare that except where otherwise explicitly stated, the work presented in this thesis is my own original work.

Philip K Robertson

Philip K. Robertson

Acknowledgements

I am grateful and indebted to my principal supervisor, Dr. J. F. O'Callaghan, for his initial suggestions for this work, and his guidance and encouragement throughout its course.

I would also like to thank my co-supervisor Prof. R. P. Brent, and advisor Dr. R. A. Jarvis, for their comments and advice at various stages.

Particular thanks are also due to Dr. D. Fraser for giving assistance and advice in many discussions during the course of this work, and to Dr. T. R. J. Bossomaier for his comments on specific aspects of it.

This work was performed using the equipment and facilities of the CSIRO Division of Computing Research (now CSIRONET). I would like to thank Dr. P. J. Claringbold, Chief of the Division, for making these facilities available, and Prof. R. P. Brent, of the ANU Computer Science Department, for supporting the joint arrangement.

I would also like to express my appreciation of discussions with, and help at various times from, those associated with the Image Systems Section of the Division: Ian Briggs, Pam Cohen, Duncan Stevenson, Garth Tier, and Murray Wilson. I also benefitted from discussions with Richard Juday (NASA) early in this work.

Financial assistance in the form of an Australian Commonwealth Postgraduate Research Award is acknowledged.

Abstract

This thesis develops a comprehensive approach to the use of colour in the display of digital image data. The approach incorporates as a primary consideration the principle that a display should exploit the normal scene analysing capabilities of the human visual system to achieve intuitive appreciation of the spatial variations of image data variables. It is realised within a computationally tractable framework; central to this framework is a perceptually uniform colour space which can be addressed in terms of perceptual attributes hue, saturation and lightness.

The development of this approach arises from the shortcomings of conventional colour display techniques for representing various types of image data, including multivariate statistical data, remotely sensed multi-spectral data, data of an arbitrary nature which may be in a form not normally observed visually, and integrations of such data types. This work rationalises conventional and ad-hoc display methods, using considerations of the operation of the human visual system to develop a display approach which encompasses these types of data. It draws from the areas of colour science, visual perception, and computer vision.

The display approach is based first, on achieving an appropriate structural representation for the data; second, on representing it in terms of appropriate perceptually significant spectral attributes; and third, on performing appropriate enhancements based on specific mechanistic knowledge of the visual pathway.

The first stage of the approach is realised by presenting the display in the form of a realistic scene. Such a display is shown to allow appreciation of two superimposed data variables with dissimilar spatial structures which under conventional display methods are difficult to interpret. One data variable is represented as the topography of a surface and the other as the colour of the surface, or of an overlaid transparency. Modelling the physical processes of reflection from, and pigmentation of, surfaces allows the realistic depiction of these scene properties; these processes are modelled in terms of their perceived spectral effects. The second stage of the proposed approach allows realisation of these models, using representations of perceptual spectral attributes in uniform colour space.

Perceptual attribute representations are also used for displaying general informative data variables, derived from remotely sensed multi-spectral data, and from statistical data. Such representations are shown to result in data displays more intuitively interpretable than those produced using conventional or ad-hoc spectral assignment methods. The use of a colour space whose metric reflects perceived colour differences also allows the proportional representation of numerical data steps by perceived gradations in colour.

The third stage of the approach is realised by deriving a representation of achromatic and chromatic opponent visual channels within uniform colour space. Considerations of the detrimental effects of the different spatial resolutions of the achromatic and chromatic visual mechanisms at high spatial frequencies lead to a suggestion for low-pass filtering chromatic information. Such filtering can be performed on the chromatic channel representation in uniform colour space; these channels also form a perceptual domain for performing other image enhancements.

It is shown that overall, use of the developed display approach can result in a substantial improvement over conventional display methods, both in high level scene comprehension of complex composite images, and in lower level spectral interpretation of imagery. The use of perceptual specification for data representations, and the increased control over the colour product resulting, can add significantly to the level of appreciation of data attained.

Table of Contents

List of illustrations

List of tables

Chapter 1
Introduction

1.1 The problem of displaying image data in colour

Our aim in this thesis is to develop a comprehensive approach to the use of colour in the display of digital image data. Image displays are used to provide a visual representation of the spatial nature of graphical, statistical, or physical data in a wide range of fields. Colour is used to increase the capacity of such a presentation to convey information. Often several image data variables are simultaneously displayed in a spatially superimposed manner; the aim of such a presentation is to portray the spatial relationships between the variables, while still allowing appreciation of the characteristics of each individual variable. Achieving this aim involves consideration of the nature and perception of colour, and draws on work in the areas of colour science, human colour perception, and computer vision.

We consider an image data variable, or simply an image, to be a spatially two-dimensional, regularly gridded, numerical representation of some chosen measured, derived or otherwise specified data variable. In general, the two spatial dimensions of an image represent actual spatial dimensions, but in some cases they can represent temporal or other non-spatial variables. As an artificially created representation, an image may exhibit characteristics or properties not inherent in the original data variable being represented. We consider the dimensionality of an image data set to be the number of distinct image data variables to be presented in a spatially superimposed manner as a composite image.

Image data variables are used to portray information in a wide range of fields. They can represent graphical or statistical variables defined over a two-dimensional area; for example, a multivariate choropleth map is formed from a set of variables which have a common spatial structure consisting of homogeneous areas and well defined boundaries. The colour figures of Chapter 4 show examples of such displays. Image data variables can also represent geological, geophysical, ecological, meteorological or other physically-based properties. Such imagery is often produced by superimposing spectral channels from a set of remotely sensed data representing reflected radiation (from such sources as the Landsat or NOAA satellite, or airborne, multi-spectral scanners). Examples of this type of data set, in which the spectral channels are often highly correlated, are given in the colour figures of Chapter 5. Such spectral channels often do not directly represent aspects of the data most informative to an interpreter; some pre-processing to generate informative data variables can be required. Composite images can also be produced by combining image data variables derived from different measurement processes, and consequently with arbitrary or dissimilar spatial structures. When several image data variables are derived from the same geographical area, spatial structure can be correlated to a varying degree over portions of the spatial field; it is very often these variations in correlation which are of interest to an analyst and which should therefore be depicted. The integration of a number of such data variables; for example, of topographic data with data depicting variations in magnetic field strength (as shown in the colour figures of Chapter 6); results in display problems due to the dissimilarity of the spatial structures of the data variables being superimposed.

Conventional methods of using colour to display these types of image data have been based around the technology of colour display devices; typically each of up to three data variables is presented on a primary colour, or combination of primary colours, of a display system. The result is not conducive to intuitive interpretation, as such arbitrarily chosen colours are not necessarily, or even likely to be, easily separable by the human visual process. Further, if the superimposed data variables have dissimilar spatial structures, as can occur when displaying variables which are not generally observed simultaneously, the composite image can be confusing and difficult to interpret at all. Attempts to improve the separability of data variables by presenting them as variables significant in a perceptual sense have been made (see Chapters 2 and 3). These methods can result in improved appreciation of certain types of data sets, but for data variables with different spatial structures, little gain is made by their use. Consequently there is a clear need to consider not only the manner of direct spectral assignment to such data sets, but also the overall manner in which the composite image is perceived by the visual system. This must incorporate both spectral and spatial considerations.

1.2 Approach taken to the problem - contribution of this work

The limitations of conventional image display methods can be explained if image interpretation is considered in the context of the information processing capabilities of the human visual system. Haber and Wilkinson (1982) suggest that "the effectiveness of information communication in a display depends on how closely the structure inherent in the information is mapped onto the modes by which the visual system processes the image". They further suggest that the visual system "attempts to interpret all stimulation reaching the eyes as if it were reflected from a real scene in three dimensions". When interpreting a real scene, the visual system relies on receiving appropriately reinforcing signals on its perceptual channels to provide information about the physical properties of the scene. This leads to comprehension of the overall scene structure. Conventional methods of image display do not necessarily generate these perceptual signals, and consequently are unlikely to depict realistic scenes. As a result, it is unlikely that a composite image formed by presenting data variables on display device primaries will achieve effective communication of information.

In this thesis we develop a comprehensive display approach which incorporates as a primary consideration the principle that any display should be recognisable as a realistic scene. This approach is based on a computationally tractable framework, and is designed to exploit the normal scene analysing capabilities of the human visual system to achieve intuitive appreciation of the spatial variations of superimposed image data variables of the type described in the previous section. Such a framework has not, to our knowledge, been previously developed to this level of detail.

In brief, the approach ensures that structural comprehension of a presented data display can occur; that is, a presented display must be recognisable as a two-dimensional representation of a realistic physical scene. We explain why in some data displays this representation is inherent, and propose techniques to achieve it when it is not. Such techniques are based on representing informative data variables as natural scene variables, and rely on characterising these scene variables in terms of their perceived attributes, the spectral descriptors of which are commonly termed hue, saturation and lightness.

We consider the process of interpretation of the colours, or spectral variations, in a scene for which the overall comprehension requirements are satisfied. We stress the importance of being able to interpret informative data variables in terms of perceptually significant spectral attributes, which have direct relation to the physical processes that depict natural scenes, rather than in terms of some arbitrary spectral attributes such as the colour primaries of a display device. Representation of data by perceptually significant attributes is not in itself new (see Chapters 2 and 3), but its incorporation within a general display framework, and its use to depict natural scene attributes, is to our belief novel. The importance of being able to represent perceptually significant spectral attributes as closely as possible leads us to base the display framework around a perceptually uniform colour space in which these attributes have a natural representation and meaningful metric.

We also look at the role of the application of knowledge about specific properties of the visual pathway in enhancing data presentations. We treat such processes, which are generally well defined only at low levels in the visual pathway due to the complexity of higher level interconnections and processes, as peripheral to the primary requirements for structural comprehension of a presented data display, but as nevertheless significant in the role of enhancing or clarifying such comprehensions if the structural requirements are met. We propose a means of realising a transformation from a perceptual level to a lower visual level (specifically, to a representation of colour-opponent channels) which can be implemented in practice. We also suggest reasons, substantiated by subjective experiments, for spatial frequency filtering at this opponent level to avoid detrimental effects associated with the creation of artificial images in the display of data. These effects arise as a result of the different contrast sensitivities of the achromatic and chromatic visual mechanisms at high spatial frequencies. We believe this aspect of the work also to be novel.

In summary, then, we propose a display approach which is based first, on achieving an appropriate structural representation for the data; second, on representing it in terms of appropriate perceptually significant spectral attributes; and third, on performing appropriate enhancements based on specific mechanistic knowledge of the visual pathway. We believe that such an approach can not only allow intuitive appreciation of superimposed image data sets which under conventional display methods are difficult to interpret, but also can substantially improve the presentation of data sets in general. This approach, and its practical realisation, are based on what is generally known, deduced or surmised about the operation of the human visual system. It draws from the areas of visual perception, colour science, and computational approaches to scene understanding (applied in the fields of artificial intelligence and computer vision). The framework is designed on the basic principles of visual system operation, rather than on specific interconnection details; it is hence not critically dependent on the current level of mechanistic knowledge of visual processes. We believe that the development of such a comprehensive approach is novel, even though it draws on many previously developed isolated applications of visual-system-dependent display processes.

1.3 Organisation of the thesis

In Chapter 2 we look at the overall operation of the human visual system, and at the visual processes involved in the comprehension and interpretation of images. We isolate factors relevant to structural scene comprehension, and the manner in which these are dependent on appropriate spatial and spectral representations in terms of perceptual attributes and their spatial distributions. We then consider the application of visual system models and considerations in image display, and develop a set of requirements for a computational framework within which the display approach can be implemented.

In Chapter 3 we develop a suitable computational framework. This framework encompasses the spectral and spatial operations involved in the structural synthesis of a realistic scene, spectral assignment at a perceptual level, and enhancements or compensations dependent on the low-level properties of the visual system.

We then treat data displays in which the requirements for structural comprehension are potentially satisfied within the data itself.

First, in Chapter 4 we consider the display of data in the form of spectral variations on a flat surface, treating both single variable (pseudo-colour) displays, or choropleth map displays (in which the spatial structure of variables simultaneously displayed is identical, boundaries between homogenous data areas being common to each variable).

Second, in Chapter 5 we consider the display of data in which the scene structure information is inherently embedded in the data itself, and structural comprehension can be achieved if this information is appropriately depicted. This type of data generally comes from the measurement of radiation reflected from a surface; the reflected signal intensity is modulated according to surface orientation and properties. Data sets of this type include remotely sensed multi-spectral scanner or photographic data.

In Chapter 6 we treat data displays in which the requirements for structural comprehension are not inherently satisfied in the data itself. Such data sets arise from the superimposition of images with dissimilar spatial structures, and the display generally attempts to portray information not normally observed in such a form. The superimposition of variations in magnetic or gravitational field strength with topographic variations forms an example of a display of this type. We realise such data as natural variables in a synthesised realistic scene, and consider also the use of such techniques to reinforce appreciation of data of the types treated in Chapters 4 and 5.

In Chapter 7 we investigate the extent to which we can improve the structural or perceptual comprehension of data displays within the developed display framework. In particular, we consider the effects of the spatial frequency characteristics of the visual system on the comprehension of an artificially presented scene (as any data display is), and consider compensation for, or avoidance of, detrimental effects.

Finally, in Chapter 8 we consider the extent to which implementation of the proposed approach within the developed framework achieves the overall display aims initially set. We isolate limitations of the developed techniques and suggest possible directions for further work.

Chapter 2
Image comprehension and interpretation

In the previous chapter we suggested that satisfactory interpretation of imagery depends on achieving an intuitive appreciation of the data variables presented in a display. Because the visual system is able to intuitively interpret visual scenes with which it is familiar, it is most likely to be able to interpret artificial data sets which are similarly presented as images depicting real scenes. On this basis, we investigate the processes involved in extracting information from real scenes, and consider appropriate ways of exploiting these processes in image display problems.

This approach involves investigating the overall operation of the human visual system. To this end, we look first at what is currently known of aspects of the operation of the visual system relevant to the colour display problem. Visual system models, and in particular those which have been used as a basis for computer-based image processing methods, are discussed. An overall approach to displaying image data in colour is developed, and in following chapters a framework to allow realisation of this approach is formalised and applied to various image data types.

2.1 The human visual system

In this section we outline the general structure of the human visual system, concentrating on aspects most pertinent to the time-invariant display of image data in colour. A comprehensive physiological and psychophysical description of the visual system operation is beyond the scope of this work; rather a selective summary is given. Detailed treatment of visual physiology and psychology can be found in Cornsweet (1971), Haber and Hershenon (1973), Evans (1974), Robson (1980) and Gouras and Zrenner (1981). Summaries of the physiological and psychophysical aspects of vision are given by Taenzer (1976), Wasserman (1979) and Tsotos (1984). Barlow (1981) places current physiological knowledge in the context of understanding vision from a standpoint of extracting, at all levels of visual operation, useful information from visual images; Marr (1982) places physiological and psychophysical experimental results in the context of overall visual function, considering the plausibility of visual operations from a purpose-specific point of view. In this work we are concerned not with investigating physiological or psychophysical aspects of the visual system in detail, but rather with understanding the overall significance of observed or measured visual properties to the processing of colour visual information. This should enable us to make appropriate use of the functions of the visual system to assist, rather than have them confound, appreciation of data presented in colour image form.

We start by considering what is known of the physiology of the visual system neural components, their interconnections and transmitting channels. This information comes from neurophysiological experiments on primates; the extrapolation to humans is justified by the believed closeness of the human visual system to that of other primates, particularly to that of monkeys. We then consider the perception of visual stimuli in terms of the perceptual attributes which form natural descriptors of these stimuli, and the concept of a structural representation of an image, arising from the spatial distribution of perceived image attributes. The relevance of these processes to the design of a colour image display approach is discussed.

2.1.1 The visual pathway - mechanistic interconnections and processes

Visual signals are derived from photo-absorption in the retinal receptors. Retinal receptors are of two types. Rods are responsible for scotopic or low-illumination level vision, and have a single broadband spectral response; at normal illumination levels they are saturated, and do not contribute significantly to visual sensation. Cones are operative at photopic or normal illumination levels; there are three distinct types, with distinct spectral absorption characteristics (Cornsweet, 1971; Gouras and Zrenner, 1981). A single response (a graded electrical potential) results from stimulation at any wavelength within the spectral response range of each type of cone; human colour vision is hence three-dimensional (or trichromatic). Receptor density varies across the retina. In the central or foveal region, which subtends a solid angle of between one and two degrees, cone density is highest, with medium- and long-wavelength-responsive cones predominating. In fact there are no short-wavelength-responsive cones in the central area of the fovea.

Signal transmission between the retina and the visual cortex is in the form of modulated ganglion cell firing rates, carrying along the optic nerve (which consists of ganglion cell axons) an achromatic and two chromatic colour-opponent channels. These colour-opponent channels provide relative measures of the red-green stimulus and of the blue-yellow stimulus; the opponent signals are generated from a combination of an excitatory input and an inhibitory input, one from each of two types of cone (Gouras and Zrenner, 1981). Ganglion cells carrying the colour-opponent signals have a centre-surround response to retinal stimuli over a circular area called a receptive field. The spatial resolution of the achromatic channel is greater than that of the chromatic channels; these spatial frequency characteristics, and their consequences when viewing artificially created images, are investigated in Chapter 7 of this thesis.

Signals carried by ganglion cells are processed in the lateral geniculate nucleus (LGN - the first stage in the visual cortex), which also has cells with a centre-surround type of response, and with receptive fields larger than those of the retinal ganglion cells (Tsotos, 1984). Some feedback (possibly negative) from the visual cortex influences these cellular responses; very little, however, is known about such mechanisms. In the visual cortex itself, cell functions are varied: simple cells respond best to precisely positioned and orientated elongated luminance contrasts; complex cells are responsive to orientation of luminance contrasts, but less to specific position; hypercomplex cells respond in a similar but more spatially constrained manner, receiving input from both simple and complex cells (Gouras and Zrenner, 1981). Investigation of the functions of cells at this level is an active area of research; simple cells responsive to input from (spatially constrained) different opponent channels have been found (Tsotos, 1984), but as yet a clear indication of the method of formation of higher perceptual level sensations has not emerged. Gouras and

Zrenner (1981) give an extensive summary of the neural processing of colour signals, emphasising the limitations of current knowledge and the degree of conjecture involved in high level modelling.

The spatial frequency characteristics of the visual pathway have been widely studied in psychophysical experiments. Proposals that the pathway is comprised of a set of channels, each tuned to a different spatial frequency, have been made (originally by Campbell and Robson, 1968) Such a structure can account for experimental results on the response to threshold level grating patterns. While the existence of such multiple channels has been contentious in vision research, Georgeson (1980) has suggested that it would not necessarily imply any particular model of spatial encoding. A more neurophysiologically-specific approach by Wilson and Bergen (1979) proposes two different sizes and types of receptive fields, and two different types of temporal response characteristics. This model can account for apparent narrow frequency response bandwidths by invoking probability summation; this presents a more plausible explanation of spatial frequency response characteristics (Marr, 1982 p.10). It is quite possible that the visual system has the ability to use higher-level-determined attention mechanisms to concentrate on particular spatial frequency ranges; for example, the ability to concentrate on focussed information (of high spatial frequency) while disregarding out-of-focus information in a scene might be such a mechanism (Ginsberg, 1980).

2.1.2 Visual perception of colour - spectral perceptual attributes

At a perceptual level, attributes of a scene such as colour or depth are recognised. The distinct spectral attributes which are used to describe spatially point-specific visual sensation at a perceptual level are known as hue, saturation and lightness (or brightness). Hunt (1977,1978) summarises the terminology of colour appearance as judged perceptually, distinguishing between subjective terms and objective terms. Subjective terms are seen as those used to describe subjectively the principal attributes of sensations of light and colour; that is, spectral perceptual attributes. We give here the formal definitions of the subjective perceptual descriptors hue, saturation and lightness (or brightness), taken from the Commission Internationale d'Eclairage (CIE), as reported by Hunt (1978).

> *Hue*: "attribute of a visual sensation according to which an area appears to be similar to one, or to proportions of two, of the perceived colours red, yellow, orange, green, blue, and purple".
>
> *Saturation*: "attribute of a visual sensation according to which an area appears to exhibit more or less chromatic colour (*colourfulness*), judged in proportion to its brightness".
>
> *Lightness*: "attribute of a visual sensation according to which an area appears to reflect diffusely or transmit a greater or smaller fraction of incident light".
>
> (*Brightness*: "attribute of a visual sensation according to which an area appears to be emitting, transmitting, or reflecting, more or less light".)

The terms brightness and lightness are often confused; Hunt distinguishes between them on the basis of whether a stimulus comes from an apparently luminous source (emitting or specularly reflecting), or whether it comes from an apparently non-luminous source (transmitting or diffusely reflecting) respectively. Because of the difficulty in an artificial display situation of always determining this distinction, and because we treat surfaces which have both diffuse and specular reflections in later chapters, we shall throughout this work use the term lightness only, with the implication that when appropriate, it should be taken to mean brightness. We shall also use the term *saturation* to describe *colourfulness* defined above, rather than *colourfulness* in proportion to brightness, because of its widespread usage in this sense in the image display literature. (This corresponds to CIE *chroma*.)

Objective terms are based on the same spectral attributes, but relate either to the magnitude of the evoked response (psychophysical terms) or to the differences between the evoked response to pairs of stimuli (psychometric terms). A perceived hue can be given a psychophysical measure in terms of its dominant wavelength; a saturation in terms of its chromaticity co-ordinates; and a lightness in terms of its luminance. Wyszecki and Stiles (1967) should be consulted for definitions of these measures. Throughout this work we use the terms hue, saturation and lightness as the perceptual attributes or descriptors of a colour, relying on context to distinguish between subjective, psychophysical objective, and psychometric objective, meanings.

The important aspect of these three attributes is that they are sufficient to describe the difference between any pair of visually distinguishable point sensations, and that one or two of the attributes only can be insufficient. Note that this does not mean that different physical stimuli will necessarily be distinguishable. Apart from spectral resolution limitations, a consequence of the finite-dimensionality of colour vision is that two non-identical physical stimuli can appear identical (metamerism). Spectral attributes are also affected by adaptation. Adaptation describes the ability of the visual system to adapt to a steady-state viewing condition, such as an illumination level, or a chromatic cast, and still distinguish spectral attributes on a relative basis, given the rendering limitations of the illumination. Thus, because of the effects of lightness and chromatic adaptation, it is difficult to specify spectral attributes for a single stimulus; rather some basis for comparison is required. Consequently the hue, saturation and lightness of a stimulus are perceived in a sense relative to a reference stimulus, and in normal scene viewing conditions that comparison would be made relative to some derived overall or average conditions. Cornsweet (1971) discusses the factors affecting these subjective perceptual attributes, treating in some detail their possible low-level physiological correlates.

The degree of locality within the visual field from which this reference is derived is not clear, and may well vary as a function of attention mechanisms. Not enough is known about the spatial receptive fields of cortical cells, or about possible feedback from cortical to LGN levels, to suggest a specific physiological basis for such an interaction. Global scene adaptation may well be supplanted by local adaptation or induction effects, which specifically describe the influence of the attributes of bordering or surrounding areas on the perceived attributes of an area stimulus. Adaptation and induction effects have been widely studied and modelled, both from a broad mechanistic, and from a perceptual, point of view (see Bartleson, 1978,1979a or Wright, 1981 for summaries). There are good physical reasons for being able to judge perceptual attributes on a relative, rather than an absolute, basis. First, within limits, spectral perceptual resolution becomes essentially independent of illumination level and contrast; second, the chromatic attributes of a stimulus can be judged, again within limits, independently of the rendering effect of a particular illumination. This makes perceptual scene analysis fairly consistent under the very wide range of viewing conditions encountered in real-world illumination variations. Consequently we must be aware, at least, that the perceived attributes of a stimulus depend on both local and global viewing conditions.

That three distinct spectral attributes can be distinguished when judging a stimulus relative to some reference does not mean that these attributes can necessarily be treated as independent under all conditions. In discriminating between samples on the basis of perceptual attributes, a set of samples can be sequentially ordered using a process of intuitive extrapolation. It is this ordering process which allows us to use such an attribute to represent the relative magnitude of a data variable. But whether this intuitive extrapolation of order can be performed under all spatial and spectral conditions is another matter. In general, under conditions of spatial adjacency, judgements of the magnitude of differences, and of sample order, can be made if the samples being compared are close in perceptual terms; that is, locally in three-dimensional perceptual spectral space. Under different conditions of spatial comparison, or for samples very different in one or more than one perceptual attribute, it becomes more difficult to make such judgements. More specifically, we can summarise the inter-dependencies, under conditions of spatial adjacency, of perceptual attributes hue, saturation and lightness in terms of their use for the discrimination of differences between, and the ordering of, samples, as follows.

(1) Discrimination between hues is consistently possible, even under conditions of varying saturation and lightness. Ordering of hues is not particularly intuitive in an absolute sense, though for small hue differences, relative ordering becomes more intuitive. (Hue orders such as those associated with the spectrum order can be learned.)

(2) Discrimination between levels of saturation is possible for both large and small differences under conditions of constant hue, but much less so under conditions of varying hue. The effect of lightness variation on saturation discrimination can also be detrimental. Saturation ordering is intuitive when discriminations can be made.

(3) Discrimination between levels of lightness is possible for small differences under conditions of constant hue and saturation, but again less so when comparing widely different hues and saturations. Lightness ordering is intuitive when discriminations can be made.

These qualifications suggest that the three variables can be considered to be independent in a spectrally local sense, but not in a spectrally global sense. In a spectrally global sense, hue discriminations can still be made relatively finely, but saturation and lightness discriminating abilities (and hence ordering capabilities) are substantially reduced.

2.1.3 Structural scene comprehension

In the previous section we identified the spectral perceptual attributes which could be associated with any particular point in a visual field. We can think of a structural representation of the information over a visual field as being the result of analysing the variation in perceptual attributes over its spatial extent. By associating structural representations with physical properties, we can interpret the information in terms of these recognised physical properties: namely, surfaces and their coverings.

The method by which this spatial analysis is performed is far from clear, though the importance of recognising edges is suggested both on heuristic grounds, and on the physiological evidence of cells responsive to edges as outlined in section 2.1.1. Similarly some mechanism for detecting gradual changes must exist for the analysis of, for example, shape from lightness variations. The association of changes in perceptual spectral attributes with physical processes giving rise to them can form the basis for developing plausible models of portions of the structural scene comprehension process; such models are discussed in section 2.2.1.

Whether such physical interpretations of spectral variations are always, or in most cases, justified is not clear; what is important is that while perceptual spectral attributes at any single point in a visual field do not necessarily indicate useful information about physical properties at that point, the attributes gauged in relation to their neighbourhoods do. In other words the spatial variations at all frequencies provide the important information about physical scene properties. The visual system is sufficiently experienced at analysing real-world scenes for the associations of perceptual attribute spatial distributions with physical properties to be intuitive. For example, surface form, or topography, can be intuitively appreciated from appropriate lightness variations (Horn, 1981); similarly shadows, textural characteristics and covering characteristics can be immediately appreciated, and an implicit understanding of the general spatial nature of the physical properties results.

In the following section we look at the extent to which considerations of visual system models and processes have been used in the development of image display approaches or techniques.

2.2 Visual system models and their use in colour image processing

Visual system models attempt to explain physiologically and psychophysically observed effects by proposing mechanistic interconnections which are substantiated by, or at least in accordance with, physiological evidence and known neurophysiological interactions. The application of visual models in image processing can be for image enhancement, compensation for visual characteristics, or supplying a measure of information content which is visually significant for data compression. Effectiveness in such a process does not necessarily validate a model on which it is based.

2.2.1 Visual models - their constraints and limitations

Models which link receptor stimuli to perceptual attributes have been proposed in the context of image processing by Faugeras (1976,1979) and Hall and Andrews (1978; Andrews and Hall, 1978). Each of these models treats opponent channels, derived from specific receptor combinations with appropriate non-linear elements, as perceptually significant; the spatial frequency bandwidth of each opponent channel is included in the model. Application in the context of this work of the perceptual colour spaces so formed is considered in the following chapter. One strength of Faugeras' work is the physical basis behind the processing approach; in providing spatial access to the opponent channel signals, he allows appropriate emphasis or de-emphasis of information closely related to the physical scene variables, and hence of the information the visual system has to extract from the scene. Gagalowicz (1982) has shown some limitations of Faugeras' derived perceptual channels when applied to the discrimination of colour texture fields; this suggests that the receptor-opponent formulation is for some purposes too simplistic in its assumptions. This conclusion is reinforced by more recent neurophysiological findings (Gouras and Zrenner, 1981).

Outside the context of image processing many models which treat specific visual factors in a receptor-opponent zone model have been proposed. Brightness perception of colours forms the basis of the vector model of Guth (1972; Guth et al., 1980), while other studies have used other empirical colour matching data to determine receptor-opponent interconnections (for example, Vos and Walraven, 1971). Paulus and Kröger-Paulus (1983) have proposed a model which is not specific in these interconnections, but rather relies on centre-surround interactions for its formulation. In any such model, if the opponent zone is treated as perceptually significant, it can provide a basis for appropriate processing, requiring only some derived analytical access. The inclusion by Faugeras, and Hall and Andrews, of spatial-frequency dependence in their modelling of the low level pathway forms a significant advance over models which consider only the spectral nature of the receptor-opponent interconnections.

A model which considers photoabsorption and signal transmission in the context of statistical communications theory, treating the brain as an optimum processor for signal detection, has been proposed by Buchsbaum (1981). Though this model does not specifically include temporal and spatial variation, this type of approach can be used as a basis for processing techniques which optimise some statistically derived detection measure. Pearlman (1978) uses such a measure of distortion, derived from a spatial-frequency-channel-specific eye-brain model, as an achromatic image processing rationale. Massof and Bird (1978) have also applied a stochastic approach using Guth's (1972) opponent-based vector colour space model, and have shown that many spatial and spectral visual properties can be explained in such a framework. Such models have a role in the prediction of measurable visual effects, but require the additional stage of relating statistically derived measures to the more complex physical attributes of scenes to be directly useful in a display context.

Land proposed a theory of vision, based on a visual system model of separately identifiable retinex images (one for each receptor type), which recognised the importance of deriving physically significant information from a visual image (see Land, 1977 for a summary). Land suggested a (one-dimensional) method for removing the (slowly varying) illumination component from a scene, resulting in extraction of surface reflectance information. Horn (1974) showed that a two-dimensional analogue of this method was theoretically feasible, and pointed out that images derived from linear combinations of the originally proposed receptor retinex channels could equally well be used. While Land's theory does not adequately explain all perceived visual effects (Marr, 1982 p.257), and now looks physiologically implausible, it is conceptually attractive in its physical basis; the removal of the slowly varying lightness component effectively amounts to processing in a perceptual domain, with potential for image enhancement operations.

Visual system modelling to a higher, or structural, level presents severe problems due to the greater complexity of the physiological processes involved. Barlow (1981) outlines the extent to which the number of cells used to process a particular piece of visual information increases dramatically at cortical levels. Marr (1982) warns of the risk of being misled in assuming mechanistic functions of visual pathway components, on the basis of response to stimuli, in an obviously complex system in which individual components may well have non-obvious functions. As a result of this lack of specificity in high level mechanistic knowledge, visual system modelling at a high level has largely involved consideration of the purpose behind visual processing.

Investigations into the possible structural analysing processes which might be performed by the visual system have, in the field of artificial intelligence, looked at the physical property carrying the information, and then considered ways of isolating that physical property. Both abrupt and gradual variations in perceived attribute provide information about the intrinsic structure of a scene. For example, Horn (1977) has shown that the shape of non-occluded smooth surfaces can be derived from non-linear local variations in lightness; lightness edges can be caused by surface height discontinuities or shadows; colour discontinuities can suggest surface material boundaries. Marr (1982, p.261) goes further to suggest that non-linear lightness changes which are not discontinuities due to shadows or surface orientation changes can be assumed to be due to surface property variation, and as a consequence that surface property information can result from analysis of non-linear local changes in lightness and spectral distribution. Rubin and Richards (1982) also suggest spectral attribute associations with the physical properties of material boundaries, shadows, highlights, surface orientation, and pigment densities. Such interpretations have been developed and used widely in the context of extracting scene characteristics from images (see Brady, 1982 for a summary). Surface shape can also be derived from consideration of textural properties under certain conditions by local analysis of lightness variations (Kanatani, 1984).

At a higher level of scene analysis, models which attempt to deduce three-dimensional geometrical scene characteristics from edge information have been proposed (Tenenbaum et al., 1974). Detection of edges can form the first stage of a primitive generating process; Marr (1982), Haralick et al. (1983), Tsotos (1984), Rosenfeld (1984), and Haralick and Shapiro (1985) summarise such modellings. These investigations lead both to refining models of higher level visual processes, and also to the development of computer vision systems for applications in robotics. Computational techniques to test hypotheses of such models have been summarised by Brady (1982).

The strength of these approaches lies in their consideration of the physical significance of the tasks the visual system performs. While mechanistic models can be used to predict observed effects, they cannot necessarily be used to realise a physical effect, in terms of perceptually significant attributes, in a realistic manner.

2.2.2 The use of visual models and considerations in colour image processing

This section summarises techniques, based on consideration of visual mechanisms and processes, which have been used in colour image processing. The purpose and rationale behind each technique only is outlined; where appropriate, subsequent chapters treat individual processes more thoroughly.

Spatial-frequency dependent processing based on specific mechanistic models

Faugeras (1976) showed that by using enhancement methods based on low level visual properties (the spatial frequency characteristics of proposed opponent channels), the apparent quality of a reproduced colour picture could be improved (see Chapter 7 for further details of this process). He also applied his model to image compression. Hall and Andrews (1978) used a similar approach, based on an earlier-developed visual model (Hall and Hall, 1977), to develop image compression methods based on an application of information theory to the visual model output. It should also be mentioned that the standard broadcast television compression of colour picture information, into a bandwidth not much greater than that required for monochrome picture transmission, results from recognition of the limitations in chromatic resolution of the visual system (see Hunt, 1975).

Representation of data by perceptually significant attributes

Recognition of the importance of perceptual attributes for representing data variables has resulted in a large number of realisations of perceptual attributes hue, saturation and lightness in various colour spaces (see Chapter 3 for a critical discussion on these realisations). The purpose behind these methods is twofold: first, to improve intuitive appreciation of data variables so represented; second, to facilitate spectral decomposition at any point in an image.

Proportional representation of data differences by perceived colour differences

Display techniques have also used perceptually uniform colour spaces (spaces in which the metric uniformly reflects perceived colour differences) to achieve proportional representation of numerical data differences by perceived colour differences. Such an approach allows an intuitive feel for the magnitude of data variations to be gained. Juday (1979) has proposed this approach for the display of multi-spectral data; Juday (1978), Juday et al. (1978) and Balon and Cicone (1979) have investigated the application of this approach to LACIE image products. Meyer and Greenberg (1980) suggest using such spaces for computer graphics in general. Others (O'Callaghan et al., 1981; Robertson and O'Callaghan, 1982,1984; Tajima, 1983) have extended the use of uniform colour spaces to various data types. These approaches are discussed in greater detail in later chapters.

Structural representations

Techniques which give three-dimensional structural representations to spatially two-dimensional data variables to improve comprehension have been used widely in graphic data display (see Bertin, 1981; Grotch, 1983), and also in image display (Arvidson et al., 1982; Robertson, 1984; Robertson and O'Callaghan, 1985). The rationale behind these techniques is that while spatially two-dimensional data presentations are not always immediately and intuitively comprehensible, a two-dimensional view of a three-dimensional representation, with appropriate monocular depth cues such as occlusion, scale or shading, can be immediately comprehensible. Image processing software packages which display variables in the form of a three-dimensional surface are available (for example, TASC or UNIRAS software). Chapter 6 considers the scope and limitations of these methods.

The preceding sections have given a summary of visual considerations relevant to the display of image data in colour, outlining the manner in which such considerations have been used in an image processing context. We now develop an overall approach to image display on the basis of these visual considerations.

2.3 Overall requirements for image display - approach taken in this work

2.3.1 Relevance of visual system operation to the display of information in colour

In analysing real-world scenes the visual system manages to process the very large amount of information it receives by its ability to interpret spatial variations in perceptual spectral attributes as corresponding to physical property variations. Patterns are intuitively appreciated, allowing extraction of the desired spatial nature of individual physical properties. Interactions, due to the overall scene structure, between physical processes are also intuitively comprehended, and compensated for, when interpreting the physical property variations. Representation of real-world scenes in a spatially two-dimensional form, such as in a colour photograph, leaves most of these associations and intuitive comprehension processes still operative. This is because most depth cues are still present (see Marr, 1982; Wolfe, 1983).

If, on the other hand, arbitrary and dissimilar spatial arrangements of each of several distinct colours are combined in a spatially two-dimensional image form, the result is confusing and difficult to interpret. Not only does the spatial structure of such an image not depict a plausible real-world scene; the spectral representation also does not have any significance as corresponding to physical properties in the presented display. We suggest that, as a general maxim for image data display, if several data variables with dissimilar spatial structures are to be simultaneously displayed, then intuitive comprehension of the spatial properties of these variables is more likely to be achieved if the chosen data representation results in a type of scene familiar to the visual system: namely, a representation which can be interpreted as surfaces and their coverings. Further, we suggest that data representations which in some way depict physical properties of a scene, by appropriately using perceptual spectral attributes, are more likely to be intuitively appreciated than those which use arbitrary spectral assignments. This is not to say that in all representations of all data variables such an approach is necessary for data appreciation; rather it is suggesting that when a presented image reaches a certain level of complexity, in terms of the number of individual data variables presented and their spatial structures, overall comprehension is more likely to be achieved if the intuitive scene analysing capabilities of the visual system are exploited.

We suggest that a data display must be designed such that the data variables can be intuitively extracted; this is a prerequisite for effective appreciation of the spatial nature of the data variations, and relies on making use of the normal visual scene analysing processes. We thus propose that a comprehensive display approach must embrace the following basic principles.

(1) A data display should be recognisable in a structural sense as a realistic scene.

(2) Individual data variables should be depicted as plausible physical properties of the scene by choosing appropriate perceptual spectral representations for them.

(3) The requirements for structural and perceptual representation should be satisfied before enhancements, or compensations for visual system mechanisms, are applied; otherwise such enhancements will have very limited use.

This leads to a three stage process when designing an image display; this process is depicted in figure 2.1. We suggest that all image data display situations can be considered within the context of these principles. The first step to realising them is to satisfy the requirements for structural scene comprehension.

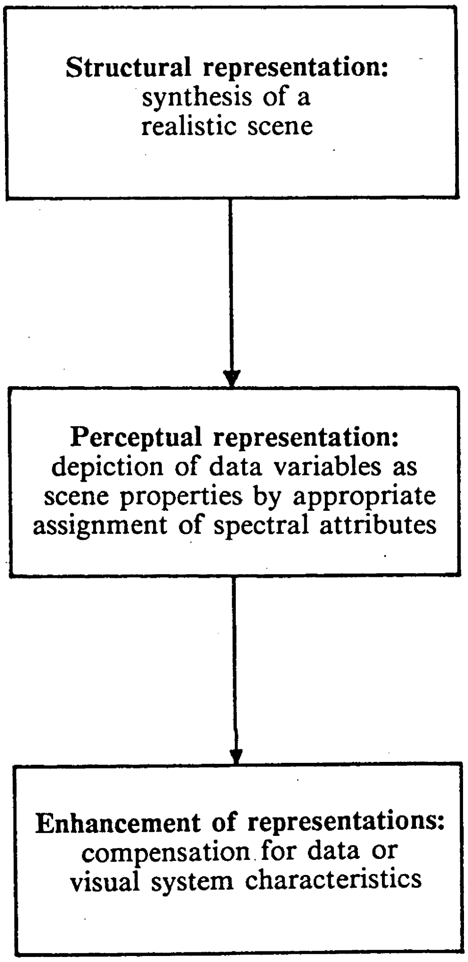

Structural representation:
synthesis of a
realistic scene
Perceptual representation:
depiction of data variables as
scene properties by appropriate
assignment of spectral attributes
Enhancement of representations:
compensation for data or
visual system characteristics

Figure 2.1 Schematic of the proposed display approach

2.3.2 Structural scene comprehension requirements

In many data displays, structural comprehension will be inherently satisfied. For example, a map displaying a single variable can be structurally recognised as a flat surface with spectral variation (representing the data variable) over that surface. Similarly a photograph, in which appropriate lightness variations portray the surface structure, will be intuitively comprehended structurally. The display of such data types is treated in Chapters 4 and 5. However, if the requirements for intuitive structural comprehension are not inherently satisfied in the data, such as in the earlier-cited example of superimposed data variables with dissimilar or arbitrary spatial structure, the display methodology must somehow provide a basis for this structural comprehension. We approach this in Chapter 6 by synthesising realistic scenes, consisting of surfaces and their coverings, and representing data variables by appropriately chosen natural scene variables. This should then allow intuitive extraction of scene variables, and hence data variables.

To achieve realistic scene synthesis, and the representation of natural scene variables, we use models of physical processes involving illumination of, and reflection from, surfaces. We also use the results from computational approaches to scene understanding; these results in general relate spectral perceptual responses to the physical processes causing them, describing the role of lightness and chromatic variations in surface topography depiction, and in reflection from the surface of materials. It is the realistic representation of these physical processes that should allow intuitive scene appreciation. This means that we must also be able to realise the appropriate spectral perceptual attributes (hue, saturation and lightness) in the synthesised scene.

2.3.3 Realisation of spectral perceptual attributes

Perceptual attributes which represent data variables either directly, or in some combination related to a chosen physical process, must be realised sufficiently closely to allow their intuitive extraction from a presented scene. Ideally we would like to consider the basic spectral perceptual attributes, (hue, saturation and lightness) as dimensions of a spectral perceptual space into which data can be mapped. Such a space would then be addressable in these terms. We would also like distance from some defined origin within the space to reflect psychophysical colour judgements, and the space metric to reflect psychometric judgements of colour differences. The extent to which we can achieve representation within such a space is considered in Chapter 3.

2.3.4 Enhancements and compensations based on mechanistic knowledge

If we can achieve the earlier-stated requirements for intuitive appreciation of the presented data, we can then consider ways of enhancing appreciation of the data characteristics, or compensating for particular known visual system limitations. This area has received widespread attention on an ad-hoc basis, both in the intensity domain (contrast enhancement, for example), and in the spatial domain (edge enhancement, for example), and also on a basis of considering visual system characteristics (Faugeras' work, for example). But we stress that if the structural and perceptual decomposition requirements are not satisfied in a scene, gain in the appreciation of data characteristics is unlikely to be made using low-level enhancements or compensations alone. We examine such low-level effects in Chapter 7, considering the spatial-frequency dependence of the contrast sensitivities of the visual system achromatic and chromatic mechanisms.

2.4 Summary

This chapter has investigated aspects of the operation of the human visual system pertinent to appreciation of the characteristics of image data variables represented in a composite colour image.

The importance of ensuring that overall structural comprehension of a presented display can take place before spectral interpretation can be performed has been stressed. A display approach which achieves this structural comprehension by presenting the display in the form of a realistic scene has been proposed. This involves representing the data variables as natural physical properties of this scene, achieved by modelling such properties in terms of their perceived spectral effects.

The importance of an appropriate spectral representation of perceptual attributes, both for realistically representing scene properties, and for direct spectral decomposition purposes, has also been emphasised. This suggests the use of a perceptual space representing these spectral perceptual attributes as central to the framework for realising the display approach. This framework is developed in the following chapter.

The role of enhancements based on, and compensations for, low-level visual system mechanisms and properties has been discussed; it is suggested that unless the structural and perceptual requirements for scene appreciation are satisfied, such enhancements and compensations can have little effect on improving data interpretability.

Chapter 3
Development of a framework for image display

This chapter develops a framework, based on a three-dimensional perceptual colour space, for the processing and display of multi-dimensional image data sets as described in following chapters. Section 3.2 looks at the requirements, derived from the approach to image display outlined in Chapter 2, for this perceptual colour space. Various types of colour space are described and examined for suitability, and a framework satisfying the requirements is proposed. Section 3.3 looks at the feasibility of performing the desired spectral and spatial transformations within this framework, while section 3.4 treats its realisation. Colorimetric terms and concepts used in this chapter are defined and explained in Appendix 1.

3.1 Basic form of the framework

As stated in the previous chapter, we assume temporal invariance both in data presentation, and consequently in visual system adaptation level.

In the following specification of transformations, we use the basic form of Pratt's (1978, p.346) notation. An operator $\Phi\{\,.....;.....\,\}$ specifies the transform; the variables before the semicolon indicate the functional dependency of the operation, while the term after the semicolon is the function to be operated on. Operator subscripts in upper case denote the domain and range spaces of the mapping, with lower case suffixes indicating purely spatial (x), or purely spectral (λ), dependencies. Spectral space co-ordinates are specified in upper case (for example, $U1, U2, U3$) while conventional lower case spatial co-ordinates x,y,z are used, z representing the third spatial dimension of depth.

We define a real-world spatially three-dimensional and spectrally three-dimensional perceptual space **P**, defining spectral variables H, S and L as cylindrical polar co-ordinates of **P**. In the general case we wish to map informative image data variables from a data space **D** into perceptual variables (in **P**); that is,

$$P(H,S,L,x,y,z) \;=\; \Phi_{\mathbf{DP}}\{D1,...,Dn,x,y\,;D(D1,...Dn,x,y)\}.$$

We may also wish to perform spatial (in a structural sense) or spectral (in a high level perceptual sense) processing in **P**:

$$P'(H,S,L,x,y,z) \;=\; \Phi_{\mathbf{PP}}\{H,S,L,x,y,z\,;P(H,S,L,x,y,z)\}.$$

These processes are considered in Chapter 6.

We define the displayed image perceptual space **U**, as spatially two-dimensional and spectrally three-dimensional. We use rectangular co-ordinates for both spatial and spectral representations of **U**. It is convenient to use **U**, with its rectangular spectral co-ordinate system, rather than **P**, with its cylindrical polar spectral co-ordinate system, as central to the framework.

Transforming from **P** to **U** can be treated as a two-stage process. First, a projection from three to two spatial dimensions is made:

$$P'(H,S,L,x,y) \;=\; \Phi_{\mathbf{PP},x}\{x,y,z\,;P(H,S,L,x,y,z)\}.$$

Second, a transformation of spectral variables from (cylindrical polar) perceptual attributes to rectangular co-ordinates is made:

$$U(U1,U2,U3,x,y) \;=\; \Phi_{\mathbf{PU},\lambda}\{H,S,L\,;P'(H,S,L,x,y)\}.$$

We consider in detail in Chapters 4 and 5 cases where no spatial projection is performed. In Chapter 5 it will also be evident that we may wish to map informative data variables directly into **U**:

$$U(U1,U2,U3,x,y) = \Phi_{\mathrm{DU}}\{D1,...,Dn\,;D(D1,...,Dn,x,y)\}.$$

Similarly in some cases we may wish to perform spectral processing (for example, to compensate for the effects of induction) in **U** rather than in **P**:

$$U'(U1,U2,U3,x,y) = \Phi_{\mathrm{UU},\lambda}\{U1,U2,U3,x,y\,;\,U(U1,U2,U3,x,y)\}.$$

In addition, as suggested in the previous chapter, we may require access to lower levels (symbolically termed **M**) of the visual pathway to allow for spatial and spectral processing for enhancement or compensation purposes. This involves a series of transformations of the form

$$M(M1,...,Mn,x,y) = \Phi_{\mathrm{UM},\lambda}\{U1,U2,U3\,;\,U(U1,U2,U3,x,y)\};$$

$$M'(M1,...,Mn,x,y) = \Phi_{\mathrm{MM}}\{M1,...,Mn,x,y\,;M(M1,...Mn,x,y)\};$$

$$U'(U1,U2,U3,x,y) = \Phi_{\mathrm{MU},\lambda}\{M1,...,Mn\,;M'(M1,...,Mn,x,y)\}.$$

Chapter 7 looks at requirements for transforming between **M** and **U**, and at appropriate mechanism-dependent processing.

Finally we must be able to realise colours specified in **U** on any chosen colour display device. This involves a transformation to display device gun-count space **J**:

$$J(J1,J2,J3,x,y) = \Phi_{\mathrm{UJ},\lambda}\{U1,U2,U3\,;\,U(U1,U2,U3,x,y)\}.$$

The derivation of this transformation is treated in section 3.4.

The overall structure of the display framework is shown schematically in figure 3.1.

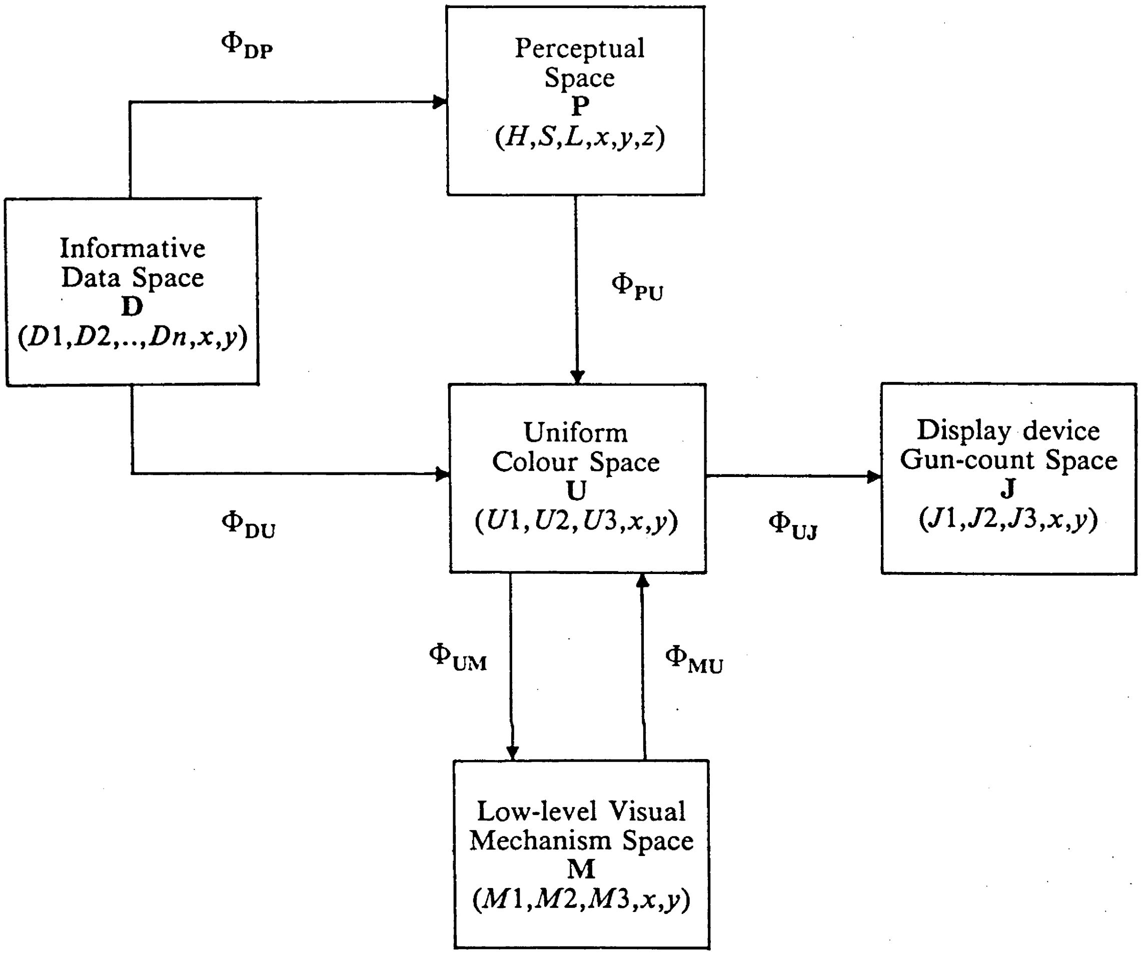


Figure 3.1 Schematic structure of the display framework

3.2 Perceptual colour space central to the framework

In this section we consider the properties required of the perceptual space **U** central to the framework, and evaluate the suitablity of various spaces derived from display device and visual system characteristics.

3.2.1 Requirements of a perceptual colour space

The perceptual space **U** which is used as the core of the display framework should have the following spectral properties.

(1) The space should be addressable in perceptually meaningful terms to allow representation of data variables by combinations of appropriate perceptually significant attributes. This should aid intuitive extraction of data variables from a display, contributing to overall scene comprehension. Addressability in perceptual terms can also allow compensation for high level visual effects.

(2) The metric of the space should have perceptual significance; that is, it should be possible to relate the size of perceived colour differences to the size of numerical data steps. This implies that there should be no discontinuities in perceived colour in the bounded region of the space accessible by a display device.

(3) It must be possible to realise colours specified in the colour space on colour display devices by providing a transformation to display-device addressing terms.

In addition, it would be advantageous if it were possible to analytically access lower visual levels from the image perceptual space **U**. This would allow spatial or spectral processing based on what is known of the low, or mechanistic, level of operation of the visual system. Compensations or enhancements applied at such a level can exploit known effects to improve image quality, and hence data appreciation.

These requirements can be used as a basis for testing the suitability of various types of colour spaces considered in the following sections.

3.2.2 Colour spaces dependent on display devices or processes

Colour display systems incorporate various types of display devices or processes, including colour television monitors, photographic film recording processes, and ink deposition processes. Combinations of these processes can also be used; for example, colour transparencies can be made by photographing a television monitor. Display devices are generally addressed in terms of three colour gun-counts, which give the relative or absolute proportions by which each of the three colour guns are to be activated. The actual colours of the colour guns are termed the primaries. The use of the terms primary and complementary in the literature is varied. In this work we use primary as defined in the field of colour science: "primary colours are the colours of three reference lights by whose additive mixture nearly all other colours may be produced" (Wyszecki and Stiles, 1967). Thus any particular display device has its own set of primaries; such primaries are distinct from those relating to the human visual system, which are known as the fundamental primaries. The complementary to any particular colour is that which when additively mixed in suitable proportions with the colour yields a match with a specified achromatic colour. Thus any colour has a complementary colour, again distinct from the fundamental complementaries relating to the human visual system. The concept of a complementary colour, however, does require definition of an achromatic colour; in this work we take the UCS grey axis (see later sections) to be achromatic. (Note that formally (Wyszecki and Stiles, 1967), the term complementary is used to refer to a wavelength, being complementary to the dominant wavelength of a particular colour; in perceptual spaces, dominant wavelength and hue are seldom uniquely related (MacAdam, 1981), and a hue-dependent interpretation is more useful.)

A display device with given primaries can produce a range of colours, or volume in colour space, determined by its primaries. It is convenient to represent this space in the form of a cube, with the primaries being the orthogonal axes of the cube, and with all realisable colours lying within the cube. This representation is shown in figure 3.2(a) for an additive display process with red, green and blue primaries. We refer to this representational space as a BGR cube. Diagonally opposite corners of the cube,

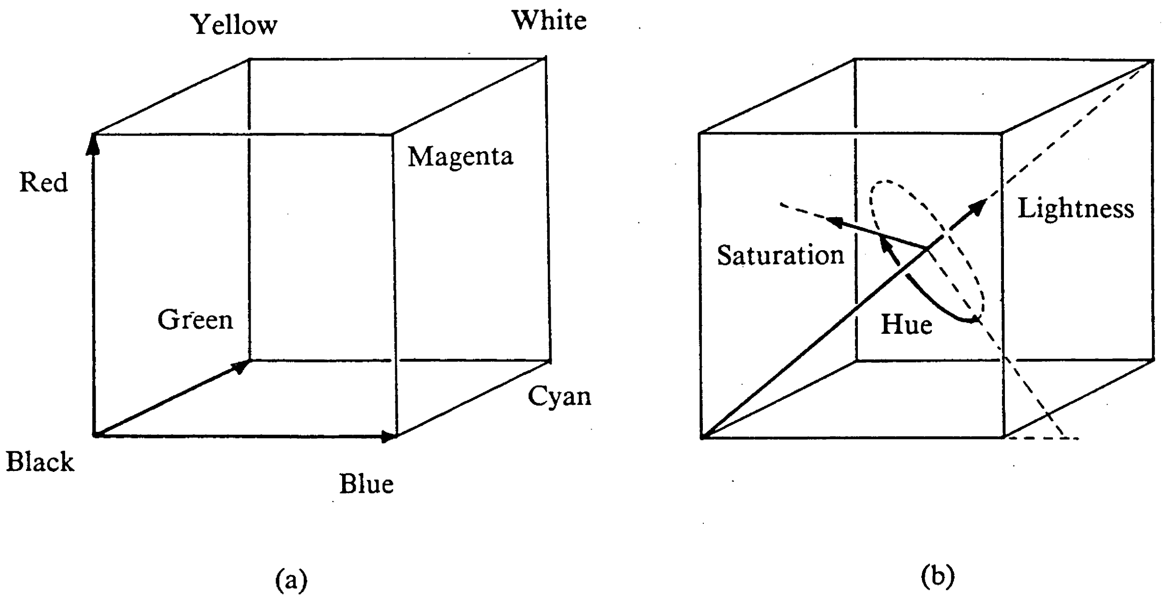


Figure 3.2 Display device BGR-cube colour space representations
(a) Addressing in terms of cube axes
(b) Addressing in terms of geometrically defined perceptual attributes

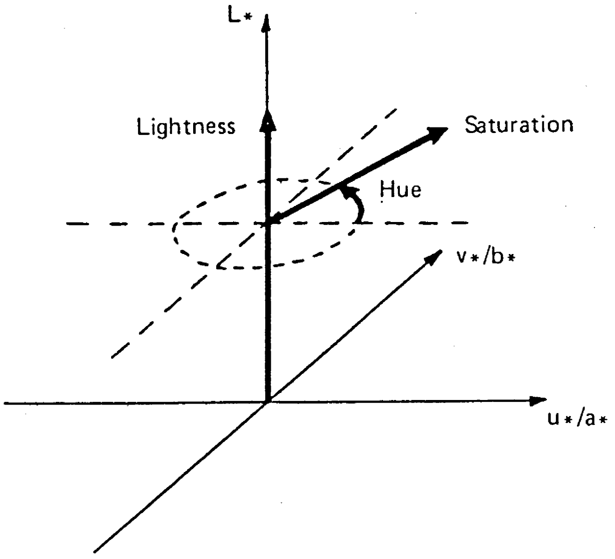


Figure 3.3 UCS representation of perceptual attributes hue, saturation and lightness

corresponding to minimum and maximum excitation of all three guns, represent colours black and white respectively. It should be noted that the white so formed is a subjective or nominal white which will change if the maximum excitation levels of the guns are changed. For directly-subtractive type display processes such as ink deposition systems, the colour guns are generally cyan, magenta and yellow, with all three being used to create black or to darken colours. The white point is at the zero excitation level of the guns in such a process.

In general, the end points of cube axes, and the relative density distributions along the axes, will not be the same for non-identical display devices. Note that relative density refers to the measured relative signal strength; if the spectral characteristics of a single gun are independent of its level of excitation, relative density can be measured using an instrument with any spectral response function which overlaps the spectral range of the gun. If this is not the case, or for display devices with cross-talk between an excited gun and the primary corresponding to a gun not excited, measured relative density is dependent on the spectral response characteristics of the instrument. Appendix 1 treats the measurement of colour more thoroughly. The result of activating more than one gun also varies amongst display device types. In additive-type devices, such as colour television monitors, the nett effect can be to increase overall lightness; in subtractive-type devices (such as those incorporating a photographic process) the nett effect is more complicated, and depends on the number of process stages and dye absorption characteristics (see section 3.4). The extent of cross-talk also varies substantially depending on the type, or even model, of display device. Consequently a given device cube specification does not in general give rise to the same colour, and further, a given device cube specification range does not in general give rise to the same colour progression, on different display devices. This means first, that the metric of the display-device cube is not perceptually significant, and secondly that perceptually significant attributes or dimensions have no well-defined description in such a space. The extent to which display-device colour spaces do in fact vary is shown later in this chapter.

These problems can be redressed somewhat by the common technique of density linearisation, which re-scales the addressing counts according to measured relative density along some chosen axis of the cube to produce a density distribution which varies linearly along that axis. The visual system has a response to the intensity of a stimulus which is close to logarithmic (Cornsweet, 1971); this means that a logarithmic density function provides a reasonable linearisation function. Density linearisation can be performed along each primary axis, but this does not necessarily, and will not in general, cause linearisation of the perceptually important grey axis. Instead, density linearisation is usually performed on this grey axis, and will generally hold for any line closely aligned with, and not far translated from, this axis. However, it will not in general hold for lines which do not satisfy those conditions. More sophisticated forms of linearisation (for example, McDonnell, 1980) are treated in section 3.4.1.

The second cited problem, that of the lack of perceptually significant addressing axes in a display-device cube, can be approached by geometrically transforming to appropriate significant perceptual variables. Figure 3.2(b) shows a display-device cube with perceptual variables hue, saturation and lightness defined within it. This type of transform, know as an HSI or HSL transform, has been widely used to introduce perceptual addressability to display-device colour spaces (Tenenbaum et al., 1974; Raines, 1977; Buchanan, 1979; O'Callaghan, 1979; Buchanan and Pendergass, 1980; Gillespie, 1980; Daily, 1983; and others). However, perceptual variables so defined are not necessarily close realisations of perceptual spectral attributes, but rather approximations to these which depend on the display device. Such geometrical representations can also have problems in their implementation, giving rise to anomalies in variable realisation (see Kender (1976) for an analysis of these anomalies).

Such implementations of perceptual variables may give rise to products more intuitively interpretable, but their limitations as described above suggest an approach which attempts to realise perceptual attributes more truly. Coupled with the problems of density linearisation throughout a three-dimensional volume, device-dependent colour spaces come far from satisfying the basic requirements of a suitable space stated in section 3.2.1.

Colour spaces can also be based on the principles of colour mixture, either additive (the Ostwald System, for example) or subtractive (the Hickethier System, for example). The Ostwald System spaces "pure" hues according to their perceptible difference around a circular perimeter, and similarly spaces greys along a lightness axis, but does not account for relative perceptibilities elsewhere in the space. These spaces are process-dependent, and although addressable in terms significant to colour formulation, and in the Ostwald System in a perceptual ordering sense, they have the other drawbacks described in this section.

3.2.3 Colour spaces based on models of the human visual system

Not all of the models of the visual system discussed in Chapter 2 isolate a perceptual space in which the perceptual attributes of colour can be realised directly. The opponent channel models treat the space which is composed of these channels as perceptual, without fully substantiating the relation between psychophysically derived perceptual attributes and this lower opponent channel level. Thus, for example, in Faugeras' model, the emphasis is not on validating the perceptual space derived from the chosen receptor combinations, but rather on showing that a space so derived can sufficiently closely approximate perceptual attributes to be useful in a picture processing context. This is rather different from attempting to represent data by perceptual attributes. Such a limitation does not mean that we cannot use these models for processing at the opponent channel level, as will be seen in Chapter 7; rather it means that they may well not provide the most suitable approximation to a perceptual space. Similarly the vector model of Guth et al. (1980), which is also specific in terms of receptor combinations, does not lay claim to complete validation at the perceptual level. It performs mechanistic discriminations which concur with psychophysically observed results, but this again does not mean that the specified perceptual variables necessarily closely represent psychophysically determined perceptual attributes. The probablistic theory of vision developed by Buchsbaum (1981) does not provide a representation of perceptual attributes which can then be inverted for realisation; similarly Land's (1977) theory of vision does not lend itself easily to this approach.

Perceptual spaces can also result from defining the space metric in terms of a line element representing a just-noticeable colour difference. This line element can be derived in terms of fundamental visual mechanisms, or it can be empirically derived from experimental matching data. Elements of the first type (which include Helmholtz, Schrodinger and Stiles line elements) are termed inductive, and they do not account for some of the basic experimental results of colour difference perception. Empirical line elements are discussed in the following section. Wyszecki and Stiles (1967) give a thorough treatment of line elements (pp.511-560; see also Vos, 1979). Ingling and Tsou (1977) have proposed a metric based on an opponent colour system line element; Rich (1980) reports on its usefulness for describing foveal aperture colour vision, but its limitations in describing the perceptual discrimination of surface colours. Thus spaces derived from these metrics suffer from the same drawback as other spaces derived from mechanistic models; they do not embody the experimentally observed perceptual attributes of a colour space as a primary consideration. On this basis, spaces so derived are not well suited to form the required perceptual colour space.

3.2.4 Colour spaces based on empirical measurements

The final type of colour space we consider is that of spaces derived from psychophysical experiments. These experiments have generally involved colour matching under controlled conditions, making judgements on the relative size of compared colour differences between pairs of samples closely similar in one or two of the perceptual attributes (hue, saturation and lightness). This type of space can be realised either as a set of samples against which unknown colours can be compared, or in terms of an absolute colour specification system (such as CIE tristimulus values - see Appendix 1).

The first type of space, of which the Munsell Color System is perhaps the best known, uses pre-prepared samples which are ordered according to hue, saturation and lightness (or in Munsell terms hue, chroma and value), and spaced uniformly on the basis of judged small colour differences, for comparision with unknown colours. Another such system, the DIN-Color System, has its lightness defined as being a logarithmic function of relative CIE lightness, rather than being absolutely defined as in the Munsell System. (The Ostwald System, as noted earlier, incorporates the ordering aspect of a perceptual space, but uniform spacing only in a very limited sense; it is also display-process dependent.) The Coloroid Color System (Nemcsics, 1980) is an order system developed to result in a balanced distribution of perceptual attributes when viewed as a whole, rather than when viewing small colour differences.

The second type of space results from attempting to explain the results of colour judging experiments by some defining formulation, and is based on the concept of using a just-perceptible (or just-noticeable) colour difference (JND) as the metric unit of a uniform scale. For example, if chromaticness is held constant, a scale of samples just perceptibly different in lightness can be constructed; such a scale is termed a uniform lightness scale. Similarly if lightness is held constant, samples just perceptibly different in chromaticness can be arranged in a lattice. While this lattice in general will not lie in a plane (the geometry of colour perception space has been found not to be Euclidean - see Wyszecki and Stiles (1967) or MacAdam (1981)), a planar triangular lattice forms a reasonable approximation to the experimental results. Such a lattice is termed a uniform chromaticness scale. Uniform lightness and chromaticness scales can be combined to form a three-dimensional space (which again only approximately describes the experimental results) which is termed a uniform colour space (henceforth abbreviated to UCS).

Many co-ordinate systems which define an approximation to a UCS have been proposed, and their relative merits have been considered by the CIE (Commission Internationale d'Eclairage) committee on colorimetry. Most recently (CIE, 1978) the CIE has recommended the use of two uniform colour spaces, one for use under reflected light viewing conditions (the CIELAB system, with (L^*, a^*, b^*) co-ordinates), and one for use in additive light source conditions (the CIELUV system, with (L^*, u^*, v^*) co-ordinates). These spaces are defined in terms of a reference nominally white maximum illumination, and details of their analytical formulations are given in Appendix 2. In addition, the CIE has supported an effort by the Optical Society of America to construct a set of uniform scales, based on a regular rhombohedral lattice (being the three-dimensional analogue of a two-dimensional triangular lattice), producing a UCS which we term OSALJG. MacAdam (1981) and Nickerson (1981) should be consulted for full details of this space.

The advantages of such spaces is not only their approximate perceptual uniformity, but also that the experiments performed in their derivation involved making perceptual judgements. This results in a natural aligning of perceptual attributes within the space. With lightness held constant, hue and saturation form polar (angular and radial respectively) co-ordinates within a chromaticness plane. Including lightness variations results in a cylindrical-polar perceptual co-ordinate system for the UCS. This is illustrated in figure 3.3.

It should be noted that a perceptual space can also be constructed using empirically derived line elements. While the same problem of dimensionality arises with this approach (the Riemannian space of three dimensions resulting from empirically derived line elements cannot in general be mapped into a three-dimensional Euclidean space in such a way that distance is preserved), an approximate mapping into a three-dimensional Euclidean space can be made. Wyszecki and Stiles (1967, Chapter 6) describe these mappings in detail, presenting the results of line element constructions in terms of discriminating (MacAdam) ellipses (loci of constant line-element distance around given chromaticity values, calculated under constant luminance).

3.2.5 Choice of a perceptual colour space

It should be borne in mind that in choosing an appropriate colour space we are not trying to develop or validate a theory or model of vision; rather we are trying to find a practical representation of psychophysically measured, perceptually significant, visual spectral attributes. We wish to use these attributes to represent data variables in order that the data variables can be appreciated in an intuitive way. Consequently it is not sufficient to show that appropriate image manipulations performed within a defined perceptual space will result in improvements in subjective picture quality. In pictures such as those processed by Faugeras, scene variables are already represented by perceptually significant attributes since they are real-world pictures. In our case, we may wish to create artificial scenes from arbitrary data sets, and to realise perceptual attributes as closely as possible, arguing that this is fundamental to the image comprehension process.

This suggests the use of an empirically derived UCS, rather than a space based on a particular visual system model, as central to the display framework. We can formally define perceptual variables H, S and L in their natural alignments in such a UCS in terms of their rectangular co-ordinates $(U1, U2, U3)$. That is,

$$\text{Hue}\,(H) = \tan^{-1}(U3/U2)\ \text{modulo}\ 2\pi;$$

$$\text{Saturation}\,(S) = \left[U2^2 + U3^2\right]^{\frac{1}{2}}\ ;$$

$$\text{Lightness}\,(L) = U1\ .$$

This defines the transform $\Phi_{\mathrm{PU},\lambda}$ from UCS (**U**) to perceptual colour space (**P**); its inverse, $\Phi_{\mathrm{UP},\lambda}$, is also hence defined.

The CIE spaces are attractive on grounds of analytical tractability; UCS co-ordinates are defined in terms of CIE tristimulus values, allowing conversion in either direction. This is of great advantage in a computational system, the alternative being three-dimensional look-up tables of measured values, with interpolation necessary to gain sufficient resolution without inordinately large look-up tables. While numerical approximations relating the Munsell System co-ordinates to tristimulus values have been developed, evaluation is computationally expensive.

The other advantage the CIE spaces hold is that they have been widely used and studied since their inception (see, for example, Judd and Wyszecki, 1975; Ohta, 1977; Stenius, 1978; Friele, 1979; McLaren, 1980; Pointer, 1980,1981). This means that anomalies, or areas of poor uniformity, have been discovered and documented; such anomalies can then be avoided or accounted for if their effects are considered sufficiently detrimental.

On these grounds we chose the CIE recommended CIELAB and CIELUV uniform colour spaces (UCS) as most suitable for the proposed framework. This does not preclude the use of other uniform spaces which can be analytically specified; in fact the following work is developed around a generalised UCS (**U**), and the implementation is designed to allow a single parameter to determine the specific UCS used. This means that any other three-dimensional (Euclidean) approximation to a perceptually uniform space can be used simply by providing appropriate UCS-to-tristimulus-value, and tristimulus-value-to-UCS, calculating modules.

We now consider access to **U** at the various levels appropriate for performing spatial or spectral processing significant in some sense to the operation of the visual system.

3.3 Spatial and spectral processing within the framework

In this section we consider the type of spatial or spectral processing which can be performed, on the basis of some physically or perceptually based rationale, within the framework. We first treat processes which result from a representation chosen to reinforce high level scene comprehension, and involve a spectral modification or operation which depends on the spatial and spectral properties of the data, and the chosen representation. Secondly we consider processes which stem from the low-level mechanistic properties of the visual system, and suggest some spatial or spectral modification of a displayed image which depends on the spatial nature of the data, and on its spectral assignment.

3.3.1 Transformations dependent on spatial representation

An image, or picture, is a spatially two-dimensional representation of information. Such an image, however, can incorporate information pertaining to a third spatial dimension which can be depicted spatially (by displaying in perspective, for example), or spectrally (by shading, for example), or both spatially and spectrally. In pictures of real-world surfaces, these spatial and spectral depictions are generally implicit, but in a synthesised image of arbitrary data, we may wish to introduce them in order to utilise depiction of an implied third spatial dimension. Thus we perform a data-dependent transformation $\Phi_{\mathbf{DP}}$ from data space **D** to the spatially three-dimensional perceptual space **P**. Within **P**, we must allow for a spectral transform $\Phi_{\mathbf{PP},\lambda}$ resulting from the spectral modification which might be involved in depicting this third spatial dimension. This transform will be dependent on the spatial geometry of the chosen representation. In addition, we must allow for a spatial transform $\Phi_{\mathbf{PP},x}$ which might result from a projection of a spatially three-dimensional perceptual space (in which a realistic image is conceptually perceived - see Barnard (1985)) to a spatially two-dimensional space (which is the actual image space), and a change from perceptual to rectangular UCS spectral co-ordinates $\Phi_{\mathbf{PU},\lambda}$. Section 3.1 defines these transforms.

3.3.2 Transformations dependent on visual system properties

We may wish to incorporate a model of a visual effect such as (photopic, as distinct from spatial frequency) induction, or local adaptation. Induction describes how the perceived spectral nature of a point is affected by the spectral characteristics of its spatially neighbouring areas. Induction models based on perceptual level processes have been proposed (Troscianko, 1977; Oyama et al., 1980); such models involve a spectrally- and spatially-dependent spectral transform $\Phi_{\mathbf{PP},\lambda}$ in the perceptual space **P**. Models which are based on lower level processes have also been proposed (Ware and Cowan, 1982; Takahashi and Ejima, 1983); such models involve a transform $\Phi_{\mathbf{MM},\lambda}$ in the low-level (mechanistic) space **M**. In addition, while a colour space defined relative to the reference illumination to some extent compensates for chromatic adaptation effects, not surprisingly, perceived colours do change with changes in adaptation. Bartleson (1978,1979a,1979b) and Richter (1980) report on the influence of chromatic adaptation on the CIE colour spaces; compensation for such effects involves a transform $\Phi_{\mathbf{UU},\lambda}$, being a spectrally- and spatially-dependent modification of spectral co-ordinates. Wright (1981) summarises studies of chromatic adaptation; more recently a non-linear mechanistic model has been proposed by Nayatani et al. (1981), while models relating to Stiles' π mechanisms (receptor-specific gain controls) have been investigated widely (see, for example, Walraven (1980), a summary by Walraven and Werner (1982), and Werner and Walraven (1982)). Compensation using such models is performed in a low level domain, and as before involves a transform of the form $\Phi_{MM,\lambda}$. These studies suggest the need to be able to transform from the perceptual space to lower visual levels.

Secondly, investigations into the spatial frequency characteristics of the visual system have produced evidence that the achromatic and chromatic mechanisms have different spatial-frequency response characteristics (see Chapter 7). Various attempts have been made to explain these characteristics using visual models, and image processing based on these models has been performed with encouraging results (Faugeras, 1976; Hall and Andrews, 1978). In light of these results, it seems desirable that the display framework should encompass transforms of the form Φ_{MM}, which allow consideration, and modification, of spatial frequency content (in the frequency domain), based on certain spectral and spatial conditions. This means that it must be possible to transform to some appropriate level (M) to apply filters as required in the frequency domain, or in some approximation to it. It should be noted that parameters relevant to the display system (such as display size and resolution, viewing distance, etc.) also affect perceived spatial frequency, and could consequently affect the choice of spatial frequency processing.

The studies described in this section strongly suggest the need for transforming between UCS and a lower visual level (M); this involves a reversible transformation $\Phi_{UM,\lambda}$, and is treated more fully in Chapter 7.

3.4 Realisation of the framework

3.4.1 Realising specified colours on display devices - device modelling

Effective use of perceptual attributes defined within a UCS requires that it be possible to produce colours specified in UCS co-ordinates on any chosen display device. This not only demands a certain level of device consistency, but also a method of transforming between UCS co-ordinates and display device gun-counts, and can be approached in one of two basic ways.

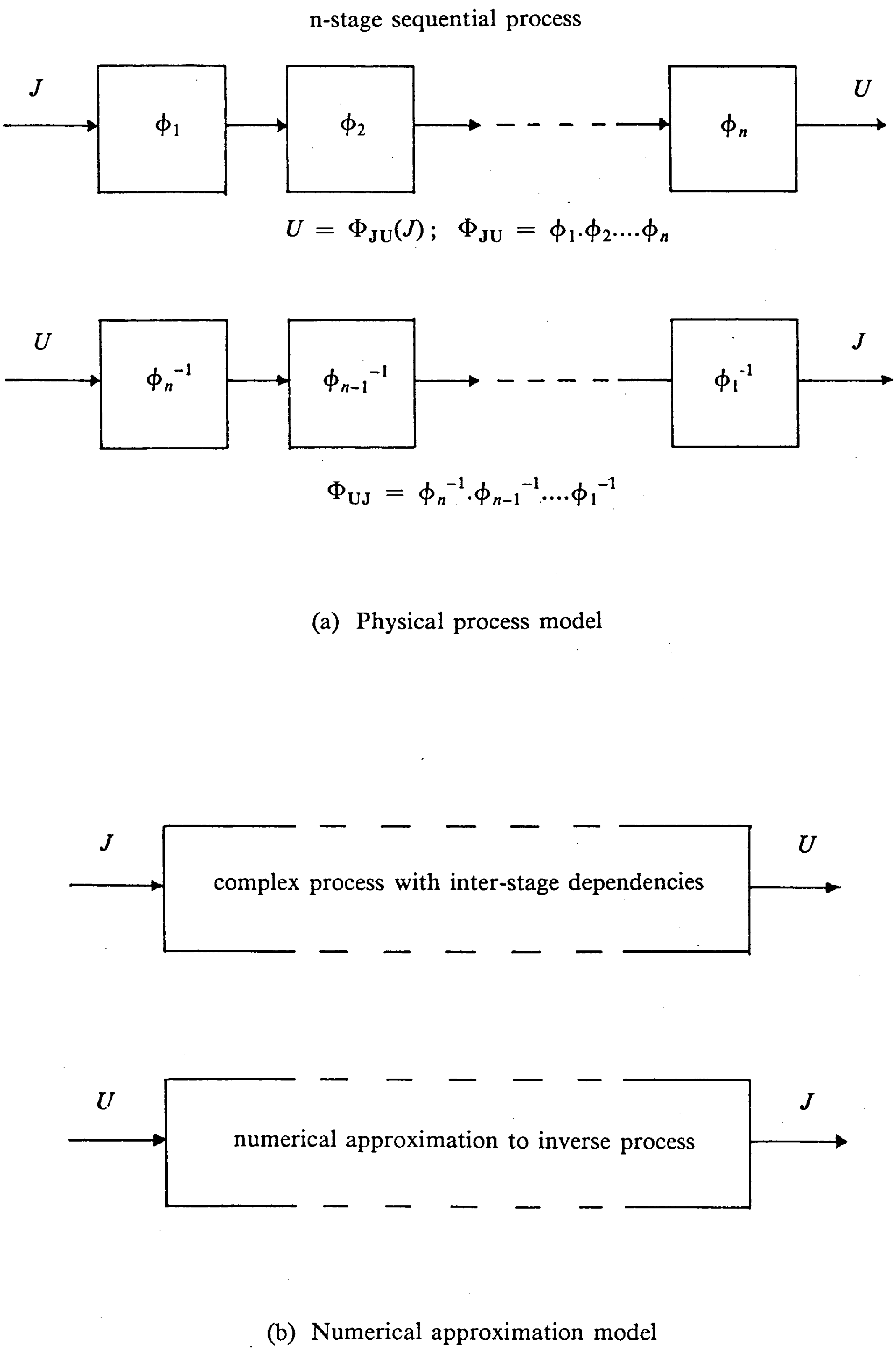
The first is to model the physical processes involved in the operation of the device; this approach is attractive from a standpoint of rigour, and for the advantage that sensitivity to variations in produced colour might well be parametrisable. It requires that it be possible to describe the signal path at every stage, together with any inter-stage dependencies such as feedback. This is shown schematically in figure 3.4(a). For an n-stage sequential process, each stage of which is described by a function ϕ_i, $i = 1,\ldots,n$, we have output UCS co-ordinates produced by input gun-counts given by

$$U = \Phi_{\mathrm{JU}}(J) \,; \quad \Phi_{\mathrm{JU}} = \phi_1.\phi_2.\ldots.\phi_n.$$

If each stage of the process is invertible, we can derive the required transform Φ_{UJ} from

$$\Phi_{\mathrm{UJ}} = \phi_n^{-1}.\phi_{n-1}^{-1}.\ldots.\phi_1^{-1}.$$

This process can be extremely complex to implement, however, for display processes which rely on latter stage compensation for early stage inconsistencies. This occurs in a photographic process where the colour balance in the final print is adjusted to compensate for film and printing paper batch characteristics. Photographic transparencies pose a slightly less complex modelling problem; Wallis (1975) treats this in detail (summarised in Pratt, 1978). Juday (1979) develops a simplified "projection-addition" model to find the gamut of a transparency photographic film product; in this model primary axis sample points are used for calibration, and spectral characteristics elsewhere in the gamut are derived using an assumption of linearity between tristimulus values and primary axis


n-stage sequential process
J
ϕ_1
ϕ_2
ϕ_n
U
$U = \Phi_{JU}(J)$; $\Phi_{JU} = \phi_1.\phi_2....\phi_n$
U
ϕ_n^{-1}
ϕ_{n-1}^{-1}
ϕ_1^{-1}
J
$\Phi_{UJ} = \phi_n^{-1}.\phi_{n-1}^{-1}....\phi_1^{-1}$
J
complex process with inter-stage dependencies
U
U
numerical approximation to inverse process
J

(a) Physical process model

(b) Numerical approximation model

Figure 3.4 Display device modelling processes

excitations. The extent to which this assumption is justified is not clear; Juday points out that it neverthless offers a substantial step towards device modelling. McDonnell (1980; McDonnell and Fowler, 1981) projects a display device cube into a parallelepiped in CIELUV space, tying the projection by the apex points, to develop a calibration for a transparency film. This approach also relies on the assumption of linear additivity, both in the use of the CIELUV space, and in the derivation of values between apex points. It is presented as a calibration technique, amounting to a more sophisticated form of density linearisation which takes visual and film characteristics into account, rather than as a means of realising an approximation to a UCS.

For display devices which are less complex, or not subject to temporal inconsistencies, physical modelling is more feasible. A colour television monitor is a device in which the processes affecting the final signal can be individually and stably characterised, and for which the linear additivity assumption is valid if the monitor is properly adjusted; consequently it can be modelled using this approach.

A second approach to device modelling also treated by Wallis (1975), and by Juday (1979), is to develop a numerical model of the entire display process, simply relating the input gun-count co-ordinates to the output tristimulus values (and hence UCS co-ordinates). This process is shown schematically in figure 3.4(b). It has the drawbacks common to numerical modelling; first, that the statistical analysis used to derive the model parameters is not necessarily performed on appropriately significant grounds, and second, that artifacts of the model can be difficult to avoid; but it does allow a process, which for some reason cannot be parametrised, to be modelled. In the following sections we describe implementation of a model of each of the above types and compare their results.

3.4.2 Modelling a colour television monitor

Meyer and Greenberg (1980) suggest a method for modelling a colour television monitor which develops a non-linear transform based on measuring the phosphor chromaticities and individual gun-voltage/excitation relationships. This approach relies on the assumption that the monitor is properly converged, and consequently that cross-talk between guns and other primaries is negligible. Under such a masking assumption, signal combination becomes additive, and a linear transform between device primary excitations and tristimulus values results. (This also requires that the shape of the spectral response to excitation be independent of excitation level; for the phosphors used in colour television monitors this is largely true (Hunt, 1975).) The voltage/excitation curve for each gun can be measured and approximated by a "gamma" curve (Neal, 1973), and the transform is hence defined. This process has been re-stated by Cowan (1983) who comments on practical methods of determing phosphor chromaticities and the "gamma" relationship. Cowan points out that a simple "gamma" relationship can be a poor approximation to fitting a monitor voltage/excitation curve, and suggests using a two-parameter curve fit, emphasising the need to perform the curve measurement to derive an appropriate form for the approximation. It should be noted that in some colour monitors, a built-in brightness level compensation circuit can render the model invalid when the average overall screen brightness is low. This means that model validity can be restricted; some form of surround brightness compensation could be applied to restore partial validity, but unfortunately such a compensation might well also have ill-defined perceptual effects.

The signal flow through each gun of a colour television monitor can be represented by

$$C_l = \phi_1(C_v) = \phi_1(\phi_2(C_j)) .$$

Subscript l denotes screen luminance, subscript v denotes applied tube voltage, and subscript j denotes input gun-counts. If $C_{l_{norm}} = C_l/C_l max$, then $C_{l_{norm}}$ can be considered as a normalised tube primary, subject to the above-mentioned conditions of zero cross-talk and linear additivity.

Thus for each gun

$$C_l = C_v^{\gamma_C},$$

or, as suggested by Cowan, for a more accurate fit,

$$ln(C_l) = a\,ln(C_v) + b\{ln(C_v)^2\}.$$

The value of γ, or those of a and b, can be found by best-fitting the measured excitation/luminance curve for each gun. The non-linear transform ϕ_1 from applied tube voltage to tube screen luminance is then specified.

If the relationship between input gun counts and applied tube voltage is linear, which is the case for a straightforward digital-to-analog converter, then ϕ_2 is given by a direct linear relationship, with appropriate range normalisation. That is,

$$C_v = k\,C_{jnorm}, \text{ where } C_{jnorm} = C_j/C_{jmax}$$

for some constant k.

Since ϕ_1 and ϕ_2 are each invertible, the overall transform ϕ_{PJ} from each normalised tube primary C_{lnorm} to each applied gun count C_{jnorm} is then defined. For all three guns ($C = R,G,B$), we term the transform Φ_{PJ}.

Following Neal (1973), if red, green and blue phosphor chromaticities are (x_r,y_r,z_r), (x_g,y_g,z_g) and (x_b,y_b,z_b) respectively, tristimulus values are determined by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} X_r & X_g & X_b \\ Y_r & Y_g & Y_b \\ Z_r & Z_g & Z_b \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} x_r & x_g & x_b \\ y_r & y_g & y_b \\ z_r & z_g & z_b \end{bmatrix} \begin{bmatrix} C_r & 0 & 0 \\ 0 & C_g & 0 \\ 0 & 0 & C_b \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

$$\text{where } x_r = \frac{X_r}{X_r + Y_r + Z_r} = \frac{X_r}{C_r}, \text{ etc.}$$

For the reference white point (set to standard illuminant D6500), we have known tristimulus values, unity primary excitation values, and known white point chromaticities. Hence C_r, C_g and C_b can be found by substituting for the white point values.

Thus the matrix Φ_{PT} transforming from normalised monitor primaries to CIE tristimulus values is specified, and can be inverted to allow the specification of the transform Φ_{UP} from UCS co-ordinates to normalised monitor primaries:

$$\Phi_{UP} = [\Phi_{PT}]^{-1}.\Phi_{UT},$$

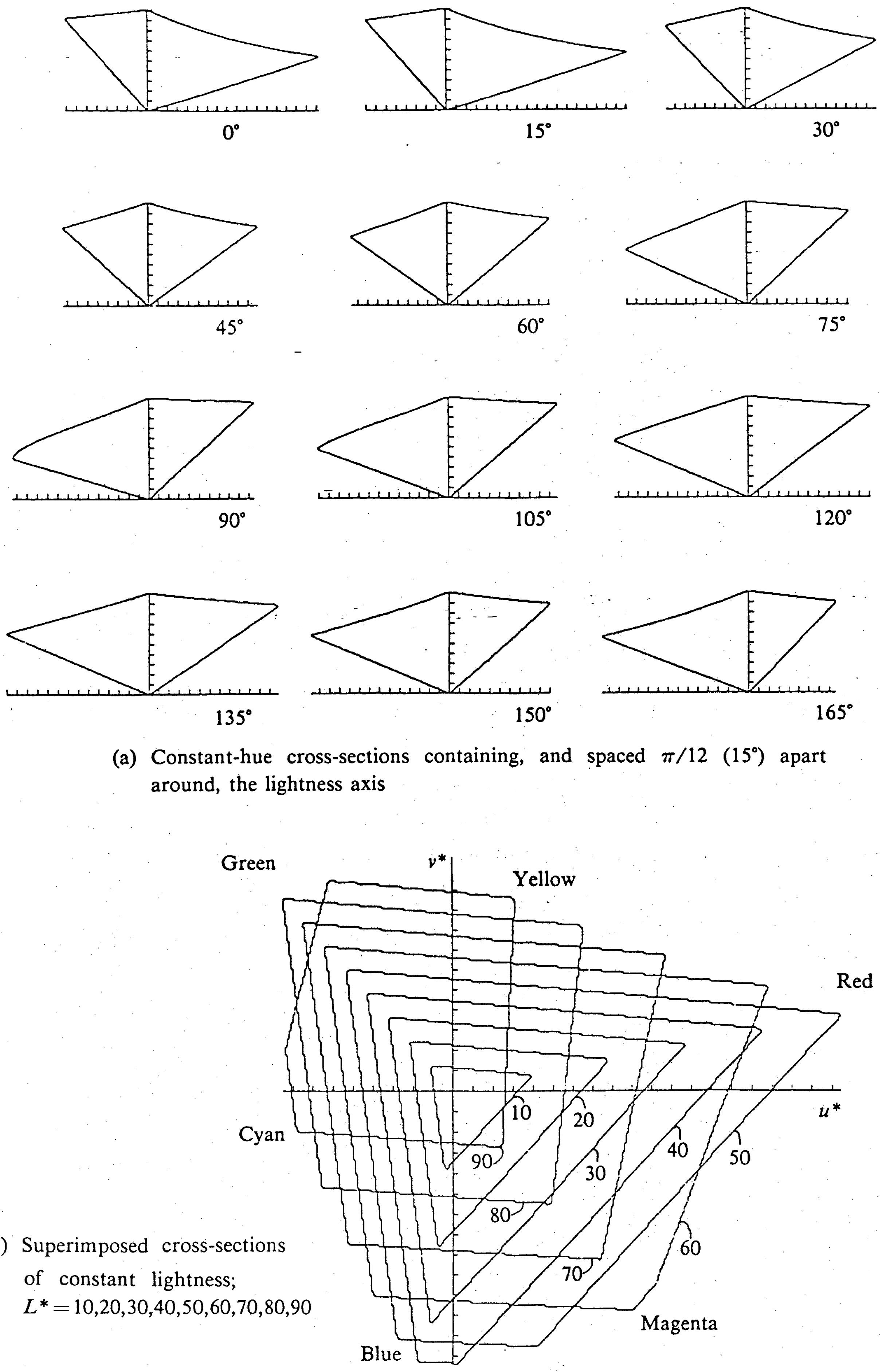
where Φ_{UT} is the non-linear transform from UCS co-ordinates to tristimulus values for the chosen UCS (see Appendix 2).

The required overall monitor model transform Φ_{UJ} is then given by

$$\Phi_{UJ} = \Phi_{PJ}.\Phi_{UP}.$$

This form of model was implemented for a colour television monitor using manufacturers' published chromaticities and measured tube "gammas". Figure 3.5 shows in graphical cross-sectional form the CIELUV gamut of the monitor modelled by this method. In 3.5(a), cross-sections including the lightness axis and spaced $\pi/12$ (15°) apart around this axis, are shown; each cross-section is of constant (twin) hue, containing a particular hue and its grey-axis complementary. In 3.5(b), a set of superimposed cross-sections of constant lightness is shown. The approximate locations of commonly recognised device-primary colours (red, green and blue) and device-complementary colours (cyan, magenta and yellow) are marked. Figure 3.6(a) shows the gamut cross-section through the three device-primaries, while 3.6(b) shows the cross-section through the three device-complementaries.

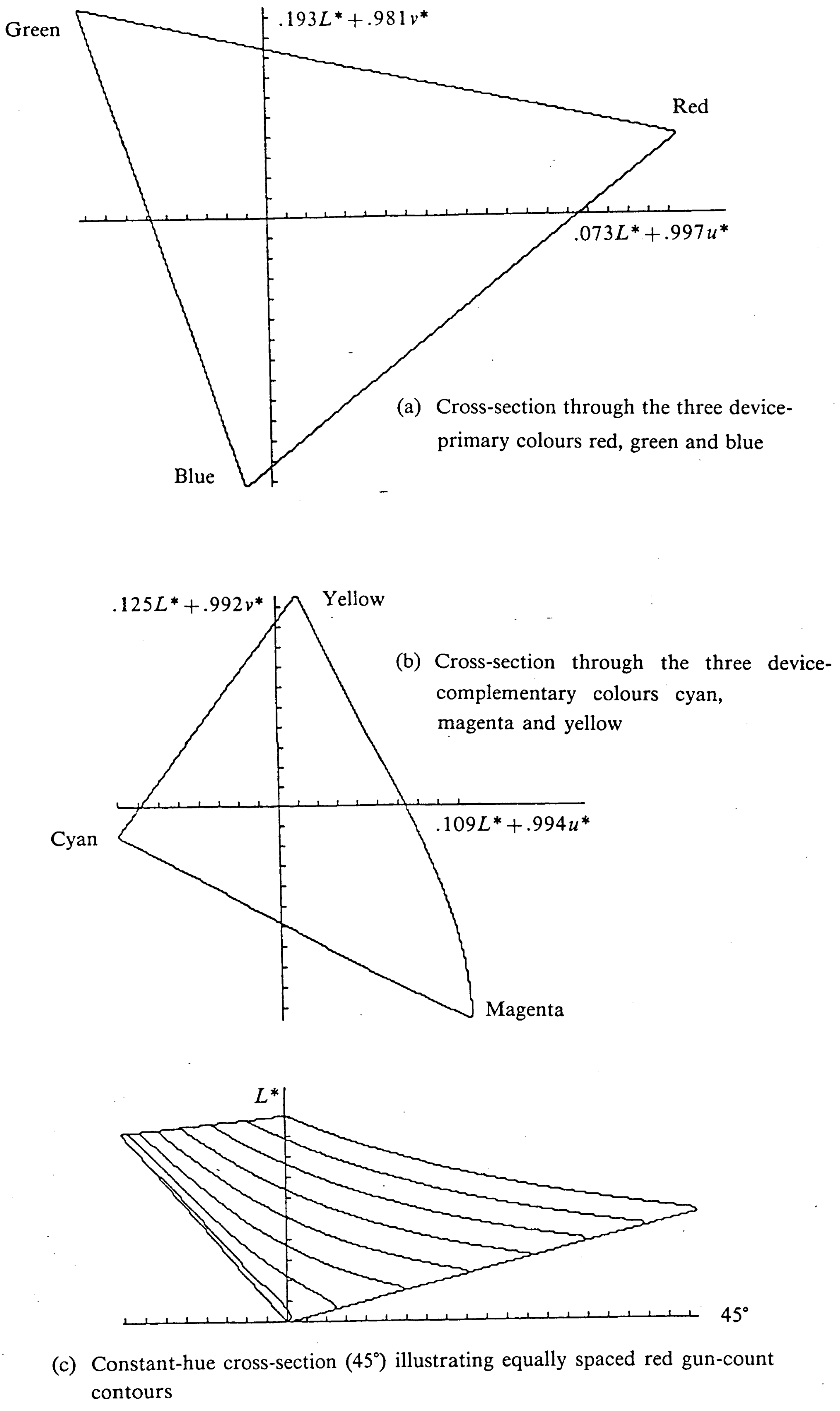
We can make several observations, pertinent to the use of these gamut shapes in following chapters, about interpretation of the model-depicted gamuts. First, the gamut boundaries in cross-sections orthogonal to the lightness axis are straight lines, while those in the cross-sections including the lightness axis are non-linear functions. In fact any gamut boundary line which is not of constant lightness will not be straight. Consequently a straight line joining gamut points of constant lightness will not leave the gamut, while a line joining points of different lightness may do so (for example, a straight line joining the red primary to the white point does not lie within the gamut). Second, the sensitivity



(a) Constant-hue cross-sections containing, and spaced $\pi/12$ (15°) apart around, the lightness axis

(b) Superimposed cross-sections of constant lightness; $L^* = 10,20,30,40,50,60,70,80,90$

Figure 3.5 Outlines of constant-hue and constant-lightness cross-sections through the colour television monitor UCS gamut. Scale mark intervals are 10 JND.



(a) Cross-section through the three device-primary colours red, green and blue

(b) Cross-section through the three device-complementary colours cyan, magenta and yellow

(c) Constant-hue cross-section (45°) illustrating equally spaced red gun-count contours

Figure 3.6 Outlines of cross-sections through the colour television monitor UCS gamut.

of the gamut to model parameters can suggest relative, rather than absolute, interpretation of the shapes. Cowan comments on the importance of deriving an accurate tube gamma-correction function, pointing out that miscalculating it results in correct chromaticities at the tube relative-level normalising point (in our case at the reference white), but incorrect chromaticities elsewhere. The monitor model developed was used in a subjective manner only; if a further reproduction stage such as taking colour transparencies from the screen were made, accurate tube "gamma" correction would be even more important. Finally, it can be instructive to look at the relationship between perceived colour differences (the metric of the UCS) and numerical gun values. This can be depicted by contouring at equal gun-count intervals in the UCS gamut shapes for each of the guns individually. Figure 3.6(c) shows the cross-section containing the red primary so contoured for that primary; similar trends are evident for the other primaries (with an appropriate hue shift).

Monitor phosphor chromaticities change with age (Hunt, 1975; Cowan, 1983), and consequently should be measured for any particular monitor being modelled. The effect of errors in chromaticity specification on the magnitude of colour differences will not be large, but alignment with the UCS chromatic axes will change.

3.4.3 Numerically modelling a photographic film recording process

We require a numerical model which can be used to calculate the gun-count values necessary to produce any colour specified in UCS co-ordinates. This involves generating a set of sample colours, the gun-count values of which are known, and measuring their UCS colour co-ordinates. This measurement can be made by visual comparison, or more accurately by using a tricolorimeter or spectrophotometer (see Appendix 1). We thus have a set of samples for which both gun-count values and UCS co-ordinates are known (UCS co-ordinates are uniquely specified by tristimulus values). We wish to find a functional approximation to the relationship between gun-count values and UCS co-ordinates of the form

$$J = \Phi_{\mathrm{UJ}}(U).$$

To simplify the numerical model, we can treat each gun separately, and find a set of functions ϕ_i ($i=1,2,3$) such that

$$J_i = \phi_i(U).$$

A model of this type was derived for a photographic process consisting of exposure of a negative (Kodak Vericolor II) through a set of three filters on a high resolution Optronics Colorwrite film recorder, negative development, and then photographic enlargement on standard Kodak colour printing paper. Processing and enlarging was performed by a commercial laboratory using standard processing chemicals and techniques.

In the model implemented a third order linear regression was used, each gun-count variable being expressed as a third order polynomial in UCS values. The polynomial fit was optimised using a minimum mean-square-error criterion. A second order model failed to explain the data variance sufficiently well, while a fourth order model resulted in residuals only marginally lower than with the third order model; this marginal improvement did not justify the risk of introducing anomalies due to the higher order terms.

In order to check the model derived, an approach to finding Φ_{UJ} which does not involve the simplifying step of treating each gun separately was also implemented. This used an entirely different numerical modelling technique, involving generating a three-dimensional spline fit to the data based on a method of minimising a generalised cross-validation parameter (Hutchinson, 1984). This approach has the advantage that at each knot, or specified data point, a simple function of low intrinsic curvature is evaluated. This substantially reduces the chance of introducing singularities in areas of low data point density. However, it has the disadvantage that in conversion calculations, the function must be evaluated at each of a representative set of points; this can be computationally very expensive if a large number of such points is required. For the three-dimensional colour volume, a fall off in closeness of spline fit occurred for less than about 60 knots, resulting in point calculations which took an order of magnitude longer than did polynomial evaluation. Consequently this technique was used only to substantiate the model derived using polynomial linear regression. In fact, if the data samples are fairly regularly spaced in colour space, it is unlikely that polynomials will cause perceptual anomalies; for a model based on irregularly spaced data, a spline fit model might be more appropriate.

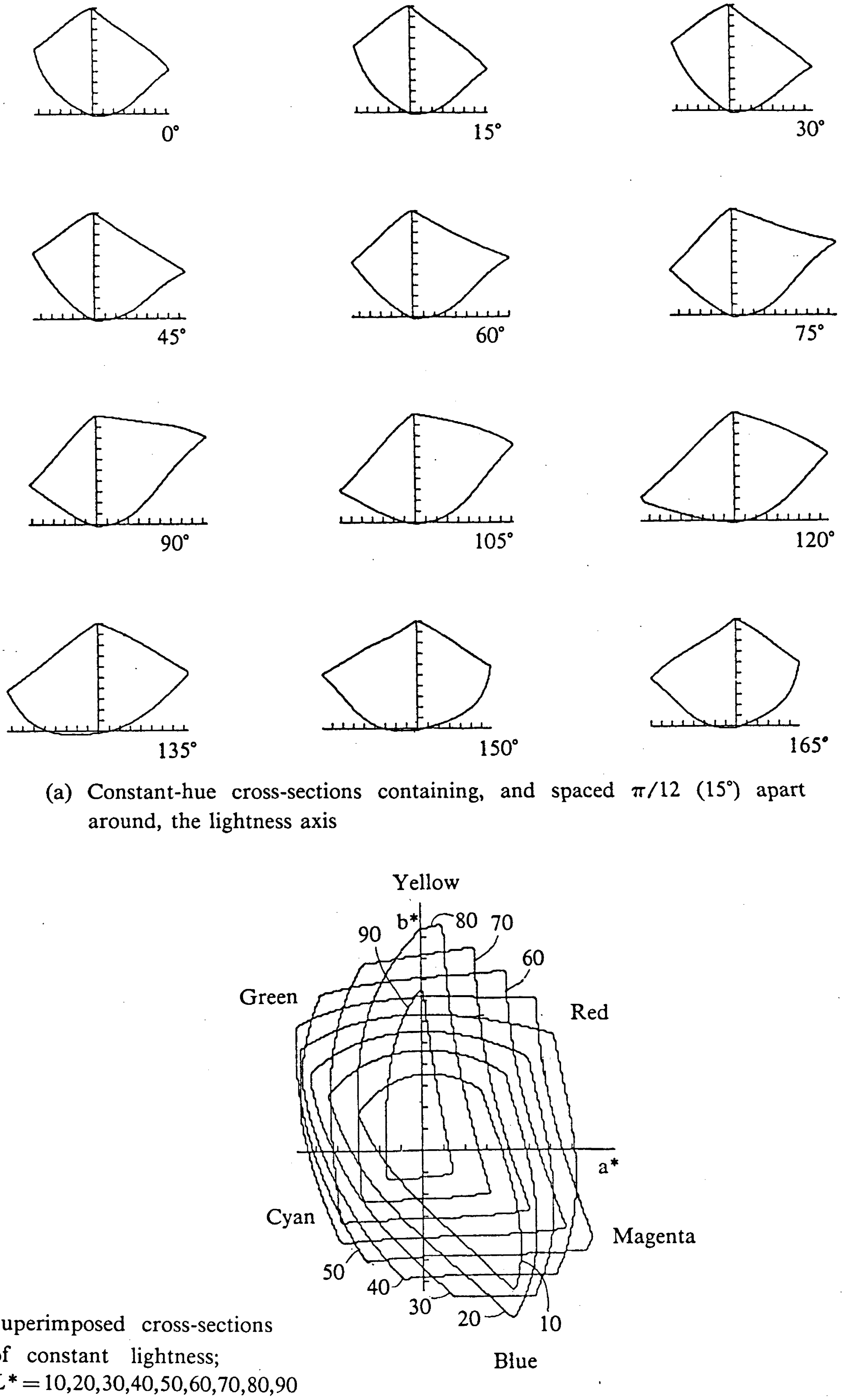
Evaluation of the regression model polynomials depends on the number of terms in the polynomial, and not on the number of data points used to generate the model. Consequently it was feasible to use a large sample data set. Initially a model based on visual comparison of samples was derived; the UCS so generated was based on about 100 samples. These samples were not regularly spaced, and there were areas of colour space sparsely covered. In order to improve the model, this UCS was then used as a basis for generating three sets of colour chips. Each set consisted of about 200 samples, evenly spaced in a lattice in the first-developed UCS, recorded on separate film sheets, and printed on separate paper sheets (but with a common enlarger filtration). This was so that models for each set, and the combined set, could be derived and compared to give some indication of process and model sensitivity. The tristimulus values for each sample were measured using a Hunterlab D25 tricolorimeter (accurate to 0.5 JND).

In fact it was necessary to apply some compensation to the measured chip tristimulus values because of uneven enlarger illumination over the extent of each print. Compensation factors were derived from edge (unexposed film) tristimulus values; proportional interpolation between edge values was performed for compensation of non-edge-adjacent samples. As it turned out, the models derived from the individual data sets gave barely distinguishable gamut profiles, and in the model finally derived, the complete set of samples was used.

In deriving the final model, several refinements were made. First, the fit optimisation process considers residuals (being the error in predicting each gun-count value of the sample set) in gun-count terms, and considers the significance of available polynomial terms on the basis of the variance of gun-count values explained. In fact residuals in UCS coordinates are a more appropriate measure of the significance of the terms, and by re-scaling input gun-counts to density linearised (along the grey axis) values, an improved model results. Note that for the reasons stated in section 3.2.3, this does not result in UCS scaling throughout the space, but is nevertheless a first approximation improvement. Second, as suggested by Draper and Smith (1964), it is appropriate to approximately normalise the polynomial coefficients by scaling the high order terms in order that accuracy is consistent; otherwise high order terms tend to have co-efficients very much less than unity, resulting in a loss of precision if evaluated in a short-word-length computing mode. Third, due to the perceptual significance of the grey axis in colour space (hues very slightly different from grey are easily recognised as such), it is crucial that the grey axis be accurately portrayed. This perceptual importance can be numerically incorporated by extra weighting of sample points on or near to the grey axis, possibly in a saturation-dependent fashion. In the developed model this form of weighting was used to remove a slight colour cast from the grey axis. The resulting grey axis can be seen in quantised form in the grey wedge on the colour prints in this thesis.

Graphical cross-sections through the CIELAB UCS gamut of the final model are shown in figure 3.7. In 3.7(a), cross-sections including the lightness axis, spaced $\pi/12$ (15°) apart in hue, are shown. In 3.7(b) a set of superimposed cross-sections of constant lightness is shown. Approximate device-primary and device-complementary colour locations are marked. Figure 3.8(a) shows graphically the gamut cross-section through the device-primaries, while 3.8(b) shows the cross-section through the device-complementaries.

The non-coincidence of the model-derived gamut zero level with the UCS lightness zero level is a consequence of the fact that for the photographic process modelled, at very low excitation of all three guns, no apparent change in lightness in the print occurs. However, when at least one gun is more highly excited, very low increments in another gun just turned on do have an observable effect; consequently very low gun-counts must still be considered. A plausible explanation for this difference in sensitivity to very low gun-counts is that there is sufficient cross-talk for a highly driven gun to raise another primary above the effective base fog level; when no gun is driven hard, however, the



(a) Constant-hue cross-sections containing, and spaced $\pi/12$ (15°) apart around, the lightness axis

(b) Superimposed cross-sections of constant lightness; $L^* = 10,20,30,40,50,60,70,80,90$

Figure 3.7 Outlines of constant-hue and constant-lightness cross-sections in the colour film/print UCS gamut. Scale mark intervals are 10 JND.

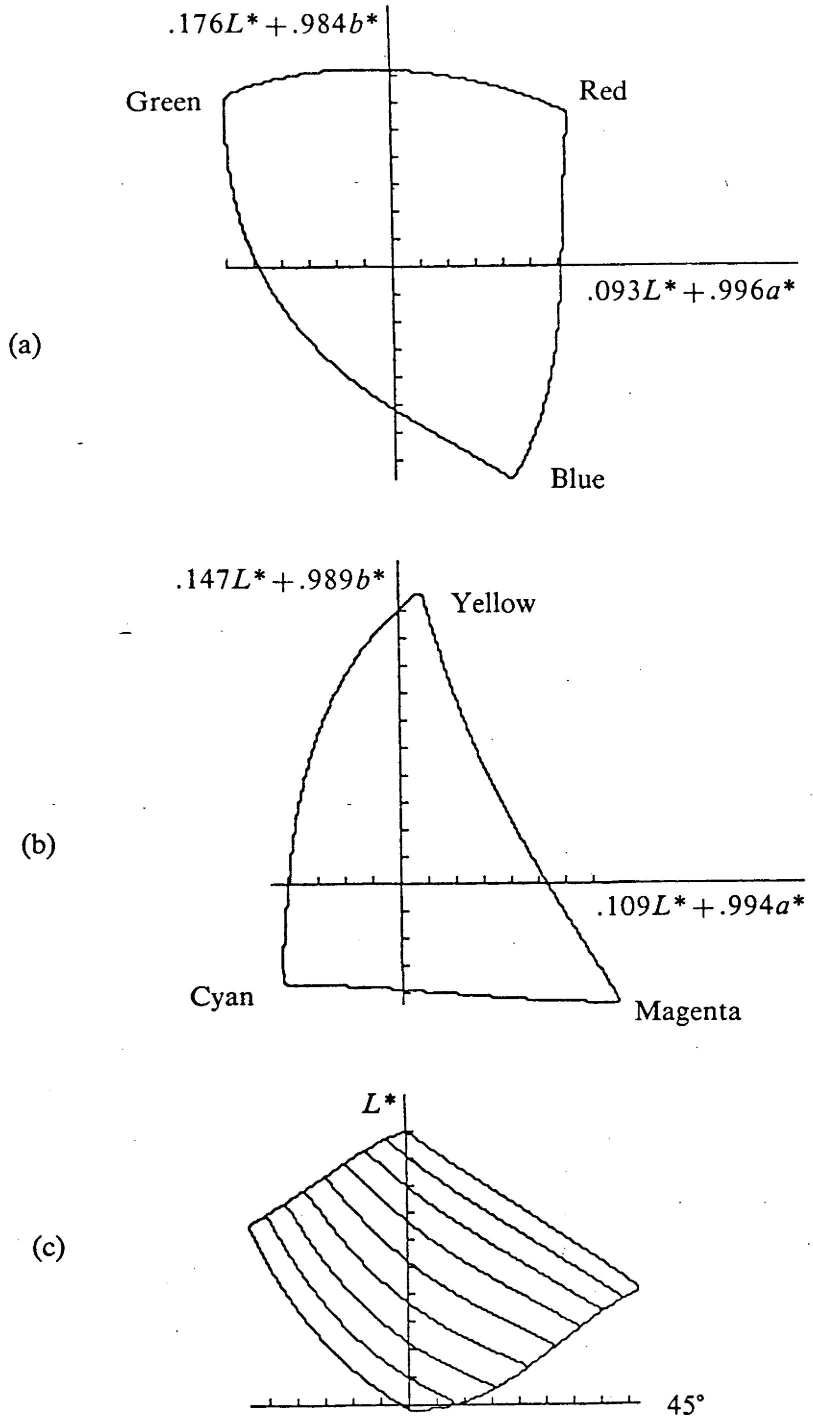


Figure 3.8 Outlines of cross-sections through the colour film/print UCS gamut. Scale mark intervals are 10 JND.

(a) Cross-section through the three device-primary colours red, green and blue

(b) Cross-section through the three device-complementary colours cyan, magenta and yellow

(c) Constant-hue cross-section (45°) illustrating equally spaced red gun-count contours stimulus at low gun-count values is still below the effective base fog level. A more careful investigation of this effect would be required to establish the validity of this explanation. Numerically, the model attempts to compress a neighbourhood in the region of the gun-count origin into the origin in UCS. With the limitations in the order of the polynomial terms, this compression is only achieved approximately, and the resulting intrusion of the gamut origin below the UCS origin is hence an artifact of the modelling process. This small intrusion was not considered to be sufficiently detrimental either to subjective graphical appreciation of the gamut shape, or to computational evaluation processes (the UCS origin lies at a gun-count value of 4 in a total range of 255, for each of the three guns), to warrant changing the model order.

The intrusion in the blue region of apparently zero lightness is a recognised anomaly of the CIELAB UCS (see Judd and Wyszecki, 1975 p.332). In practice it makes little difference when mapping data into UCS, as this region of UCS is seldom used; some care only with distribution tails extending into this region is required. As for the colour monitor gamut, cross-sections contoured in gun-count values can provide a useful gauge of the gun-count ranges which cover the bulk of the UCS volume, reinforcing the earlier-made statements about the inappropriateness of using gun-count values as a data-representing metric. Figure 3.8(c) shows a constant-hue cross-section including the red primary contoured for the red gun-count; other guns again give similar (hue shifted) results.

Figure 3.9(a) shows a set of colour cross-sections including the lightness axis, spaced $\pi/12$ (15°) apart around this axis, for the colour film/print process modelled. The angle is measured from the positive a^* axis, towards the positive b^* axis; in the figure, angle increments from left to right and downwards as in figure 3.7(a). The degree of hue-constancy achieved in the device-modelled UCS can be seen in these cross-sections. Also shown, in figure 3.9(b), is a set of constant lightness cross-sections for lightness values of 20 (upper left), 40 (upper right), 60 (lower left) and 80 (lower right). Figure 3.9(c) shows the cross-section which includes all three device-primaries, and that which includes all three device-complementaries. In this figure, as in all other colour figures in this work, a grey wedge evenly spaced in UCS terms along the lightness axis is given as a printer's

Image descriptions for the following colour plate

Figure 3.9 Cross-sections through the colour film/print UCS gamut

(a) Constant-hue cross-sections containing, and spaced $\pi/12$ apart around, the lightness axis. Hue angles, measured from the positive a^* axis to the positive b^* axis, increment from left to right and downwards, as in figure 3.7(a).

(b) Cross-sections of constant lightness; $L^*=20$ (top left), 40 (top right), 60 (bottom left), 80 (bottom right).

(c) Cross-sections through the three device-primary colours red, green and blue (top), and the three device-complementary colours cyan, magenta and yellow (bottom).

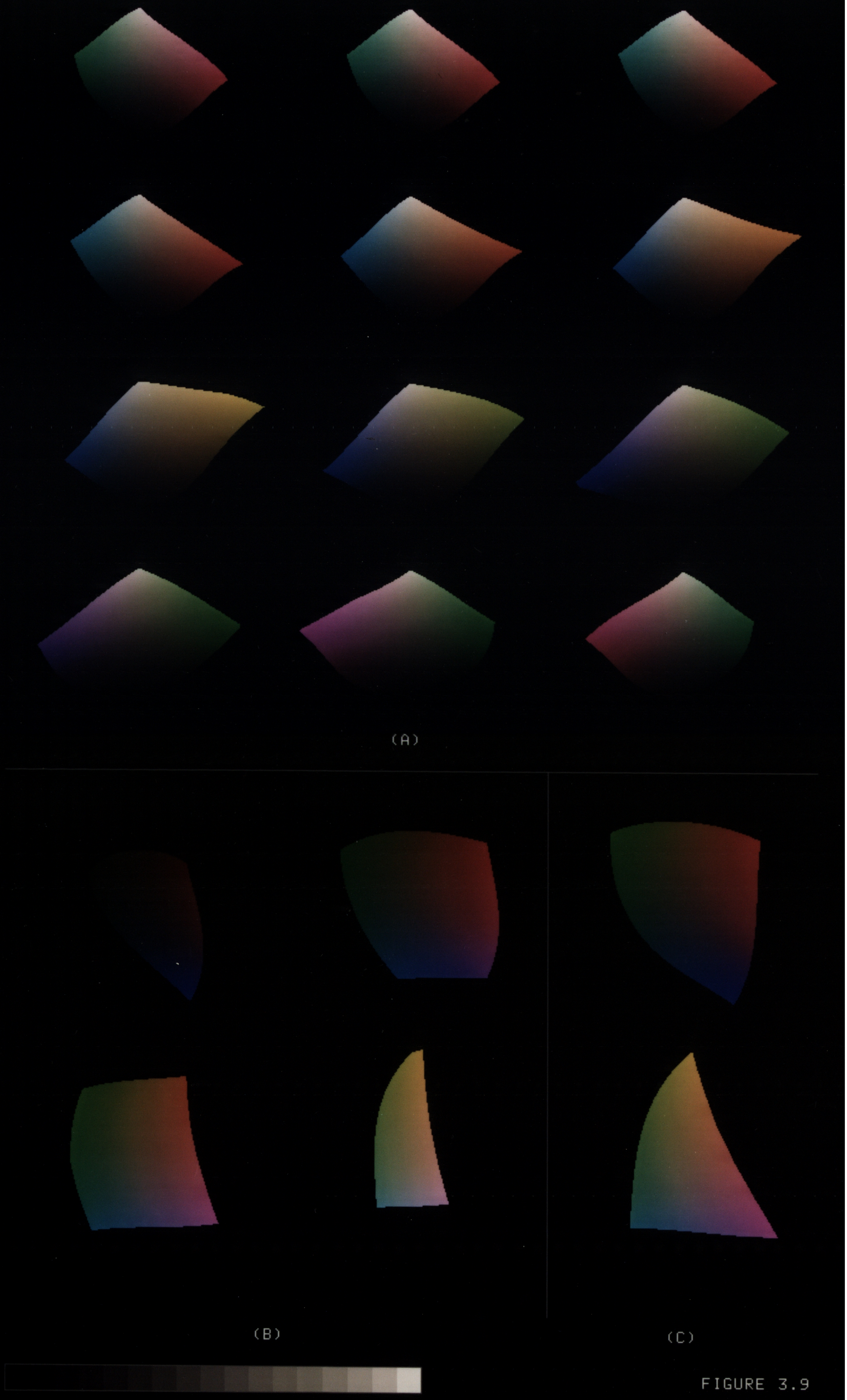


FIGURE 3.9

guide, and subsequently to act as an visual indicator of introduced casts or variations. Because of the perceptual sensitivity to casts in a neutral grey, this forms a reasonable indication of product control.

3.4.4 Comparison between CIELUV colour monitor and CIELAB film/print gamuts

Figure 3.10 shows a set of superimposed constant-hue gamut cross-section outlines for the colour film/print and colour television monitor UCS gamuts. We can compare the shapes of these gamuts, subject to the caveat that we assume the appropriateness of the CIE recommended spaces for each device type. Investigations comparing these spaces have been made (Ohta, 1977; Pointer, 1980,1981), and it transpires from these studies that the spaces do in fact have subjective dissimilarities. However, some features, borne out by experience in practice, can be of value when considering their application to display.

First, the general form of the colour monitor gamut is sharply defined, reflecting the lack of cross-talk in that type of device. The photographic process gamut, however, is well rounded, indicating cross-talk effects as being significant. Second, the monitor gamut has its widest extent at high lightness values, in keeping with the characteristics of phosphorescence with incident energy. Device-complementary colours have a higher nett energy, and hence lightness, than their two constituent primaries. In the photographic process gamut this need not be the case. In a subtractive process, in general activating two sets of colour dyes results in increased absorption and hence reduced nett reflectance. However, since a two-stage process is involved (negative exposure, followed by print exposure), the effects are more complicated, but loosely speaking tend to cancel and have a similar overall result as for the colour monitor. These factors can be relevant when choosing colour sequences to represent the magnitude of a displayed variable, as described in the following chapter.

As indicated by Pointer (1981), the CIELUV and CIELAB spaces are not consistent in their co-ordinate representation of hue; for example, in the region of the positive v^* axis, hues are greener than in the region of the positive b^* axis. This type of inconsistency

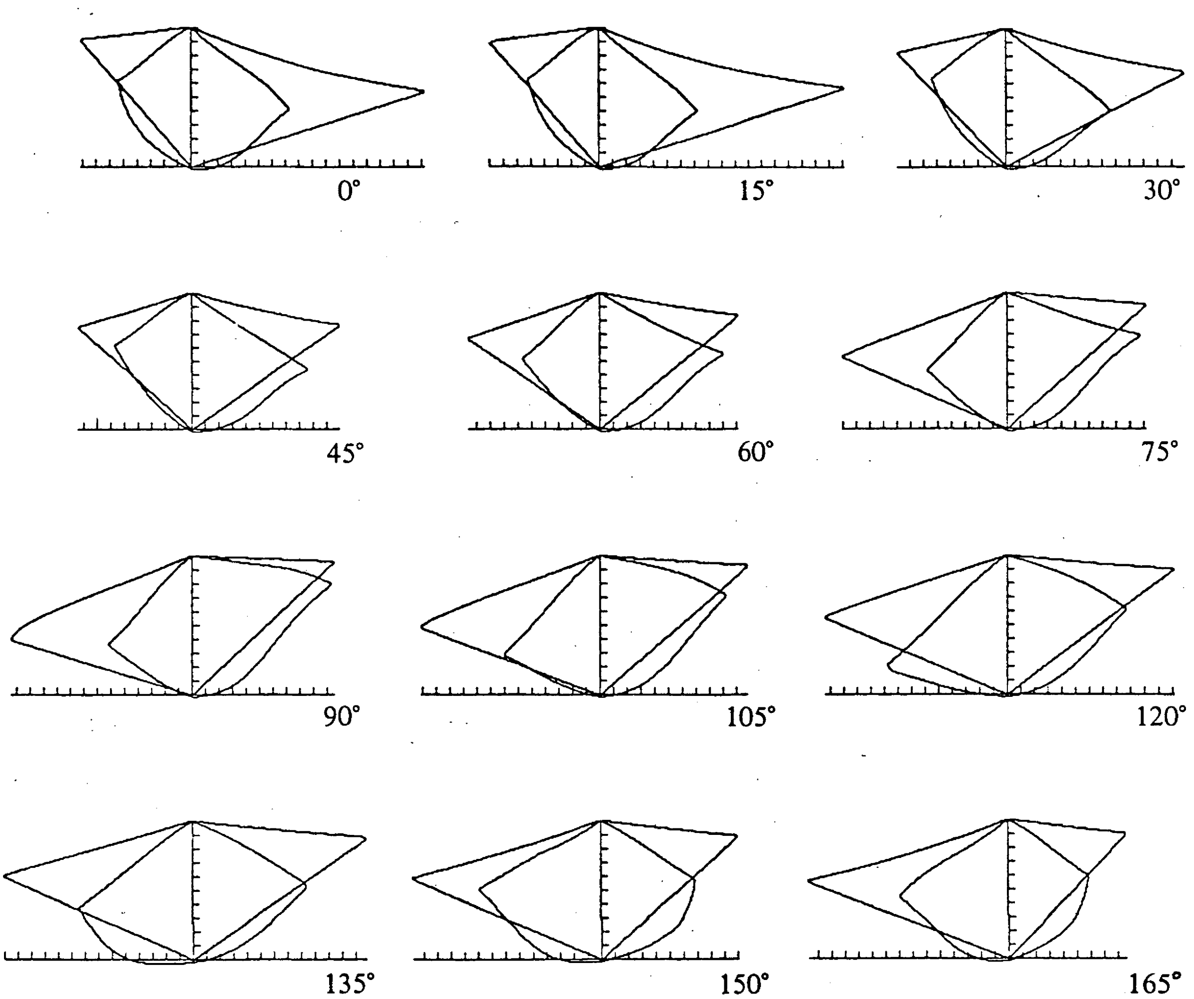


Figure 3.10 Outlines of superimposed cross-sections through the colour film/print and television monitor gamuts. As in figures 3.5(a) and 3.7(a), cross-sections are spaced 15° apart around the L^* axis.

limits the extent to which device-independence of displayed data can be achieved if different spaces are used for different device types. The CIE recommendations for the use of different spaces for subtractive and additive types of display processes are based on subjective judgement of consistency of representation; the inherent physical difference between these processes prevents true reproduction of a set of stimuli on different device types. Thus true device-independence for colour displays may well be confined to comparison of display devices of a similar type; that is, to displays viewed under either reflected or transmitted light conditions, or to displays which are produced by additive sources. The requirements for exact reproduction of colour are treated thoroughly by Pratt (1978) and Horn (1984).

3.5 Summary

This chapter has developed a framework suitable for realising data displays which can be addressed perceptually. Central to the framework is an empirically derived UCS, within which perceptual attributes are naturally aligned. A spectral transformation between perceptual attributes and UCS co-ordinates follows directly from this natural alignment. The UCS metric is defined in terms of perceived colour differences, allowing the proportional representation of numerical data steps by perceived colour differences.

The framework allows for the direct mapping of data into UCS, or into perceptual attributes hue, saturation and lightness (developed in Chapters 4 and 5), and for processing in the spatial and spectral domains for perceptual or lower-level visual purposes (developed in Chapters 6 and 7).

The problem of realising UCS-specified colours, in the form of a transformation to display device gun-counts, has also been addressed; suitable models for two types of display device have been described and implemented. This means that data can be mapped into a perceptually uniform colour space, and the result displayed on either type of display device in a device-independent manner (within the limits of colour accessibility on any particular device).

Chapter 4
The generation of colour sequences for univariate and bivariate maps

4.1 The problem of colouring statistical maps

Statistical maps provide a means of displaying the two-dimensional spatial distribution of statistical variables, using an appropriately chosen graphic such as degree of shading or colour. Developments in computer graphics technology have meant that it is now feasible to use a full colour range, rather than overlaid primary colours, to represent statistical variables in map products. The choice of a colour scheme is crucial to the comprehension of such maps. Fienberg (1979) has highlighted the unsatisfactory state of developments in this area, indicating the need to examine the effectiveness of certain colouring schemes and to develop a theoretical structure for colour utilisation.

Wainer and Francolini (1980) establish an empirical approach to judging the effectiveness of schemes for representing statistical variables in colour. Following Bertin (summarised in Bertin, 1981), they define three distinct levels at which a statistical map can be comprehended. First is at an *elementary* level, in which a direct translation from a perceived variable to a quantitative component is made; this is important in both univariate and bivariate displays. At a second level, the *intermediate* level, trends between two perceived variables are related; in effect the local inter-relationships between two variables are understood, with the univariate analogue being the appreciation of local distribution. At the third or *superior* level, the entire structure or distribution of one variable is compared with that of the other, again with the univariate analogue being appreciation of global distribution. While studies such as Olsen's (1981) stress the value of supporting aids such as legends and textual descriptions to achieve these levels of comprehension, what is of primary importance is that quantitative variables be represented by graphic variables which are in some way intuitively, or perceptually, significant. Wainer and Francolini, consistently with Bertin's terminology, refer to such a variable as a "retinal variable", though the term "perceptual variable" might be more appropriate, bearing in mind the extent to which visual information is processed between the level of physical stimulus on the retina and the psychophysical perceptual level.

In developing a theoretical framework for realising different kinds of graphic variables in colour, Trumbo (1981) identifies four principles which facilitate the comprehension of statistical maps; these principles are stated below.

> (1) *Order*: "If the levels of a statistical variable are ordered, then the colours chosen to represent them should be perceived as preserving the order."
>
> (2) *Separation*: "Important differences in the levels of a statistical variable should be represented by colours clearly perceived as different."
>
> (3) *Rows and Columns*: "If preservation of univariate information or display of conditional distribution is a goal, then the levels of the component variables should not interact to obscure one another."
>
> (4) *Diagonal*: "If display of positive association is a goal, scheme elements should resolve themselves visually into three classes: those on or near the principal diagonal, those above it, and those below."

The first two principles should be satisfied if maps are to be interpreted at an *elementary* level. The third and fourth principles are applicable to bivariate maps, being interpretable in terms of the two-dimensional matrix representing the joint distribution of the variables. Trumbo uses these principles to develop various schemes for colouring bivariate maps, realising the schemes within the framework of two colour models, the Ostwald solid and the Hickethier solid (or unit cube). As described in Chapter 3, the disadvantage of such realisations is that their parameters, and hence perceptual attributes, are dependent on the colour characteristics of any particular display device used.

This chapter approaches the problem of realising Trumbo's and other schemes in uniform colour space (see also Robertson and O'Callaghan, 1984). The principles delineated by Trumbo are used as a guide for choosing the representation, and the levels of comprehension specified by Wainer and Francolini are used as a subjective measure of appropriateness of representation. Section 4.2 examines the generation of colour sequences for univariate and bivariate maps in accordance with Trumbo's schemes within UCS, including techniques for sequence specification. The relative merits of using such a colour space as a framework for generating map sequences are considered in section 4.3; limitations due to spatial and empirical constraints are also discussed.

4.2 Generating colour sequences in uniform colour space

The importance of being able to specify colour sequences in terms of perceptually significant variables is clear from the previous section. Consequently we use the perceptual variables hue (H), saturation (S), and lightness (L) developed in Chapter 3, and illustrated in figure 3.3. These variables form a perceptual colour space (HSL), and are cylindrical-polar co-ordinates of any chosen uniform colour space (UCS), lightness being the vertical cylindrical axis, hue the angular variation, and saturation the cylindrical radius. HSL space is not cartesian; as saturation decreases, hue distinctions become smaller; and conversion to UCS co-ordinates is necessary for determining colour difference magnitudes. As pointed out in Chapter 2, it should be recognised that the perceived hue, saturation, and lightness of a colour will in fact be spatially-dependent on surround lightness or colour; this is discussed further in section 4.3.

4.2.1 Univariate sequences

The mapping of one statistical variable into a line in UCS gives a pseudo-colour display. To satisfy Trumbo's principles of *order* and *separation*, the mapping should be continuous. It also follows from Trumbo's principles that the mapping should be *one-to-one* and *onto*. (Note that at this stage we draw no distinction between quantised and continuous statistical variables, treating quantised data as uniformly spaced between the end points of an ordered sequence. In fact there are reasons to consider quantisation step size; these are discussed in section 4.3.) Although the line should be straight (in UCS) to preserve *order* and equal *separation*, the use of carefully chosen curved lines can give data representations with greater colour space utilisation. In general, provided that the curve is continuous and smooth, and that it has low curvature at any point, the requirement that perceived colour differences uniformly reflect data steps can be met within the inherent inaccuracies of UCS. To ensure preservation of *order*, the use of certain types of curves should be avoided. For example, if a curve is closed, or nearly so, when projected to a plane in UCS, the end-point separation perpendicular to the plane should be large compared with the radius of the curve. In practice this imposes constraints on curves following the path of a coiled spring, or bolt thread, around the lightness axis; the pitch must be large, or the hue range small. Thus while univariate sequences designed for best intuitive interpretation would increment monotonically in perceptual variables H, S, and L, the fact that these variables do not form a cartesian space when the space metric is defined in terms of perceived colour differences means that sequence generation can be a compromise between satisfying the high level requirement for intuitive comprehension on one hand, and the lower level sequence separation and spacing on the other.

By restricting the range of colour space used to any plane including the grey axis we can satisfy both requirements, since such a plane is spanned by perceptual variables lightness and saturation; a straight line path which increments regularly and monotonically either in lightness, or in saturation, or in both, can be chosen, hue remaining constant. Figure 4.1(a) gives an example of a 7-element pseudo-colour sequence generated along a straight line path in UCS between colours blue and yellow in the film/print gamut, together with a 7-level univariate map produced using the sequence. A blue complementary to yellow in the device-model UCS was used to keep the sequence in a constant-hue plane. The path is shown in this constant-hue gamut cross-section in figure 4.3(a), with tick marks indicating the points at which the colour sequence elements in 4.1(a) lie on the line. The broken line shows the projected UCS path of a BGR-cube-specified sequence between the same end points, with tick marks again indicating sequence element points. The advantage of using paths monotonically incrementing in UCS (S and L together), with explicit control over lightness increments and inter-element spacing, is evident. For a colour television monitor this advantage is much more pronounced; central sequence

Image descriptions for the following 2 colour plates

Figure 4.1 Univariate map colouring schemes - examples of legends and maps using:

(a) a 7-element sequence from blue to yellow, regularly spaced along a straight line in UCS;

(b) a 4-element sequence from blue to yellow, regularly spaced along a straight line in UCS;

(c) a 7-element sequence from blue to yellow, monotonically incrementing in hue (clockwise), saturation, and lightness (sequence elements are regularly spaced in UCS co-ordinates);

(d) a 7-element sequence from blue to yellow, monotonically incrementing in hue (anticlockwise), saturation, and lightness (sequence elements are regularly spaced in UCS co-ordinates);

(e) a 40-element sequence from blue to magenta, monotonically incrementing in hue (clockwise), saturation, and lightness (sequence elements are regularly spaced in UCS co-ordinates);

(f) a 40-element sequence from blue to magenta, monotonically incrementing in hue (clockwise) and saturation, and incrementing regularly but not (globally) monotonically in lightness (sequence elements are regularly spaced in UCS co-ordinates).

Figure 4.2 Bivariate map colouring schemes - examples of legends and maps using:

(a) a grey (dark) to blue univariate sequence, being one variable of the bivariate scheme of (c);

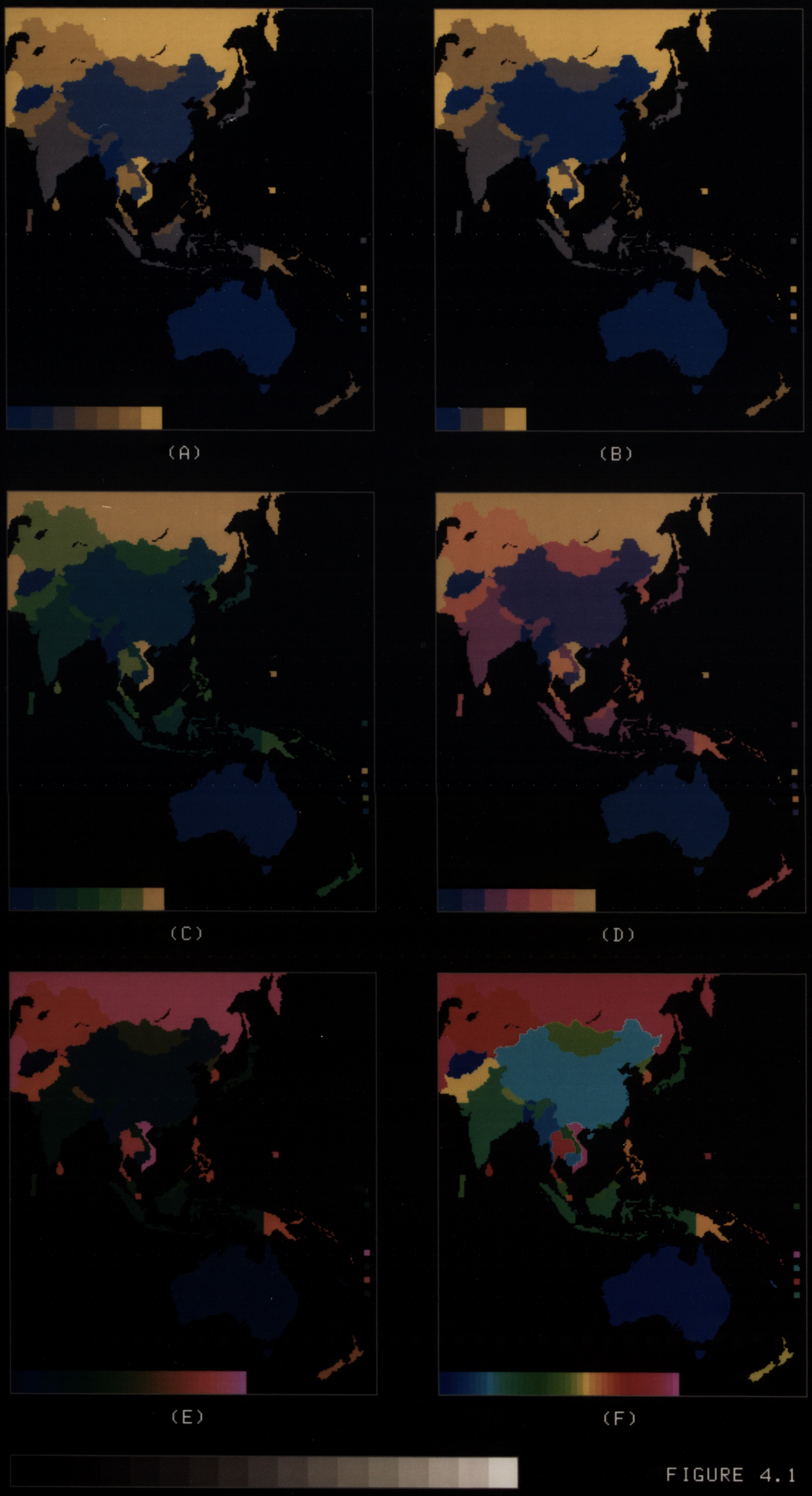
(b) a grey (dark) to orange univariate sequence, being one variable of the bivariate scheme of (c);

(c) a bivariate scheme represented by a parallelogram with one diagonal on the grey axis, lying in a plane of constant (twin) hue in UCS (see figure 4.7);

(d) a bivariate scheme represented by a vertical cylindrical section with its central axis on the grey axis in UCS (see figure 4.8);

(e) a bivariate scheme represented by a vertical spiralled cylindrical section with its central axis on the grey axis in UCS (see figure 4.9);

(f) a bivariate scheme represented by a double conic (one inverted) section with its central axis on the grey axis in UCS (see figure 4.11).



(A)

(B)

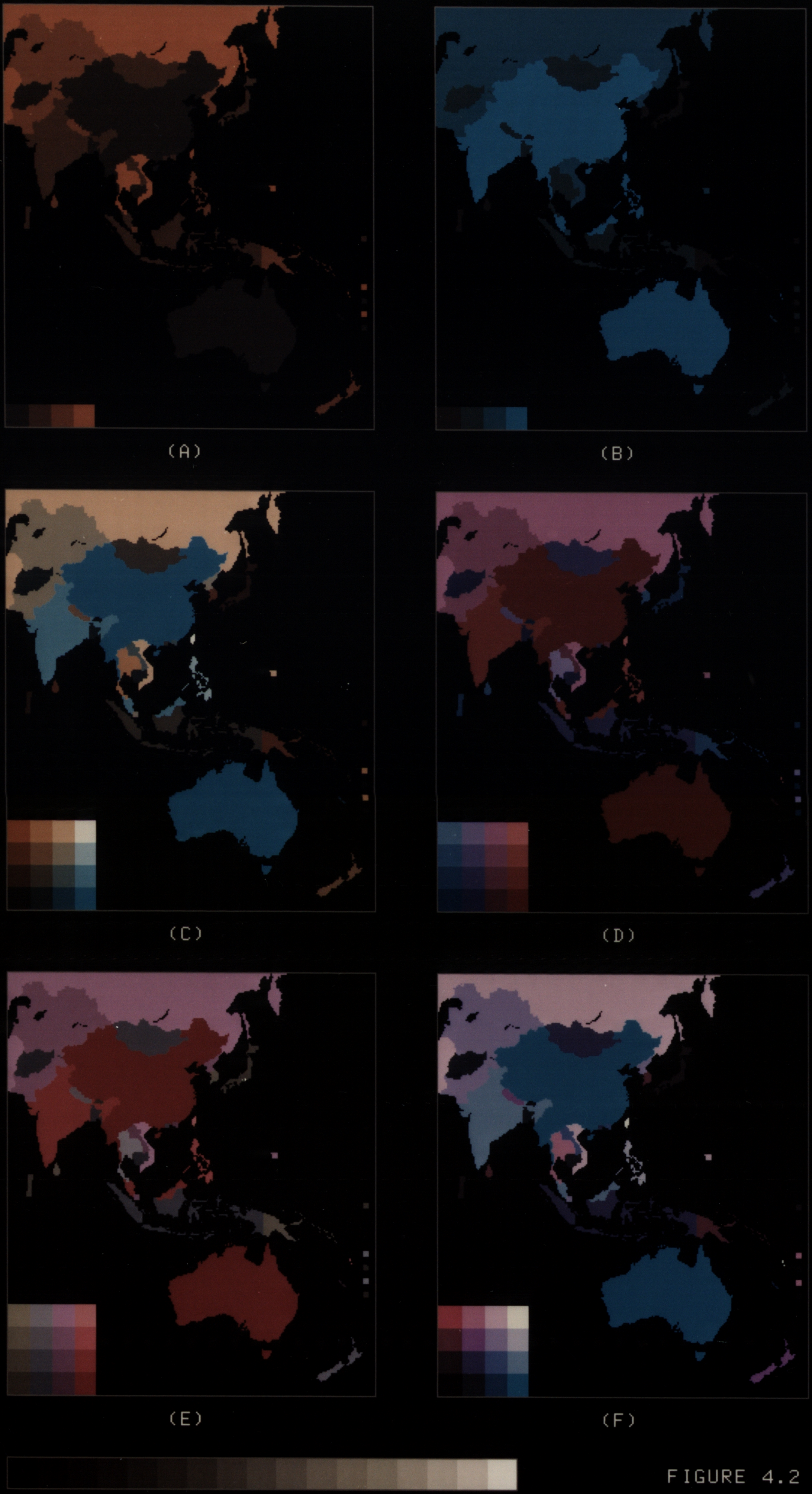
(C)

(D)

(E)

(F)

FIGURE 4.1



(A)

(B)

(C)

(D)

(E)

(F)

FIGURE 4.2

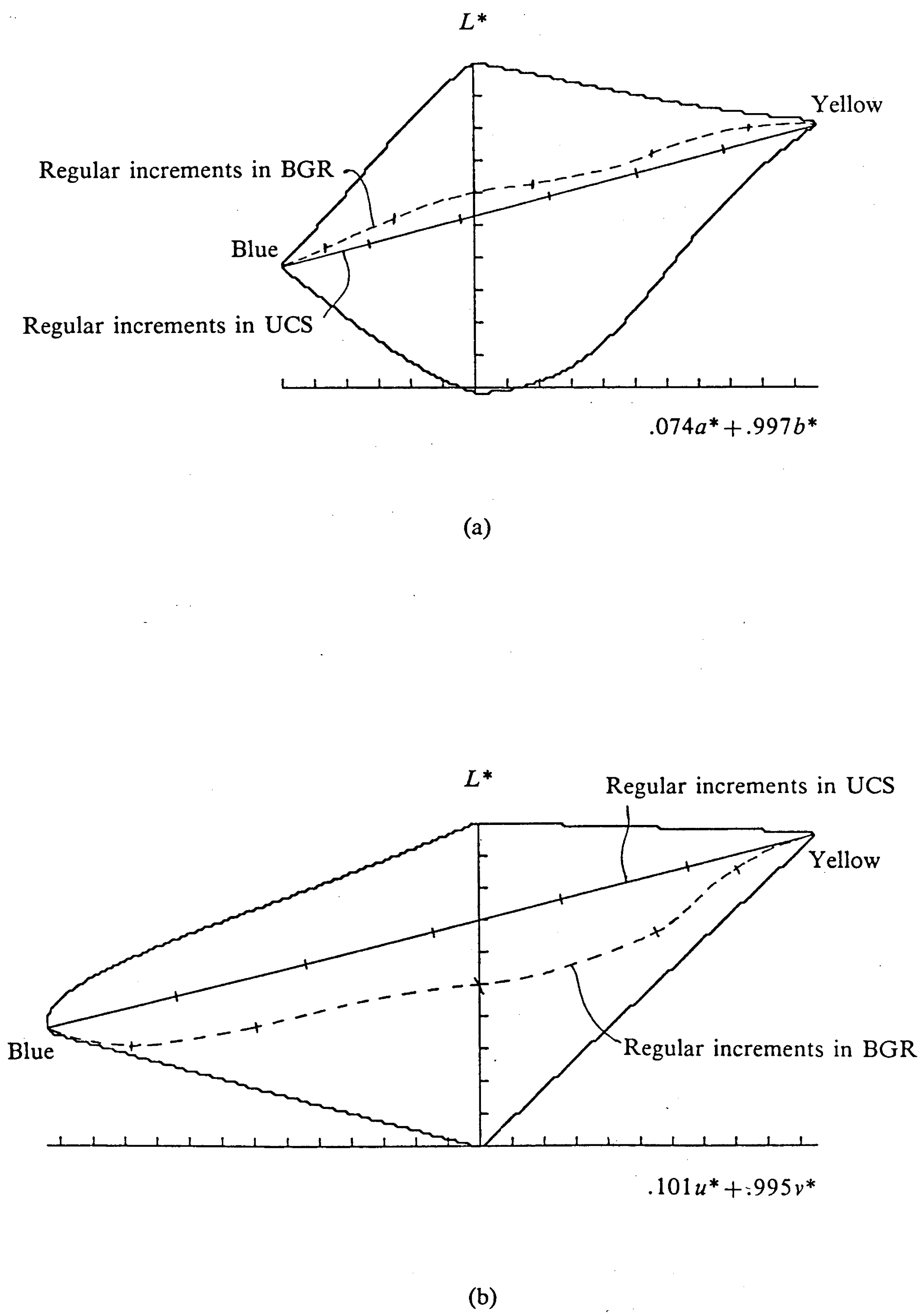


Figure 4.3 Univariate sequence paths through constant-hue gamut cross-sections

(a) Path in the film/print gamut defining the univariate sequence shown in figure 4.1(a) (in bold line), together with the projected BGR-cube-specified path between the same end points (in broken line). Tick marks indicate sequence element locations on the path.

(b) As for (a), but for a similarly defined path in the colour television monitor gamut.

elements in a BGR-defined sequence appear appreciably darker than external elements, specifically because for a colour television monitor the energy output from the guns is not linearly related to numerical (and hence voltage) controls. Instead a power (or "gamma") relation generally applies, with the result that, for example, mid-level excitation of each of two guns can produce a lower nett energy (and indirectly, perceived lightness) than full excitation of a single gun. The equivalently defined paths in the colour monitor gamut are shown in figure 4.3(b); low lightness levels can be seen in central sequence elements in the BGR-cube-defined path. In the device-modelled UCS, however, these "gamma" characteristics are measured and incorporated, as described in Chapter 3.

The common interpretation of a neutral colour (grey) is also used to advantage in this type of sequence, allowing for an immediately apparent rough data dichotomy into "highs" and "lows" either side of grey.

The map shown has variously-sized areas, and a 7-element sequence may mean that due to induction effects (see section 4.3.1), some doubt might arise over the actual colour of small areas. Figure 4.1(b) shows the same data reduced to a 4-level representation; improved *elementary* level interpretation is achieved at the expense of level resolution.

If level resolution and *elementary* interpretation are both important, hue variation can be introduced to increase sequence element *separation* since a longer path in UCS can be used. If appropriately chosen, such a path can still maintain sequence *order*. This involves relaxing the mapping constraints described above, allowing curved paths in UCS defined by sequences monotonically incrementing also in hue.

A sequence in HSL space which increments monotonically in each of H, S, and L has an equation of the form $aH + bS + cL = 1$ where a, b, and c are constants. In UCS, this becomes a cylindrical spiral of constant pitch. The method implemented to generate such sequences is straightforward, requiring specification of the range of the statistical data and its corresponding path in UCS. The colour scheme can be defined by the start and end points, or by the start point and range, for each of H, S and L. It should be noted that while the path through colour space is defined in terms of H, S and L, the spacing between elements along the path must be regular in terms of UCS co-ordinates. This is because at high saturations, a hue increment represents a large colour difference, whereas at low saturations the same hue increment represents a small colour difference. To avoid path parametrisation, a number of sequence elements m larger than the required number n is generated in H, S and L increments, and then approximate re-scaling to achieve regular spacing in UCS co-ordinates is performed. Errors due to this approximation are kept less than one just noticeable difference (JND - the UCS metric unit). Hence we generate an HSL sequence of m elements, index i ($i = 1,...,m$), the values of which are given by

$$H(i) = H_b + H_r\left(\frac{i-1}{m-1}\right) \qquad S(i) = S_b + S_r\left(\frac{i-1}{m-1}\right) \qquad L(i) = L_b + L_r\left(\frac{i-1}{m-1}\right) ,$$

where subscripts b and r denote specified base values and ranges for each of H, S, and L. We then convert each element of the sequence of m elements to UCS co-ordinates $\{U_k,\ k=1,3\}$ (using the transformation given in Chapter 3), and calculate the cumulative piecewise-linear path distance to sequence element i, D_i, in UCS from

$$D_i = \sum_{h=2}^{i} \left\{ \sum_{k=1}^{3} [(U_{k,h} - U_{k,h-1})^2] \right\}^{\frac{1}{2}} ,$$

where $D_1 = 0$.

Taking the average distance along the path between each pair of n required elements as

$$d_{av} = D_m/(n-1) ,$$

for every j, $j = 1,...,n$ we find an i such that

$$D_i < (j-1)d_{av} \leqslant D_{i+1} \ , \qquad i \in [1,m-1]$$

and set $U'_{k,j} = U_{k,i}$ for $k=1,3$.

This generates a set of n elements approximately equally spaced in UCS. The closeness of the approximation depends on the value of m, and on the actual path in UCS. In practice, the value of m was chosen to keep errors in element spacing less than one JND in UCS.

Paths with a hue variation are more closely constrained by the shape of any particular display device gamut, and some indication of the limits of saturation is required to be able to specify practical paths. Figures 4.4(a) and (b) illustrate constant lightness cross-sections through the UCS gamuts of each of the the colour film/print recording system and the colour television monitor, with paths which monotonically increment in H and S, each touching the gamut boundary in two places. On the colour monitor it is not feasible to use fully saturated primary colours in a sequence specified as incrementing monotonically in H and S because intermediate points would be over-saturated; if they were projected to appropriate points on the gamut edge, the resulting sequence would not be monotonically incrementing in terms of perceptual attributes. The same holds if increments in lightness are also incorporated. In the gamut of the film/print recording system such a path comes closer to including more maximally saturated colours.

For specified hue and lightness base values and ranges, the display device gamut limits the saturation of the sequence elements, and hence restricts the saturation base and range values, but not uniquely. Considerations such as maintaining sequence *order*, and implying an increasing sense of magnitude, will govern the choice of these values, but to define an appropriate path in any particular display device gamut without some form of graphic aid is not easy. Interactive choice of path sequence can be made using gamut shapes as an initial guide. Given the start and end hues and lightnesses of a required sequence, the HSL gamut can be projected to a two-dimensional rectangular representation to show the gamut boundary (and hence limiting) saturation at any hue angle along the path. Such a projection is shown in figure 4.5; the colour film/print gamut boundary along a path defined by monotonically incrementing H and L is depicted. (It should be noted that this is not the boundary which is obtained by taking a UCS gamut cross-section through the specified end points and containing the line perpendicular to both the lightness axis and the line joining the end points.) Hue increments in an anticlockwise direction around the gamut from blue to yellow (on the left hand side) and back to blue (on the right

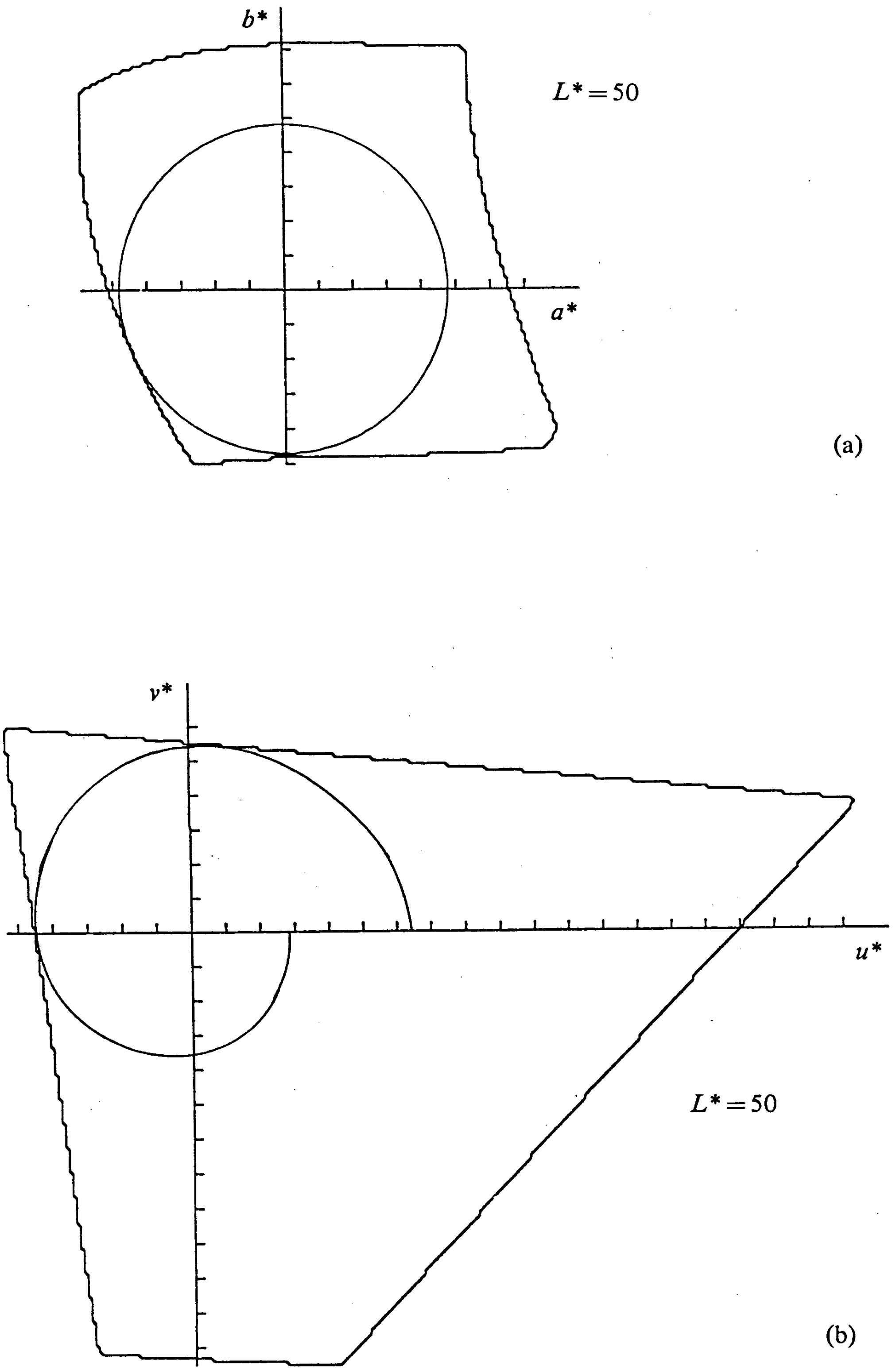


Figure 4.4 Saturation limits of univariate sequence paths in constant-lightness gamut cross-sections

(a) Colour film/print gamut

(b) Colour television monitor gamut

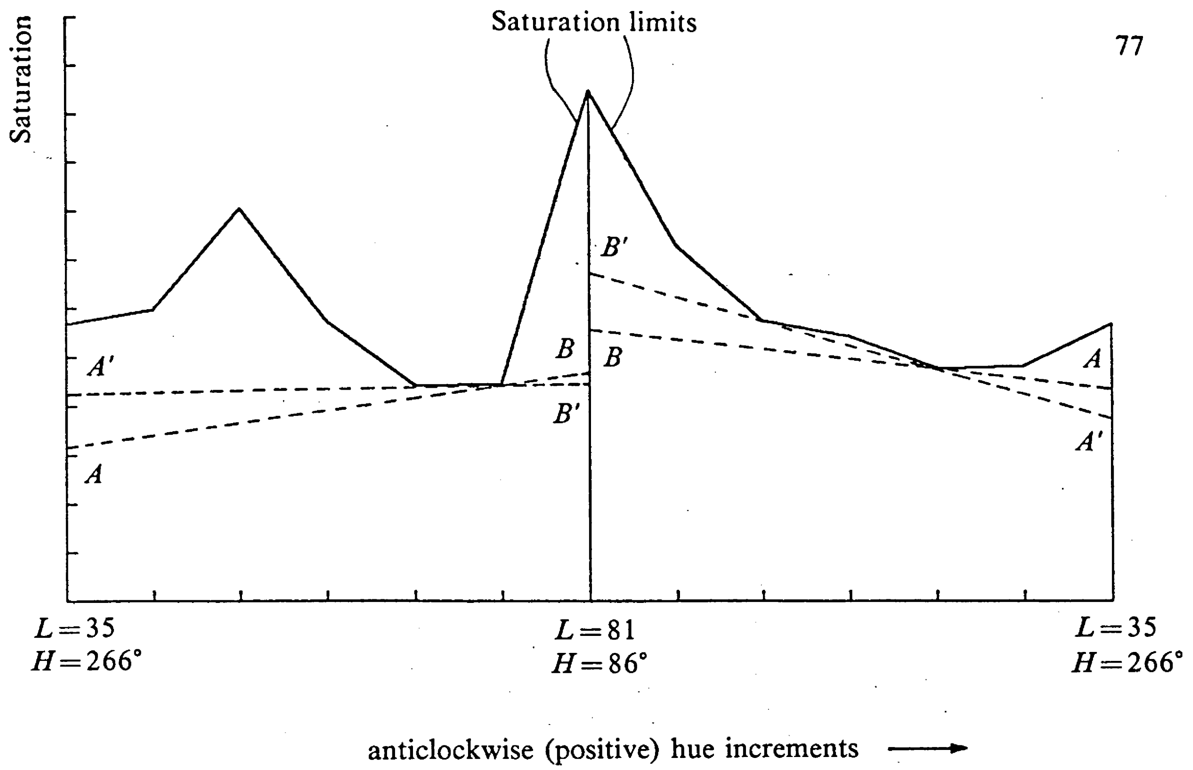


Figure 4.5 Graphical projection of gamut saturation limits for univariate sequence specification. The maximum saturation at hue and lightness values determined by the specified end points of the sequence¯ is shown. For a sequence monotonically incrementing in hue, saturation, and lightness, any straight line (such as *AB* or *A'B'*) which remains inside the boundary projection can be used.

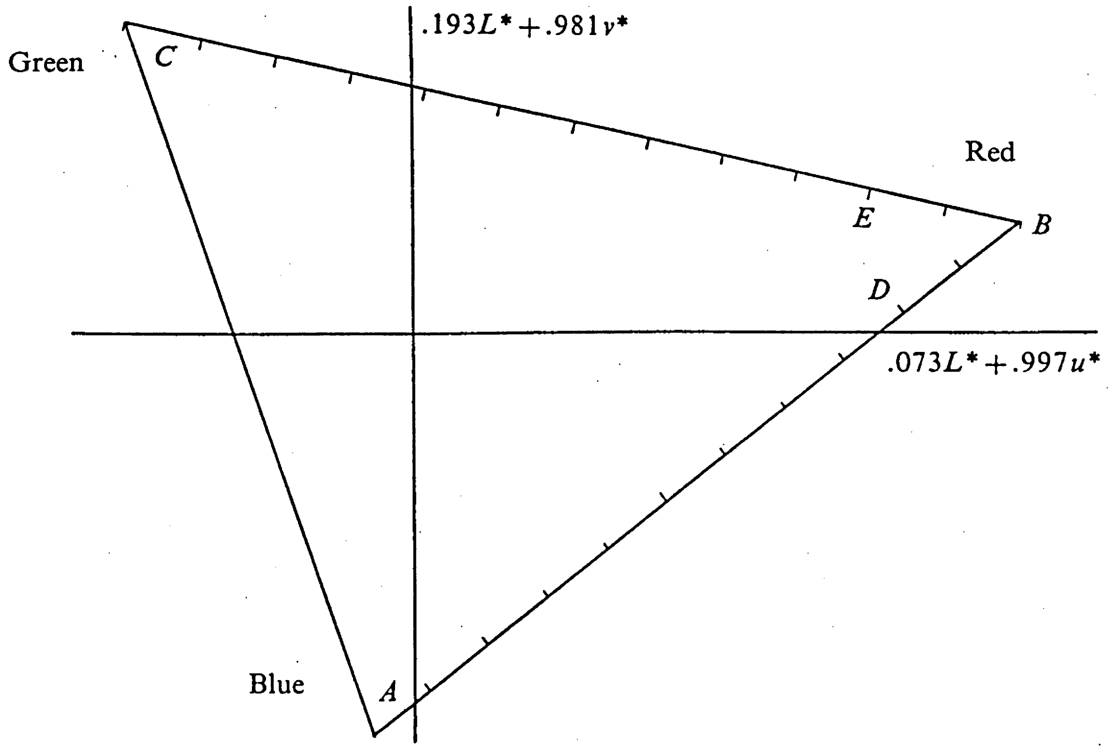


Figure 4.6 Poorly-chosen univariate sequence paths in the colour monitor gamut. In the illustrated path, sequence points *D* and *E* can be closer in colour than either *D* and *B*, or *B* and *E*.

hand side). A sequence between these end points can then be chosen by taking any (graphically) straight line which remains inside the gamut boundary, such as the lines *AB* or *A'B'* shown. Perceptual or data-dependent considerations can be used to determine the path encompassing the most appropriate range in saturation, and sequences for any specified path can then be produced. In the practical generation of such sequences, checks can be made in UCS on curvature and on spacing between non-adjacent sequence elements, in order to trap such paths as spirals with very low saturation throughout which are unlikely to satisfy Trumbo's first two principles (*order* and *separation*).

Figures 4.1(c) and 4.1(d) show maps produced using 7-element sequences resulting from monotonically incrementing each of *H*, *S* and *L* both clockwise and anticlockwise around the film/print gamut, between the colours blue and yellow. These sequences correspond to the paths *AB* in figure 4.5. In each case the defined path covers half the available hue range, though the path through red is slightly less saturated, and hence element *separation* should be marginally smaller. Spacing along the path is regular in UCS. The loss of saturation necessary to keep colours within the gamut can be seen when comparing end-point colours with those of 4.1(a). For a colour television monitor this loss of saturation is greater because of the less convex gamut shape.

When a sequence with a large number of elements is required, increasing the hue range improves level discrimination in a map. Figure 4.1(e) shows a map produced using a 40-element sequence encompassing a hue range of about 300°, between the colours blue and magenta, monotonically incrementing in *H*, *S*, and *L*. Here again the gamut shape limits saturation and lightness, but the resulting colour progression remains intuitive. The penalty of increasing *separation* by increasing the total hue range is that sequence continuity, and the intuitive concept of sequence *order* (Trumbo's first principle), become less well preserved in the map itself; appreciation of the overall distribution (*superior* comprehension) is less well achieved. The lightness increase through the range helps to maintain this *order*.

The mapping constraints can be further relaxed in order to maintain a higher level of saturation in sequences incorporating a hue variation. By using the maximum saturation available on a display device at any specified hue and lightness, we can define a path by

$$H(i) = H_b + H_r\left(\frac{i-1}{m-1}\right) \qquad S(i) = S_{max}(H(i),L(i)) \qquad L(i) = L_b + L_r\left(\frac{i-1}{m-1}\right) ,$$

where subscripts b and r as before denote specified base values and ranges, and S_{max} is the maximum saturation in the chosen display device gamut at specified hue and lightness values. This amounts to ignoring the perceptual significance of saturation as an ordering attribute, no longer requiring either linearity or *one-to-oneness*, in order to achieve greater element *separation*. Again, in generating such a sequence, re-scaling along the path is neccessary to achieve the required regular spacing in UCS. Saturation was maximised in this manner in the map and 40-element sequence shown in figure 4.1(f). In addition, the UCS path of the sequence was constrained to pass through the device-primary and device-complementary points (in the order blue, cyan, green, yellow, red and magenta); consequently lightness is no longer monotonically increasing globally, but rather monotonically increasing or decreasing between the specified points. The increased saturation along the sequence can be seen when comparing it with that of 4.1(e). Direct value (*elementary*) interpretation is improved at the expense of appreciation of the overall distribution (*superior* comprehension).

Knowing the shape of a device gamut in UCS is also helpful for detecting unsuitable paths for colour sequences. Figure 4.6 shows a cross-section through the colour television gamut which includes the three device-primary colours. Tajima (1983) suggests straight line segments between device-primary hues in UCS for generating pseudo-colour sequences (path AB), with two such segments (path ABC), to include a greater hue range. The drawback with such a path is that because of the highly pointed nature of the colour monitor gamut at the primaries, two non-adjacent quantised samples could appear very close in colour (points D and E). In such gamuts it seems more important to maintain the *order* and *separation* of colours at the expense of some saturation, indicating that it is more appropriate to use paths monotonically incrementing in H, S and L.

4.2.2 Bivariate sequences

Bivariate sequences pose the problem that two suitable univariate axes in UCS must be found, with the constraint that they must have a common end point. Ideally these two axes should be orthogonal in perceptual (HSL) colour space. As in the univariate case, in practice we can relax these constraints to achieve particular associations between perceptual representations and specific data. However, requiring intuitive comprehension of both the individual and joint distributions of two variables imposes more severe limitations on appropriate perceptual alignment, and on curvature and *one-to-oneness.*

The following bivariate colouring schemes all have sequence elements in positions in UCS corresponding to a (possibly warped) two-dimensional matrix lying on the geometrical surface specified. This matrix hence forms the bivariate map legend.

Trumbo suggests two types of scheme, each of which partly satisfies his bivariate principles. The first is a square lying in a plane, or curved surface, with its principal diagonal along the grey axis in colour space as shown in figure 4.7. Considerations of *order* and equal *separation* suggest a square lying in a plane in UCS as being the most suitable bivariate representation, though as in the univariate case, a square mapped on to a surface with low curvature (compared with inter-element spacing) might also be appropriate. Similarly a parallelogram with one diagonal on the grey axis, and restrictions on acuteness of angle, might be satisfactory. The H, S, and L values for such an $m\ x\ n$ bivariate legend with two-dimensional indices i and j ($i = 1,..,m$, $j = 1,..,n$) are given by

$$H(i,j) = \begin{cases} H_b & \text{for} \quad i/m \leqslant j/n \\ H_b + H_r & \phantom{\text{for}} \quad i/m > j/n \end{cases}$$

$$S(i,j) = S_b + S_r\left(\frac{i-1}{m-1} - \frac{j-1}{n-1}\right) \qquad L(i,j) = L_b + L_r\left(\frac{i-1}{m-1} + \frac{j-1}{n-1}\right),$$

where as before subscripts b and r indicate base values and ranges respectively. In this type of scheme only the third principle (*rows and columns*, ensuring preservation of univariate information) is not satisfied, as the rows and columns of the matrix are not aligned with a single perceptually significant attribute (H, S or L). To determine whether

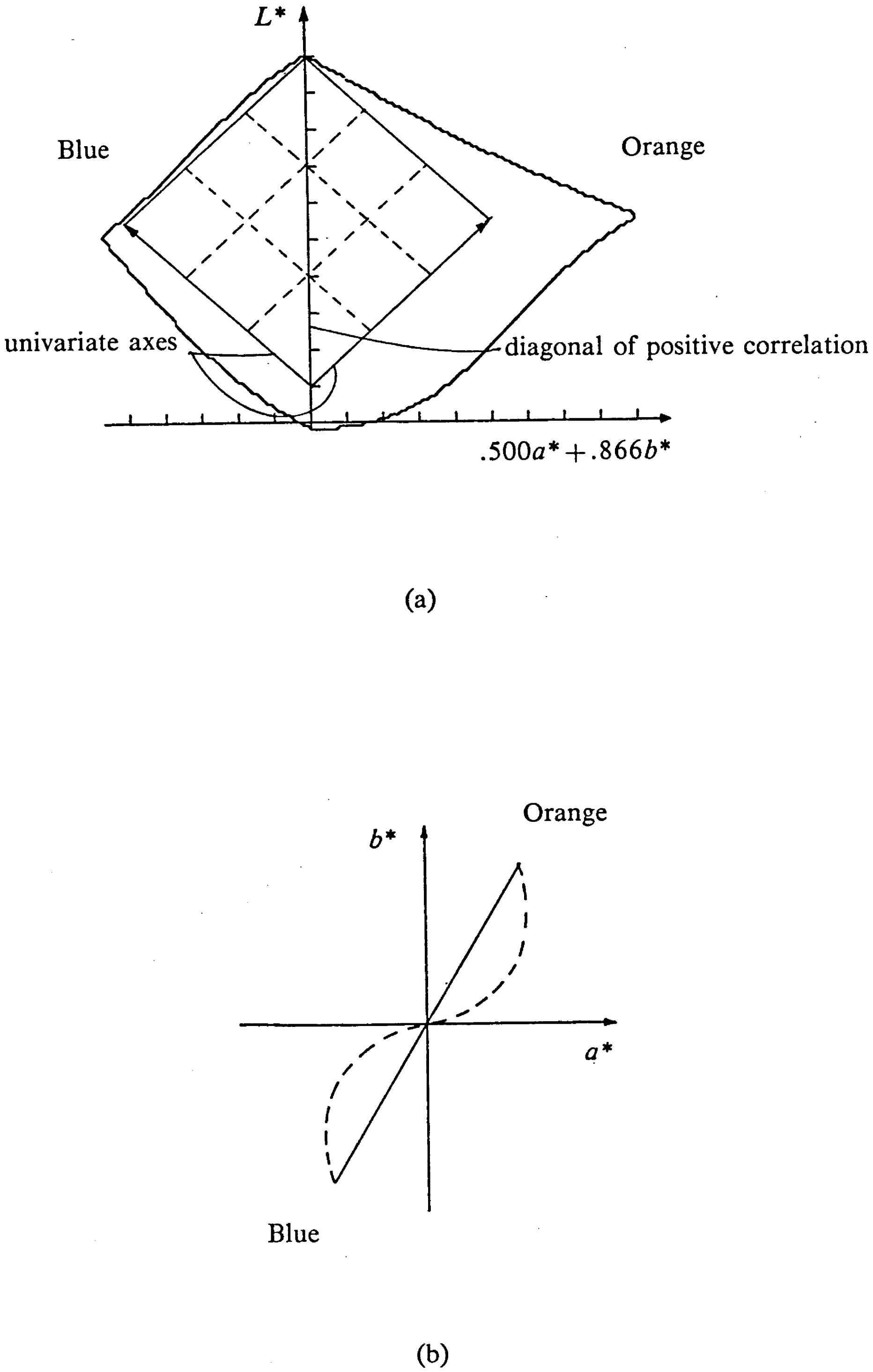


Figure 4.7 Schematic of a bivariate legend in a constant-hue plane represented by a paralleogram with one diagonal on the grey axis in UCS, illustrated in legend and map form in figure 4.2(c). A plan view is shown in (b); the broken line represents a possible scheme to include a hue variation which is inappropriate on perceptual grounds (see text).

in fact a correlated increase in lightness and saturation can be treated as an intuitively meaningful attribute would require some form of empirical evaluation study such as those of Wainer and Francolini or Olsen. Figure 4.2(c) shows a bivariate map with its legend derived from a parallelogram lying in a plane in UCS with the positively correlated diagonal along the grey axis; to distinguish the lowest values from the background black, the grey range used was 10-100 (0 is black, 100 is white). Saturation was maximised for the gamut cross-section used, under the constraint of equal saturation for each variable. This form of scheme symmetrically represents the two variables, allowing the use of univariate maps in conjunction with a bivariate map; figures 4.2(a) and (b) show the two different univariate maps whose information is combined in the bivariate map of 4.2(c). The colour ranges used are consistent between maps, though in the univariate maps the lowest level was raised to 20 to better separate it from the background black. This bivariate map is used in subsequent examples.

Trumbo also suggests the use of a curved plane with its diagonal on the grey axis, using the curved slice to include a variety of hues as shown by the broken path in figure 4.7(a). This scheme gives a hue *order* which is not intuitive, and a hue utilisation which is not *one-to-one*; the small increase in *separation* achieved over the scheme described above appears to be more than offset by the sacrifice in hue *order*.

The second type of scheme suggested by Trumbo is that of a surface which forms part of a cylinder in colour space, as shown in figure 4.8. Such a legend has *H*, *S*, and *L* values, for an $m \times n$ legend with indices i and j ($i = 1,..,m$, $j = 1,..,n$) given by

$$H(i,j) = H_b + H_r\left(\frac{i-1}{m-1}\right) \qquad S(i,j) = S_b \qquad L(i,j) = L_b + L_r\left(\frac{j-1}{n-1}\right) ,$$

where subscripts b and r again denote specified base and range values. In this case the third principle (*rows and columns*) is satisfied, as the rows are lines of constant lightness and saturation and vary in hue, while the columns are of constant hue and saturation and vary in lightness. Univariate information is thus represented by perceptual attributes. The fourth principle (*diagonal* - the depiction of positive and negative correlations) is not satisfied. Figure 4.2(d) shows an example of an implementation of this scheme in UCS.

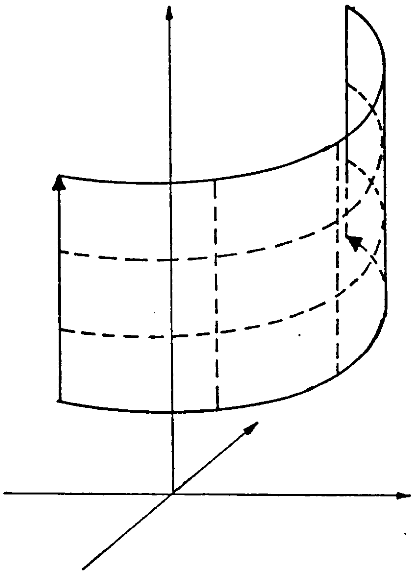


Figure 4.8

Schematic of a bivariate legend in a vertical cylindrical section with its central axis on the UCS grey axis.

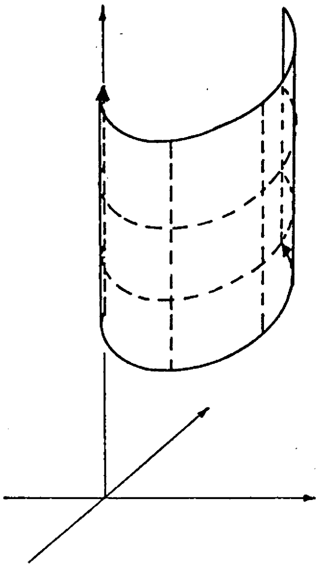


Figure 4.9

Schematic of a bivariate legend in a vertical spiralled cylindrical section with its central axis on the UCS grey axis.

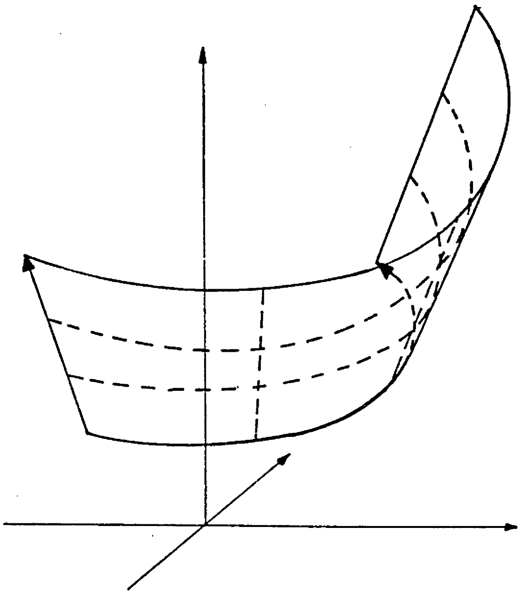


Figure 4.10

Schematic of a bivariate legend in an expanding vertical cylindrical section with its central axis on the UCS grey axis.

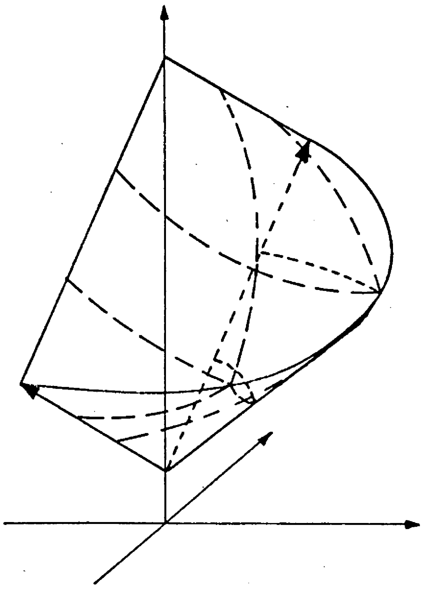


Figure 4.11

Schematic of a bivariate legend in a vertical double-conic (one inverted) section with its central axis on the UCS grey axis.

A variation on this scheme, producing a legend which is an expanding cylinder, or in plan view (along the *L* axis) a spiral, allows for a grey axis to be used as one variable. This is shown in figure 4.9, and has hue and lightness as defined for Trumbo's second scheme. Saturation is given by

$$S(i,j) = S_b + S_r\left(\frac{i-1}{m-1}\right),$$

thus introducing a hue-dependence. A legend resulting from this scheme, and a map produced using it, are shown in figure 4.2(e).

Alternatively, lightness-dependence can also be introduced (figure 4.10) using saturation given by

$$S(i,j) = S_b + S_r\left(\frac{j-1}{n-1}\right).$$

This scheme can take advantage of the conical aspect of the UCS gamut shape of a colour television monitor (see figure 3.5), though it is less effective for a display device with a less conical gamut shape; consequently it is not shown in legend and map form for the film/print gamut. A disadvantage of this scheme, inherent in using univariate axes along the circumference and axial directions of a conic section, is that *separation* between the elements of one variable becomes dependent on the value of the other variable, thus introducing a perceptual correlation. In the schematic example shown, inter-element *separation* of the variable depicted by hue is also dependent on the value of the variable depicted by lightness; the relative merits of the increased saturation must be weighed against this drawback.

In addition to Trumbo's two recommended schemes, a third was implemented. This consists of a section from a double cone (one inverted), with univariate information represented by a progression in hue, saturation and lightness combined. Specifically, for an $m\,x\,n$ legend, we have

$$H(i,j) = H_b + H_r\left(\frac{i-1}{m-1} - \frac{j-1}{n-1}\right)$$

$$S(i,j) = S_b + 2S_r\left(\frac{L_b - L(i,j)}{L_r}\right) \qquad L(i,j) = L_b + L_r\left(\frac{i-1}{m-1} + \frac{j-1}{n-1}\right) .$$

Diagonals of positive correlation are depicted by constant hues, while diagonals of negative correlation are depicted by constant lightness and saturation. This scheme is effectively a modification of Trumbo's first scheme to include a range of hues, in order to increase sequence element *separation*. Figure 4.11 shows the form of the surface, while figure 4.2(f) shows an example of the bivariate legend which results, and a map using it.

Realisation of each of these schemes can be made in the HSL co-ordinate system, with hue, saturation and lightness base levels and ranges being specified as in the univariate case. For legends which have non-zero ranges of both saturation and hue, for equal perceived spacing, re-scaling to a UCS metric must be performed. As in the univariate case, generating an expanded set of elements and then re-sampling at UCS-distance determined intervals produces a legend with adjacent elements approximately equally spaced in UCS; the closeness of the approximation can be improved by increasing the expansion factor. In fact, of the schemes shown, only those of figures 4.9 (displayed in 4.2(e)) and 4.10 require this re-scaling. As implemented, the method of legend generation allows for specification of colour scheme type and HSL parameters, returning suggested modifications to these parameters if the desired ranges result in colours outside the gamut of the display device being used. Guides to appropriate attribute ranges for each scheme are also provided, with excessive range or curvature being trapped (excessive hue range and hence hue spacing between sequence elements makes comprehension of *order* very difficult).

4.3 Discussion and evaluation

4.3.1 Uniformity and sequence element separation - the influence of induction

In the sequences produced uniformity is not always closely achieved, as indicated by the apparent variations in sequence element *separation*. This could be caused by poor display device modelling (see Chapters 3 and 8), but it should also be pointed out that the defining equations for the CIELAB and CIELUV uniform colour spaces only approximately predict the experimental results leading to their formulations; exact uniformity would thus not be expected. Also relevant is that these spaces have a metric which is derived from experiments on the perception of just noticeable colour differences. This makes them suitable for displaying fine gradations of data, but it does not follow that more widely spaced colours with equal Euclidean spacing will necessarily have equally perceived differences. It is quite possible that if sequence element spacing is to be large, a UCS such as the OSALJG space might be more appropriate, since in the empirical derivation of this space, the judged colour differences were comparatively large. Alternatively, a system such as the Coloroid Color System (Nemcsics, 1980) might be more appropriate if a full range of hues is used, since it is based on uniformity when viewing the colour space as a whole, rather than when viewing small colour differences. However, the widespread acceptance of the CIE uniform colour spaces, the knowledge of their limitations and anomalies (see, for example Pointer, 1981), the availability of measuring instruments which use them, and their analytical tractability make them particularly attractive for use at this time.

Another consideration is that exactness of colour differences may well be more than offset by the effect of adjacent colours on the perceived nature of any particular colours. This spatial effect, known as induction, can alter the apparent lightness, hue or saturation of a surrounded colour quite substantially. This is particularly evident in maps with areas of different sizes; the perceived colour of the smaller areas is heavily affected by surrounding homogeneous large areas. This can be seen in the map chosen for illustrating colouring schemes in figures 4.1 and 4.2. While schemes to compensate colours according to their surrounds have been proposed (see Wyszecki and Stiles, 1967; Oyama et al., 1980; Ware and Cowan, 1982), such an approach has the drawback not only of being dependent on viewing distance but also that direct legend comparison becomes impossible. Olsen's findings on the importance of presenting legends with maps suggest that such an approach seems at this stage impracticable. Bearing in mind the influence of induction on perceived colour in any map, exact uniformity of the colour space is unlikely to be critical.

These factors have a bearing on the number of discrete levels appropriate for representing variables in a statistical map. If the *elementary* level of comprehension is of primary importance, the number of discrete levels must be such that induction effects are small compared with inter-level spacing. Figure 4.1(a) shows a 7-level map in which the inter-element colour spacing is probably too small to avoid misinterpretation caused by induction. It was for this reason that the map in 4.1(b) was produced, the number of discrete levels being reduced to 4. It would be quite feasible to use some form of induction model to derive expected maximum perceived excursions of a colour under given spatial distributions, and limit the number of discrete levels accordingly. However, this might not be necessary if achieving the higher levels of comprehension is more important; a moveable legend for direct adjacent comparision might solve the problem of *elementary* level identification. Trumbo and others have suggested that for bivariate maps, the use of more than four classes for each variable is likely to confuse rather than provide additional information, and it is for this reason that four classes were used in the examples shown.

4.3.2 Limitations on the independence and use of perceptual attributes

Implicit in our use of perceptual attributes hue, saturation and lightness have been two assumptions: that the attributes can be considered as independent, and that such independence is maintained under varying conditions of spatial arrangement. Neither assumption is obviously or necessarily valid; indeed many known effects, such as the indistinct appearance of isoluminant chromatic borders, or the difficulty in judging the relative lightness of two very different hues, suggest that the assumptions should be treated cautiously. Such conditions could be used as constraints on the choice of suitable schemes.

The existence of three perceptual attributes suggests the possibility of producing trivariate maps. However, as pointed out by Trumbo, the HSL co-ordinate system does not form a cartesian space; at low saturations hue distinctions are meaningless, while at extreme lightnesses, hue and saturation are degenerate. Thus such use of the space would have to be limited to mid-lightnesses and outer saturations. Given these factors, and the difficulties experienced in comprehending bivariate maps, trivariate maps relying on colour for discerning between variables are unlikely to satisfactorily achieve any of the three comprehension levels cited earlier.

From a subjective viewpoint, the univariate scheme of figure 4.1(a) appears to achieve the *intermediate* (local distribution appreciation) and *superior* (global distribution appreciation) levels of comprehension well, but direct interpretation at an *elementary* level is less well achieved unless the number of discrete levels is made small. Sequences incorporating hue variation, such as those of 4.1(c) and (d), result in improved *elementary* interpretation, but as the hue range encompassed increases (4.1(e) and (f)), the higher levels of comprehension are less well achieved. Allowing lightness to vary non-monotonically through the sequence (4.1(f)) further reduces the degree to which the higher comprehension levels are achieved.

In the bivariate map representations, Trumbo's first scheme (a plane of constant hue (and its complementary) with its diagonal on the grey axis) appears to best achieve the higher two levels of comprehension while less clearly depicting the actual specific values at the *elementary* level of comprehension. The scheme with hue variation representing one variable, and lightness variation representing the other, lends well to comprehension at an *elementary* level, but poorly at higher levels. This gives credence to being cautious about assuming the independence of hue, saturation and lightness, and suggests that considerations of the visual purposes of these perceptual senses such as those of Rubin and Richards (1982) or Marr (1982) might lead to the development of more appropriate schemes.

In fact according to Bertin (1981), the chromatic attributes of colour are useful to differentiate properties, but not to convey a sense of order or relative magnitude in a statistical map. He suggests that lightness is the dominant attribute for this purpose. However, this does not preclude the use of lightness as the ordering attribute in each of two represented variables, and hue and saturation as attributes to convey the balance between the two variables. This, in effect, is the use of perceptual attributes in the bivariate schemes shown in figures 4.2(c) and (f), and might well explain the subjective preference for these schemes over those shown in figures 4.2(d) and (e). These observations reinforce the need to consider carefully both the suitability of graphic representations in the form of perceptual attributes, and the realisation of these attributes in colour in as controlled a way as possible.

The choice of univariate and bivariate colouring schemes may in practice be governed by other considerations. For example, it is often desirable to display a particular statistical variable both in univariate form, and with another variable in bivariate form, using the same colouring scheme for each (as shown in figures 4.2(a), (b) and (c)). Such a univariate display forms a useful adjunct to a bivariate display, and consequently a bivariate scheme in which the univariate axes both satisfy the same criteria of, say, *order* and *separation* (such as the schemes shown in figures 4.2(c) and (f)), may be preferable to one in which such criteria are satisfied differently for each variable (such as the schemes shown in figures 4.2(d) and (e)).

4.4 Conclusions

The purpose of this chapter is to show how schemes such as those proposed by Trumbo for colouring univariate and bivariate maps can be realised within a computational framework based on uniform colour spaces. The essential features of such spaces; the perceptually significant ordering method and intuitive addressibility, and the uniformity of the metric in terms of perceived colour differences; make it feasible to use them for specifying colour sequences suitable for displaying statistical variables in map products.

The framework described in Chapter 3 allows for the generalised mapping of data of up to three dimensions into a uniform colour space, and realisation on various colour display devices. The application of this approach to producing univariate and bivariate maps is straightforward, involving only constraints on the mapping to satisfy basic tenets, such as those specified by Trumbo, appropriate to map sequence generation. It is difficult to measure the effectiveness of the colour sequences produced. As pointed out by Wainer and Francolini, the success of colour schemes for mapping is likely to be judged partly on the basis of the experience of experts, and partly on the basis of empirical experiments. However, from a subjective viewpoint, despite uniformity limitations the colour sequences produced appear to satisfy the principles leading to their specification. Consequently it seems worthwhile using such an approach for the generation of colour map sequences, both for investigative purposes and for high quality products.

Colour schemes produced using the approach described in this chapter have been given practical use in an interactive colour mapping system (O'Callaghan and Simons, 1983). In general, it has been found that such schemes offer a distinct improvement over those realised without the use of uniform colour space and display device modelling. These results suggest that the essential features of uniform colour spaces stated above can in fact be usefully realised within the developed framework. The display device- and process-independence of this approach also make it more feasible to reproduce maps on various types of display device with a controlled consistency of representation.

Chapter 5
Colour representation of remotely sensed reflected radiation

5.1 Remotely sensed reflected radiation - the general display problem

In this chapter we consider the colour display of image data from aircraft or satellite scanner systems which detect the radiation reflected from the earth's surface. Such displays are used to examine terrain structure and surface cover (including soil and vegetation properties). As mentioned in Chapter 2, information allowing the depiction of terrain structure, or surface topography, is inherently embedded in such data sets.

5.1.1 Data types and sensing systems

Remote sensor systems convert detected radiance to a numerical value. Radiance is the radiant flux (per unit solid angle) leaving the source in a given direction per unit projected source area in that direction; it is thus dependent on the source irradiation and orientation (see Chapter 2, Volume II of Colwell, 1983). The spectral bandwidths of sensors are chosen to discriminate between the reflectance characteristics of physical properties of interest; in order both to span a sufficient spectral range, and to achieve sufficient spectral resolution within that range, multiple sensors sensitive to different spectral bands are used, generating multi-dimensional data sets. The Landsat satellite multi-spectral scanner (MSS) generates four spectral bands in the visible and near-infrared parts of the spectrum.

If the spectral bands of a multi-sensor system overlap, some spectral correlation in the data is inherent. In addition, if the spectral reflectance distributions from physical objects overlap the spectral bands of more than one sensor, further spectral correlation between bands results. Thus while multi-sensor systems such as the Landsat or airborne multi-spectral scanners produce data sets which have an absolute dimensionality equal to the number of different sensor response distributions, the individual spectral data channels are often highly correlated.

The aims in displaying such data sets can be specific, such as looking for, or at, certain relationships between spectral characteristics and spatial structures; they can also often be ill-defined, relying on the processes of visual inspection and detection to guide the analysis. Remotely sensed multi-spectral data sets are widely used for interpreting geological, geophysical, ecological and other physically-based properties, and because of their multi-dimensionality, colour is generally used as a display aid. The problem is how, with the variety of data types and analysis purposes, colour can most effectively be used; this chapter looks at how the perceptual attributes of uniform colour spaces can be used to generate colour displays more easily and intuitively interpretable than those produced by more conventional display methods.

5.1.2 The inherent depiction of image spatial structure

In image data depicting radiance from a surface, the underlying topographic structure of the surface is generally evident. This is because of the dependence of the measured radiance on surface orientation; the visual system is able to extract topography, and hence overall structure, from these variations in measured radiance. While variations in surface albedo (reflectance/incidence ratio) also affect the overall lightness, in general the continuity and smoothness of local variations not only allows the visual system to extract the topography, but also derive information about surface cover, from such an image.

The fact that surface-orientation-dependent lightness variations are largely independent of wavelength means that the surface topography in images derived from a sensor with, say, a spectral response well into the infra-red region, can be appreciated as well as that in images derived from sensors in the visual-system response range. This applies to both passive and active sensor systems, provided that the sensors detect the radiance from the surface, rather than an emission from the surface or from below it. Thus, for example, in images derived from detecting reflected radar, the surface structure is apparent (see, for example, Elachi, 1982; Daily, 1983).

Consequently the surface structure is inherently embedded in most images derived by sensing radiation reflected from the earth's surface. Provided that chosen display techniques preserve this inherent scene structure information, the structural comprehension requirements for the display will be satisfied.

5.1.3 Depiction of surface covering - the problem of spectral assignment

Choice of an appropriate spectral representation for remotely sensed data will depend on the data and its interpretation. One rationale for representation is that it should appear in image form as close as can be achieved to the way it would appear to a human observer; this can be termed a "true colour" representation. Pratt (1978) and Horn (1984) treat the sensor system and display system requirements for exact reproduction of a colour representation. However, sensor systems responsive over a greater spectral range, or with finer spectral resolution, than that of the visual system can detect information not normally observed visually, and to restrict a representation in such a manner would waste available information. In addition, a sensor system may not have the appropriate range or bandwidths to adequately cover the visual range.

As a result, displays often approximate a "true colour" representation; the value of such an approximation is that some degree of familiarity of representation can aid the intuitive appreciation process, but the drawback is that such a representation can be misleading. This suggests the importance of considering in the general case the nature of the data, and the perceptual processes to be used to interpret it; we treat this problem in this chapter.

5.2 Mapping from data to colour space

This section considers rationales for the mapping of multi-spectral data into UCS. These are based on consideration of the data nature, and the perception of it by the visual system. Mapping methods are developed, together with constraints imposed by data-dependent or perceptual considerations.

5.2.1 Generation of informative data variables

Important to the process of interpretation is that the data be presented in a form desired by, or useful to, the interpreter; in other words, that informative data variables be depicted appropriately in colour. Ideally these informative data variables are generated by modelling the physical processes causing and affecting the detected signal. This can involve not only the modelling of radiation reflection from the materials being studied (see Volume II of Colwell, 1983 for a full treatment of this topic), but also the characterisation of, and compensation for, atmospheric and sensing equipment effects (see Chapter 5, Volume I of Colwell, 1983). Alternatively, if no such model can be confidently built, an approach which derives variables that are significant in a statistically-based information sense can be used. The choice of methods to generate informative data variables lies in the realm of the analyst. In the following sections we shall consider the placing in UCS of informative data variables generated by various approaches.

5.2.2 Representation of informative data variables

Consistent with the approach outlined in Chapter 2, we first consider appropriate perceptual representations for informative data variables. Most important is that the chosen perceptual representation should aid the process of interpretation by making use of the normal scene analysing capabilities of the human visual system. It is consequently appropriate to use representations as natural as possible; for example, representing overall reflectance by lightness, or gradual changes by variations in saturation, or sharply defined changes by combined hue and lightness boundaries; in order to produce a scene structurally similar to one which might be seen in the real world. This is recognised by Daily (1983) who uses the spatial-frequency characteristics of data derived from reflected radar measurements to determine display representation; high frequency variations (as indicative of geological structure) are mapped into lightness, and low frequency variations into a geometrical approximation to chromatic components. While this type of approach facilitates scene comprehension, the danger that misinterpretation might occur must also be borne in mind. Examples of this real-world type of representation are shown in section 5.3.1.

Another type of familiar representation, and hence one which the visual system has some experience in coping with, is the conventional false colour type of display commonly used in satellite image presentation. This consists of placing up to three spectral bands on the blue, green and red guns of a display device, generally with the shortest wavelength band on blue, and the longest on red. Despite the lack of representation of significant data variables by appropriate perceptual attributes, products generated by this method have become familiar over time, and hence more interpretable than arbitrary assignments might be. Such products suffer the disadvantages cited in Chapter 3; namely, data differences are not uniformly reflected by colour differences, and the results are display-device dependent; but if a mapping of the data into UCS is made, these drawbacks can be overcome. Examples of this approach are given in section 5.3.2.

5.2.3 Mapping informative data variables into uniform colour space

It was pointed out in Chapter 3 that perceptual attributes hue, saturation and lightness do not form a three-dimensional Euclidean space; consequently it is not possible to map three informative data variables into HSL space in a linear and orthogonal manner. Instead, we shall consider the mapping of informative data variables into UCS, and note the resulting perceptual restrictions when treating data sets of various dimensionalities.

Mapping constraints - the dimensionality and distributions of real data sets

In mapping informative data variables into UCS, we follow the basic principles of spectral assignment proposed in Chapter 2. These are that variables should be represented by perceptually significant colour attributes, and that perceived colour differences should be proportional to numeric data differences. Good utilisation of colour volume will maximise the discernibility of these differences. The range of the mapping is the gamut of colours which can be produced by any particular display device, and is consequently a bounded volume lying within the three-dimensional UCS accessible by the human observer. This bounded volume is usually largely convex. For uniqueness of representation of informative data variables in colour, the mapping should *one-to-one*, limiting its domain in data space to be at most three-dimensional. If the entire colour gamut of the display device is to be accessible, the mapping should also be *onto*.

These principles suggest that the mapping be linear, and to optimise discernibility between variables, orthogonal. However, in practice it may be desirable to relax these constraints for several reasons. First, it is reasonable to assume that in general, an informative data variable derived from remotely sensed data will have a distribution which reflects a physical property in some way. This distribution is thus unlikely to be naturally bounded, and may well be irregular in shape. There then arises the question of whether the linearity condition should be relaxed to allow, for example, a compression of the distribution tails into the outer shell of colour volume, or whether linearity should be maintained and *one-to-oneness* sacrificed by mapping the data outside the domain to, for example, the colour volume boundary. This problem of dealing with what we term saturated pixels (pixels which, when transformed, give rise to device-addressing values outside the physically utilised range) is discussed more specifically in section 5.4.2.

The second constraint which might be in some way relaxed is that of orthogonality of informative data variable axes within UCS. In practice, uniform colour spaces approximate uniformity, and the influence of factors such as induction can affect locally judged colour differences quite appreciably. It might well be desirable to achieve a natural or well established association of particular colours or colour ranges with physically significant informative data variables for reasons of interpretation. Mapping two such variables in UCS might not be possible under the orthogonality constraint, and it may be more important to give these variables some chosen perceptual representation than to maintain orthogonality. Further, if informative data variables have distributions which are particularly sparse in one area of colour space, and discernibility in that area is of little interest to an analyst, it might be advantageous to further separate the variable axes in more densely occupied areas of colour space. It becomes apparent that choosing the mapping involves consideration of the distribution and nature of the informative data variables, and of the dimensionality of the data space.

One-dimensional case

The mapping of one informative data variable into a line, or one-dimensional subspace, (to maintain *one-to-oneness*) of UCS gives a display commonly known as pseudo-colour. Linearity requires that this line be straight, but as in the case of univariate map sequences described in Chapter 4, the use of carefully chosen curved lines can give satisfactory data representation with a wider colour range than can be achieved with a straight line. The same smoothness and low curvature requirements apply, although because of the presence of distribution tails, more care must be taken to keep the ends of the chosen line distinct.

Scaling of the informative data variable along the chosen line should be performed to achieve fullest utilisation of the colour range of the line, and hence maximum discernibility of numerical data differences, while considering the loss of information in the distribution tails due to saturation. Total and cumulative histograms are useful for determining the extent and shape of the data distribution, and failing any physically based rationale for determining exact colour assignment to particular values and hence scaling or positional factors along the line, histogram means or percentile values can be used.

Two-dimensional case

Two informative data variables can be mapped into a plane in UCS with the aid of two-dimensional histograms, or scatter-plots, of the data values, using knowledge of the extent of the device gamut in the chosen plane of UCS. Provided that the main extent of the scatter-plot lies within the boundary of the chosen plane, saturation will be low. Plots showing display device volume cross-sections (figures 3.5 and 3.7), or colour images of these cross-sections (figure 3.9), are useful for this purpose.

While any plane in UCS can be used, planes with the fullest variation in lightness (that is, including the lightness axis) are attractive because of the importance in perceptual terms of lightness. In addition the complementary hues on either side of the lightness (or grey) axis accentuate any natural division in the data if that division point is placed on the lightness axis. To include a wide hue range, and also some lightness variation, a plane including the three display device primary colours can be used (see figure 3.9(c)).

As with the one-dimensional case, scaling in each direction is likely to be dependent on placement in UCS due to the physical significance of data values and their desired colour representations, and on the importance of discriminating between outlying points on the distribution which might be saturated under maximal expansion. Since the UCS colour volume accessible by most display devices is not symmetrical, the particular plane most suitable for use might be determined on the basis of the fullest utilisation of colour space as determined by the shape similarities between the two-dimensional data histogram and any particular colour plane. The significance of the relationship between the data variables might determine whether scaling factors for the two variables should be kept the same.

Three-dimensional case

For mapping a volume in informative data space into a volume in colour space, the conditions of linearity and orthogonality are particularly important. It is unlikely that chosen colour assignments for all three informative data variables will be in keeping with orthogonality of variable placement; in practice the more dominant of the variables should be aligned as intuitively suggested.

Deciding on the placement and scaling of variables in three dimensions can be a problem as there is no easy representative analogue of the one- or two-dimensional histograms. One possible rationale for specifying the mapping might be to place the centroid of the data space distribution at the centroid of the display device colour volume, and find the rotation and scaling factors to give maximal colour volume utilisation while keeping the proportion of saturated points below some specified level. This could be done under specific constraints, such as a fixed orientation for a particular axis. As in the two-dimensional case, scaling can be consistent for each variable, or differential to allow greater use of colour volume.

However, practical results suggest that fullest possible utilisation of colour volume is not necessarily the most important consideration when interpreting three-dimensional data; more important is the ability to associate perceptually meaningful colour attributes, or specific perceptually discernible colour ranges, with particular data characteristics. Bearing in mind the computational complexity of optimal volume-in-volume fitting, an approach guided more by considerations of assigning appropriate colour ranges to each variable was taken.

5.2.4 Derivation of the required mapping

As a result of mapping various two- and three-dimensional data sets into planes and volumes of UCS, a flexible technique for specifying, and hence generating, the mapping from informative data space to UCS, was developed. This technique involves the specification of the position and the direction of each informative data variable axis in UCS, and generates the resulting linear transformation. Orthogonality need not be maintained. Axial specification is made by means of any two points in both informative data space and UCS, from which positional and directional information can be derived.

We wish to transform L variables from an M-dimensional data space $\mathbf{D}$, to an N-dimensional colour space $\mathbf{C}$. The alignment of a data variable in $\mathbf{C}$ can be determined by specifying the points in $\mathbf{C}$ into which each of two points which define the variable axis in $\mathbf{D}$ are to be mapped. If the data variables are in fact the axes of data space (that is, the transform from data space to informative data space has been performed), this amounts to defining the alignment of the data space axes in $\mathbf{C}$. However, in the general case, we may wish to define the mapping directly from raw data, and consequently we develop the method to encompass this.

In practice, we will wish to specify points in the data distributions which are physically significant, such as particular percentile points of cumulative histograms. Consequently the origin in $\mathbf{D}$ will not necessarily be a physically significant point, and its location in $\mathbf{C}$ may not be specified. Specifying two points for each variable axis means that the transform is overspecified, and further, may be inconsistently so. We hence need a rationale for resolving conflicting specifications arising from possibly arbitrary axial point specifications. Because of the importance of maintaining the directional alignment of variable axes as specified, we give alignment priority, and allow displacement of each axis in the plane orthogonal to its alignment.

We require a linear transform which maps a vector $\mathbf{x}$ in an M-dimensional data space into a vector $\mathbf{y}$ in an N-dimensional colour space. Such a transform has the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b},$$

where $\mathbf{A}$ is an NxM -dimensional transform matrix and $\mathbf{b}$ is an N-dimensional bias vector.

To find $\mathbf{A}$, we consider a set of L pairs of vectors ($L \leqslant M, L \leqslant N$), each of which satisfies the transform $\mathbf{y} = \mathbf{A}\mathbf{x}$. The directional vectors derived from the axial specification points form such a set. If $\mathbf{x}_l^1$ and $\mathbf{x}_l^2$ are the vectors defining the points in $\mathbf{D}$ on each axis l which are to be mapped into points in $\mathbf{C}$ defined by vectors $\mathbf{y}_l^1$ and $\mathbf{y}_l^2$, the directional vectors are given by

$$\mathbf{x}_l^d = \mathbf{x}_l^2 - \mathbf{x}_l^1$$

$$\mathbf{y}_l^d = \mathbf{y}_l^2 - \mathbf{y}_l^1,$$

and we have

$$\mathbf{y}_l^d = \mathbf{A}\mathbf{x}_l^d$$

since the bias has now been removed.

Let $\mathbf{X}$ be the set of L vectors $\{\mathbf{x}_l^d\}$, and $\mathbf{Y}$ be the set of L vectors $\{\mathbf{y}_l^d\}$, represented for convenience as column vectors of a matrix. Then we have

$$\mathbf{Y} = \mathbf{A}\mathbf{X},$$

where $\mathbf{Y}$ is an NxL -dimensional matrix, $\mathbf{A}$ is as before an NxM -dimensional matrix, and $\mathbf{X}$ is an MxL -dimensional matrix. $\mathbf{A}$ can then be derived from

$$\mathbf{A} = \mathbf{Y}\mathbf{X}^{-1},$$

where $\mathbf{X}^{-1}$ is the inverse of $\mathbf{X}$. This depends on $\mathbf{X}$ being invertible; if $L < M$, $\mathbf{X}$ can be augmented by adding linearly independent column vectors and appropriately re-building after inversion (the columns of $\mathbf{X}$ will only be linearly dependent if alignment of an axis has been multiply specified).

To find the bias vector **b**, we apply the earlier stated rationale which gives axial alignment precedence over translation. Under this condition, the zero value of each defined variable axis must lie on an $(N-1)$-dimensional hyperplane in **C** which is orthogonal to the variable axis. A hyperplane in N-dimensional space has an equation given by

$$\mathbf{w}.\mathbf{z} + \omega = 0,$$

where $\mathbf{w} = (w_1,....,w_n)$ is an N-dimensional weight vector orthogonal to the hyperplane, and $\mathbf{z} = (z_1,....,z_n)$ is a vector in N-dimensional space. We can find the constant term ω by noting that the components in **C** of the vector $\mathbf{y}_l^0$, defining the point of intersection of each variable axis l with its orthogonal hyperplane, are given by

$$y_{i,l}^0 = y_{i,l}^1 - x_l^1\left(\frac{y_{i,l}^2 - y_{i,l}^1}{x_l^2 - x_l^1}\right), \text{ for } i = 1,N.$$

We then have a point $\mathbf{y}_l^0$ on the hyperplane, and its orthogonal directional vector $\mathbf{y}_l^d$. Substituting into the equation of the hyperplane, we can obtain the constant term ω from

$$\omega_l = -(\mathbf{y}_l^d . \mathbf{y}_l^0),$$

and the equation of the hyperplane orthogonal to each variable axis l, and containing its zero value, is then determined. We then find the intersection of L $(N-1)$-dimensional hyperplanes in **C** to obtain the transform bias vector **b** (and hence the origin of the data space co-ordinate system in **C**).

If $L = N$, this gives us a unique solution from

$$\mathbf{b} = -\mathbf{Y}^{-1}\omega$$

where as before **Y** is the matrix of L column vectors $\{\mathbf{y}_l^d\}$, and ω is the L-dimensional vector $(\omega_1,....,\omega_L)$. The solution depends on **Y** being invertible, which in turn requires that no two variable axes are co-aligned in **C**.

If $L < N$, we have the solution being $(N-L)$-dimensional (assuming we invert **Y** by augmenting with linearly independent vectors as before), and some further constraint is required to choose a single point **b** in this $(N-L)$-dimensional space. We can use a minimum distance criterion to find the **b** which minimises (for example in a least-squares sense) the distance from $\mathbf{b}$ to each variable axis defined. In practice, however, this need not be done, since $\mathbf{C}$ is at most three-dimensional (that is, $N \leqslant 3$). If $L = 1$, the choice of origin is straightforward, and taken as the zero value of the specified single variable axis. If $L = 2$, an appropriate point can be found by taking the line which is perpendicular both to the line defined by the hyperplane intersection and to the line joining $\mathbf{y}_1{}^0$ and $\mathbf{y}_2{}^0$, and defining $\mathbf{b}$ to be at the point where this line joins the intersection of the hyperplanes.

A final situation, where $L > N$, can arise. In this case perceptual associations between variables and UCS dimensions will be lost, or partially so. However, to allow for such a mapping if required, we note that using L constraints to establish the bias vector results in over, and possible conflicting, specification. Consequently we simply choose any set of N variables (such as the first N) and use them to derive the bias vector. Note that derivation of the alignment matrix is not affected by the over-specification problem.

The advantage of specifying each variable alignment by two points in each of data and colour spaces is that it allows the depiction of any physically significant values as particular colours while also allowing colour range specification. A set of cross-sections of the available colour volume of any particular display device shows the range in any direction in colour space, while the extent of each informative data variable can be taken from appropriate points on its one-dimensional histogram.

The success in practice of this approach depends on two factors: first, that the colour volume is largely convex and reasonably smooth; and second, on the shape of the informative data variable distributions. If the distributions are more or less physical in nature (as ideally they should be if the variables represent a physical quantity) and tail off towards the edge in something approaching a normal or Gaussian manner, excessive saturation should not occur. However, an artificially produced distribution might be very non-physical in nature (for example, three equi-probable distributions will produce a cubical three-dimensional distribution), and generate a poor fit within a particular display device colour volume, giving rise to substantial saturation problems.

5.3 Application to real data sets

In this section we consider the mapping in UCS of first, informative data variables which are to be displayed in a controlled and closely specified way; and second, remotely sensed data sets which under conventional display methods generate products with poor chromatic contrast.

5.3.1 Representation of model-based physical variables in uniform colour space

Use of multi-spectral data gathered by satellite or airbone scanner is widespread in the ecological survey of large areas of land. In one such study of an area in the Broken Hill region of Australia (Graetz et al., 1982), informative data variables representing the degree of vegetative ground cover (*cover*) and ephemeral vegetative growth or greenness (*greenness*) were derived from Landsat MSS data, using calibrating ground-truth radiometric measurements on control sites. Five images were radiometrically normalised in this manner to study the change in each of these variables over a period of time. *Cover* and *greenness* were thus the informative data variables to be placed in UCS. A colour assignment which would allow appreciation of gradual or subtle changes in each variable, while at the same time allowing separation of the two variables, was required. Consequently the perceptual attributes lightness and saturation were chosen to represent *cover* (inversely) and *greenness* respectively. Choosing lightness to represent *cover* is in keeping with the normal scene viewing situation in which not only surface orientation, but also surface covering, affects the overall reflectance and hence perceived lightness. Saturation, and hence *greenness*, was realised within a cyan-red UCS gamut cross-section, in order to be consistent with the conventional false colour representation of ephemeral growth (characterized by strong infra-red region reflectance and hence MSS band 7 response) in red. This cross-section is shown in figure 5.1, together with the graphic outline of the two-dimensional histograms, or scatter-plots, of each of the five data sets. The scatter-plots are outlined at the 8% (of the maximum bin-count) level to show the shape of the bulk of the data; outlying points are hence not depicted. Placement was determined to keep no more than

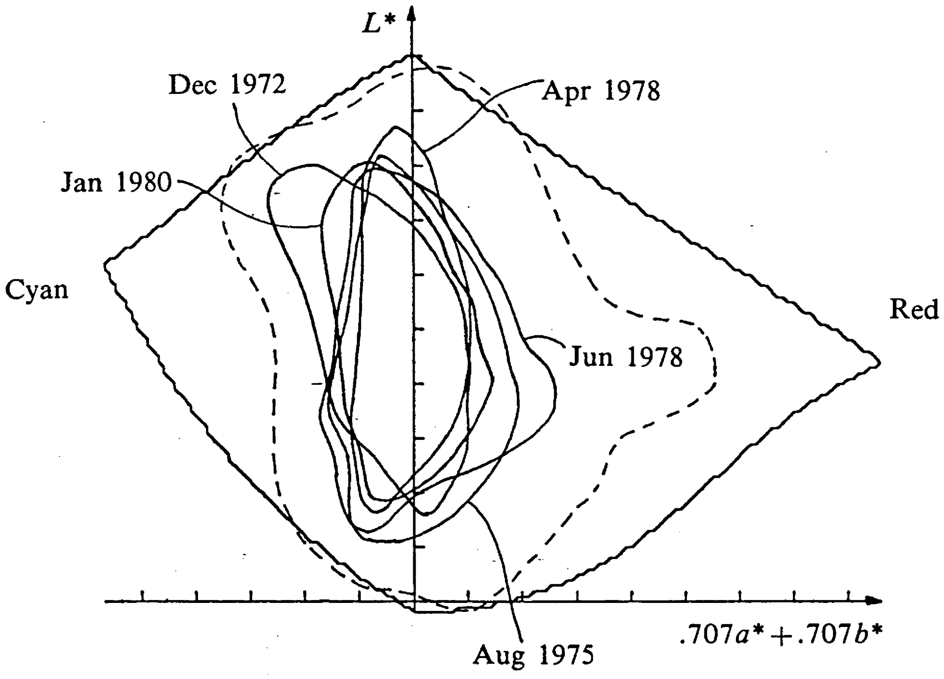


Figure 5.1 Scatter-plots in a constant-hue gamut cross-section. Scatter-plots of five normalised data sets representing Landsat-derived informative data variables *cover* and *greenness* are shown in a cyan-red gamut cross-section of the film/print gamut. Scatter-plot boundaries are defined at the 8% level.

1% of the points in any image saturated; effectively, the outer boundary of the union of the scatter-plots, contoured at a lower level (2% of the maximum bin-count), was kept within the chosen gamut cross-section. This union boundary is shown in the broken line in figure 5.1.

In the more general case of a single image, this placement can usually be achieved simply by using specified percentile points of the one-dimensional histograms, and placing them at appropriate points in the colour gamut cross-section. This approach relies on assumptions on the individual and joint distributions of the data variables, which can be invalid for data variables with individual distributions which are highly non-Gaussian in nature, or when data sets are highly correlated. Consequently, not only the data channel distributions, but also their correlations, must be examined before specifying placement and scaling factors. In practice, for typical Landsat products it was found that in the film/print gamut, stretching the 2% points on the one-dimensional histograms of the data variables between points approximately 0.6 of the distance from the centre of the grey axis out to the gamut boundary generates well balanced imagery with less than 1% of the image points saturated. As noted above, this is a rule-of-thumb only, and will depend not only on data characteristics, but also on gamut shapes.

The full temporal set of normalised *cover-greenness* images is shown in the upper half of figure 5.2. The lower half of figure 5.2 shows a set of images designed to show the change in *cover* over the period of time under study; in this case the degree of saturation represents the extent of the change. The variable actually mapped into saturation was the difference between the final scene (1980) *cover* and that of each other year; the hue boundary thus indicates clearly whether the cover in 1980 was better (magenta) or worse (green) than in each other year. This positioning is achieved simply by specifying the equal value point (in the data difference channel) to lie on the lightness axis in UCS. The final image in the sequence is the reference *cover* image for 1980 which, in each other image in the *cover/change-in-cover* set, is placed in the lightness direction to give the images spatial form. In both temporal sequences each individual image has a limited lightness and chromatic range, since the union of the five data sets was used to determine expansion.

Image descriptions for the following 3 colour plates

Figure 5.2 Broken Hill temporal sequences - informative data variables mapped in UCS for interpretations of gradations in data property.

In the upper set of images, a temporal sequence of five data sets depicts the degree of vegetative ground *cover* and ephemeral *greenness*, in a Landsat subscene of the Broken Hill region of Australia. *Cover* is represented by lightness; dark areas indicate good ground *cover* and light areas indicate poor *cover*. *Greenness* is represented by saturation in a two-hue range from cyan, through grey, to red; areas with high *greenness* appear a saturated red.

In the lower set of images, the change in ground *cover* over the period of the study is depicted by saturation, in a two-hue range from green, through grey, to magenta. Magenta hues indicate that the *cover* in 1980 was better than in the year with which it is being compared; green hues indicate the converse. In each of the images, the *cover* in 1980 is displayed in lightness.

Figure 5.3 Upper set of images: Example of chromatic expansion of correlated data with low dynamic range by mapping principal components orthogonally in UCS.

The data is from a Landsat MSS scene of the Broken Hill region in Australia, acquired in August 1975.

(a) A conventional false colour representation of bands 4, 5, and 7 maximally expanded on blue, green, and red guns respectively.

(b) A composite of principal components 1, 2, and 3 maximally expanded on green, red, and blue guns respectively.

(c) Principal components 1, 2, and 3 mapped orthogonally and maximally expanded in UCS.

(d) The effect of smoothing low order principal components: in the upper left sub-image no smoothing was applied; in the upper right sub-image, PC3 was smoothed using a $3x3$ box filter; in the lower left sub-image, PC2 and PC3 were smoothed using $3x3$ box filters; in the lower right sub-image, PC2 and PC3 were smoothed using $3x3$ and $5x5$ box filters respectively; (see text, section 5.4.2).

Figure 5.3 Lower set of images: Example of proportional data variables mapped in UCS for maximum chromatic contrast

The data was acquired in an aerial radiometric survey of the Broken Hill region of Australia. Proportional variables were derived from radiometric counts representing emissions from uranium, thorium, and potassium radio-isotopes.

(a) Relative proportions of uranium and thorium radio-isotope counts are represented by saturation in a constant-hue red-cyan cross-section (red - high uranium count; cyan - high thorium count). The total relative radio-isotope count is displayed in lightness (see figure 5.7(a)).

(b) As for (a), but for thorium and potassium in a green-blue cross-section.

(c) As for (a), but for potassium and uranium in a blue-orange cross-section.

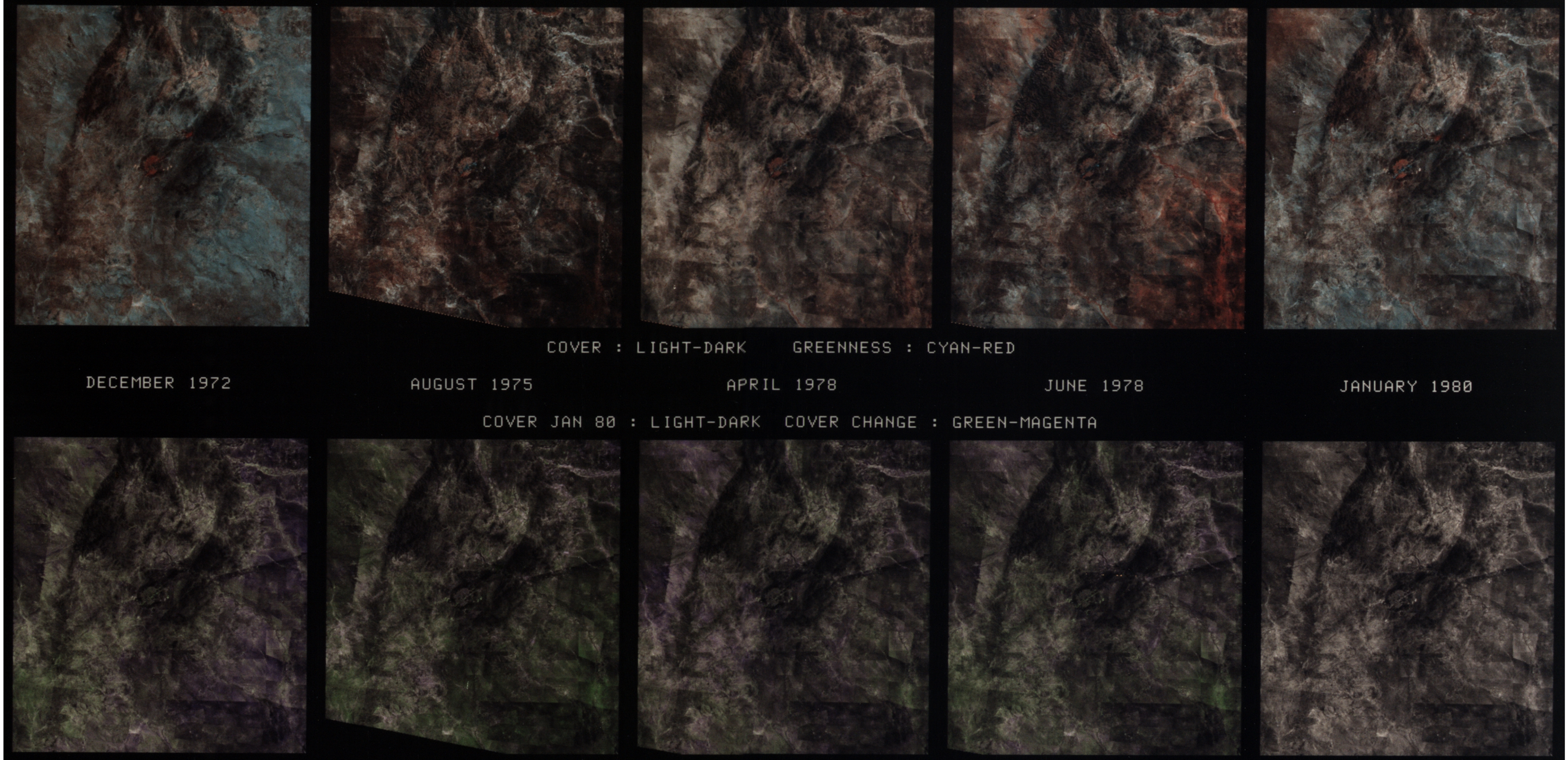
(d) Relative proportions of each of three variables (uranium-red; thorium-green; potassium-blue) mapped into saturation $2\pi/3$ apart. The total count is mapped into lightness (see figure 5.7(b)).

Figure 5.4 Simpson Desert mosaic - chromatic expansion of correlated multi-modal data.

(a) A conventional false colour composite of bands 4, 5, and 7 maximally expanded on blue, green and red guns respectively.

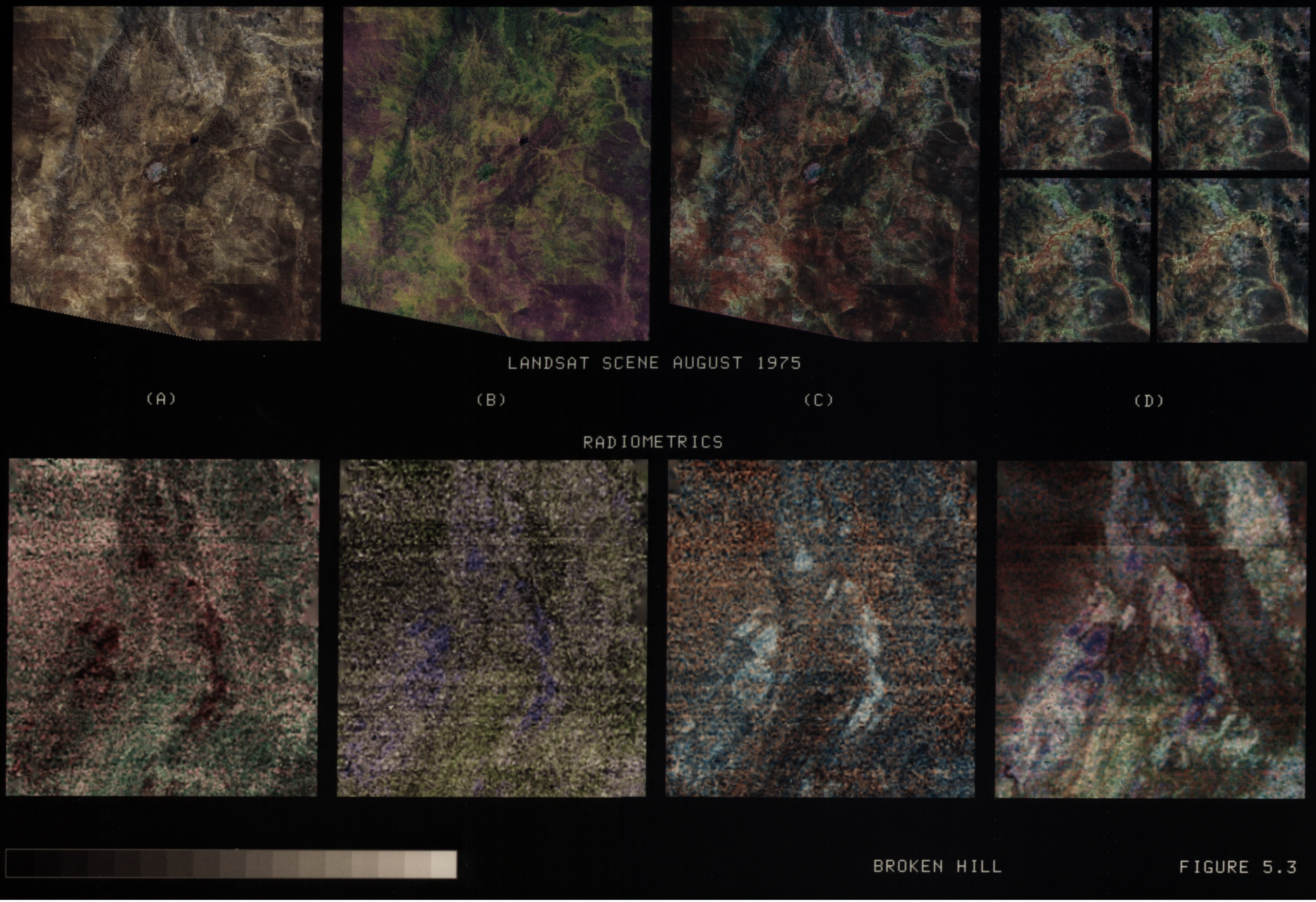
(b) Principal components 1, 2, and 3 mapped orthogonally and maximally expanded in UCS.

(c) Detailed illustration of the improvement in chromatic contrast achieved.


COVER : LIGHT-DARK   GREENNESS : CYAN-RED
DECEMBER 1972
AUGUST 1975
APRIL 1978
JUNE 1978
JANUARY 1980
COVER JAN 80 : LIGHT-DARK COVER CHANGE : GREEN-MAGENTA

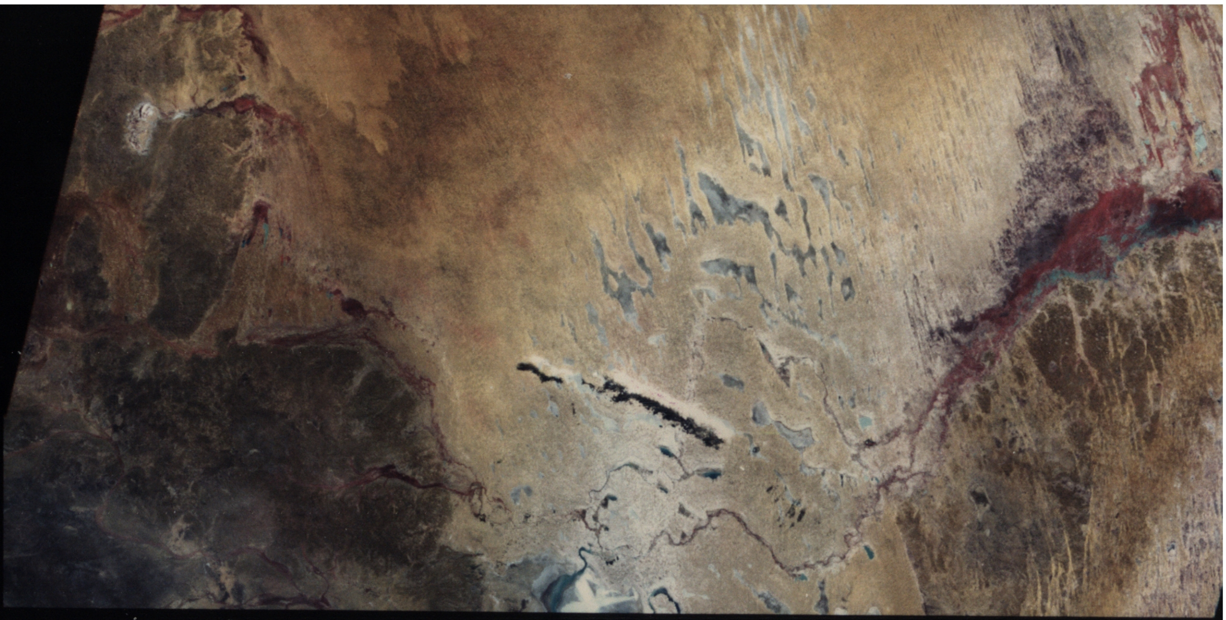
BROKEN HILL TEMPORAL SEQUENCE

FIGURE 5.2

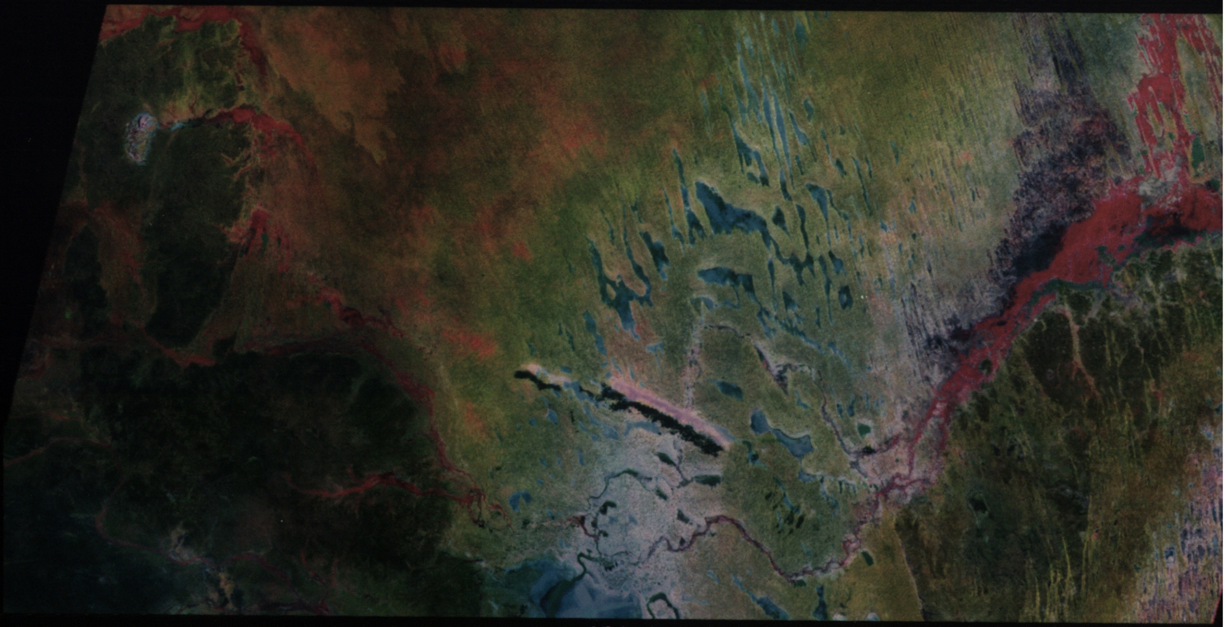


BROKEN HILL

FIGURE 5.3



(A)



(B)

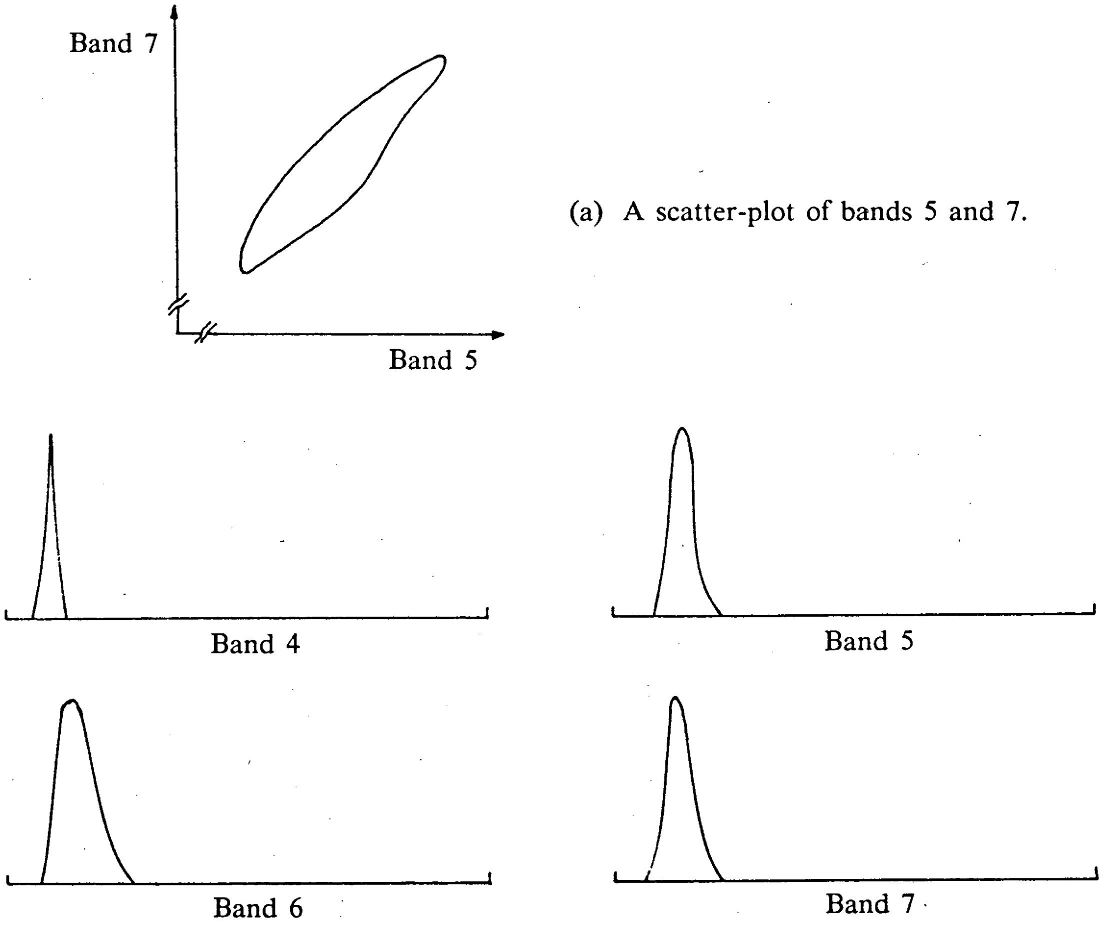
The sensitivity of the human visual system in discriminating opponent hues can also be exploited as a form of classification aid in imagery in which a distinction is to be made between two close spectral specifications. In conventional display it is not generally possible to cause a particular line in the data spectral space to be of a neutral hue, with opponent hues either side. In UCS however, such a specification is straightforward.

This technique was found particularly useful for resolving salt-affected soils from bare soils; often the spectral signatures derived from training sites are very close for these types of soil. Use of this technique resulted in the required distinction on the basis of hue, while appropriate placement of the lightness axis allowed discrimination between vegetated and bare salt-affected areas.

In each of the applications described in this section, the informative data variables have not only been represented by appropriately chosen perceptually defined colour ranges, but also such that gradations in colour have uniformly corresponded to gradations in ground cover characteristics; these aspects of the representation can be crucial for detailed interpretation of such data.

5.3.2 Representation of statistically decorrelated imagery - chromatic contrast enhancement

A common problem when displaying Landsat or other similarly derived data sets is that if the data channels are highly correlated, the conventional approach of assigning a channel to each of three colour guns of a display device produces an image with poor chromatic contrast. This is illustrated in figure 5.5(a), which shows a two-dimensional scatter-plot of two channels from a Landsat subscene of the Broken Hill region acquired in 1975. Table 5.1 shows the band correlations for this scene; correlation between bands 5, 6 and 7 are very high. This is typical of imagery of arid regions with dry, flat, uniform vegetation; not only is the spectral variation small, but the overall dynamic range is also very low (as can be seen from the spectral band standard deviations given in table 5.1). Figure 5.5(b) shows the one-dimensional channel histograms of this image; the proportion of the full 256-level range occupied by each histogram is an indication of the dynamic



(a) A scatter-plot of bands 5 and 7.

(b) Histograms of the four spectral bands. Dynamic range, and hence signal-to-noise ratio, is indicated by the radiometric range encompassed by the data.

Figure 5.5 Histograms of correlated spectral bands with low dynamic range from Landsat MSS data of the Broken Hill region in Australia.

	Band 4	Band 5	Band 6	Band 7
Mean	24.08	37.07	39.40	30.90
Standard deviation	3.68	7.51	9.01	7.52

Correlation matrix				
	Band 4	Band 5	Band 6	Band 7
Band 4	1.00			
Band 5	0.81	1.00		
Band 6	0.76	0.96	1.00	
Band 7	0.75	0.92	0.95	1.00

Table 5.1 Spectral band statistics and correlations for a Landsat MSS subscene of the Broken Hill region in Australia.

range in the data (and hence the signal-to-noise ratio) since the level quantisation is of the order of the radiometric resolution. Consequently radiometric expansion of such data does not increase the effective dynamic range. The result is that a three-dimensional histogram of any three of the four bands is sausage-shaped. Fitting such a histogram into a BGR cube by the conventional method of stretching each band over the numerical range for each colour gun results in the use of a small portion only of colour space. For Landsat data there is an added disadvantage that one MSS band cannot be displayed.

Statistically-based processes such as transformations to principal components have been widely used to produce a set of uncorrelated data channels from any number of input channels. These uncorrelated channels can be ordered according to their "information" content, the usual measure being statistical variance. Such transforms are used for image compression on the basis of statistically-measured "information" content, rather than some perceptually-based measure; the implicit assumption in their use is that failing any other rationale, the statistical measure is a reasonable one to use. A principal component analysis (PCA) finds the linear combination of the input channels which has the greatest variance, terming it the first principal component (PC1), and subsequent components are the combinations with greatest variance subject to the condition of being orthogonal to previously defined components. These combinations are derived from the eigenvectors of the data set covariance matrix. (See Huang (1975), Pratt (1978) or Rosenfeld and Kak (1982) for further details on orthogonal image transforms and their bases.) In fact a PCA is equivalent to a Karhunen-Loeve transform, with a shift of co-ordinate origin included (Gerbrands, 1981).

Thus if the channels of an image are correlated, statistical decorrelation techniques will generally result in an effective reduction in data dimensionality, lower order components portraying progressively less significant "information" (based on a measure of variance), and appearing largely as noise. On this basis, then, when applied to four band Landsat data, the lowest order component can be discarded. Display of the first three components appropriately expanded then results in an image with improved colour utilisation. (In fact there are image noise problems associated with highly expanding low-order principal components; this is treated in section 5.4.3.) However, there is still a problem in displaying such an image, as presenting the first three principal components on the colour guns of a display device results in an image not only very different from a natural scene, but also quite different from conventional false colour products which have become familiar to interpreters of satellite imagery. In addition, such an image no longer has the surface topography depicted by variations in lightness as occurs in false colour Landsat composites, resulting in the scene structure not being obviously depicted. The upper set of images in figure 5.3 show various representations of the Broken Hill Landsat subscene. In 5.3(a) a conventional false colour composite of bands 4, 5 and 7 is shown, while in 5.3(b) a composite of principal components 1, 2 and 3 of the four band image is displayed on green, red and blue guns respectively. Incorporation of the fourth band, and some colour expansion, are achieved at the expense of both familiarity of representation, and clarity of surface topography depiction; interpretation is not intuitive.

Juday (1978,1979) suggests placing the principal components of an image in UCS in order to optimise colour expansion; in an earlier work Taylor (1974) proposes a similar approach. These approaches concentrate on optimising colour contrast, and achieving representation of the magnitude of data differences by the size of perceived colour differences. Interpretation can consequently be difficult, particularly since it is not obvious to an analyst what the principal components represent, in terms of real-world or even original-data-channel properties, even if some attempt at colour separation can be made. Kaneko (1978) takes this approach one step further by recognising the physical significance of the first principal component as indicating overall reflectance, and mapping this component into lightness in UCS. By placing the principal components of an image in UCS in a way which considers relevant real-world data associations, the resulting image can be more easily interpreted. We extend this approach by further considering the perceptual processes used to interpret images, and also consider experience-based representations. This should produce imagery which not only has the advantages gained by placing maximally expanded principal components in UCS, but also is substantially easier to interpret intuitively.

We consequently display decorrelated components by developing rationales which are based both on the use of perceptual processes for analysing scenes, and on maintaining some consistency with conventional false colour products, thus compromising between the various conflicting display requirements. Due to the closeness of the MSS spectral bands, the first principal component (PC1) represents characteristics common to each band; this is largely the overall nett radiance which, as described earlier, is not highly wavelength dependent in the visual and near-infra-red regions. Alignment of PC1 with the lightness axis, or close to it, gives a natural representation of this overall reflectance. In qualitative terms, the second principal component (PC2) generally indicates the most obvious distinctions between the data channels. In practice this is often the difference between long and short wavelength reflectances, and if these difference can be physically explained, a perceptual association with the physical effect could provide a rationale for the placement of PC2, orthogonally to PC1. Alignment of PC3 with the third orthogonal axis is then defined, and the sign can be decided again on physical considerations. For example, in vegetated areas the dominant wavelength-dependent component is often the result of the long wavelength reflectance of high chlorophyll plant life, suggesting its placement in a direction in UCS corresponding to a conventional false colour red representation of long wavelength reflectance. Such an approach can be successful if the general nature of the data is known, but if it is not, a more quantitive and less heuristic approach is required.

A quantitative approach can be based on using the matrix showing the transformation from the original data channels (MSS bands) to principal components. The alignment of PC1 can be derived by taking the vector sum of the vector directions of the display device blue, green and red guns in UCS, each weighted according to the formation of PC1 from each of MSS bands 4, 5, and either 6 or 7 respectively. This rationale is based on the conventional representation of maximally expanded bands 4, 5, and 6 or 7 on blue, green and red guns respectively of display devices. Such an alignment will in fact generally be close to the UCS grey axis. PC2 can be aligned similarly, but with the added constraint that it be placed orthogonal to PC1. Transforming to an axial system based around the chosen direction for PC1 then gives the component of the appropriately weighted sum orthogonal to it, and PC2 is then so aligned. The alignment of PC3 is then determined as being orthogonal to PC1 and PC2, but its direction or sign can be determined by again working out the appropriate weighted sum and choosing the positive direction. In practice, it was found most convenient to actually align PC1 with the lightness axis; PC2 and PC3 are then aligned in constant lightness directions, making intuitive and computational specification of their alignments more straightforward.

Clearly if orthogonality were not imposed, and the axes were aligned in UCS exactly as determined by their formations from the original bands, little would be gained (in fact, only the inclusion of the excluded band and uniformity of representation of numerical data differences by colour differences). However, imposing orthogonality of placement of the principal components means that colour contrast is substantially increased. Hence false colour representation is not closely achieved, but rather its main characteristics are. Bearing in mind the substantial variations in false colour representation due to widely varying scene properties, this is not likely to be crucial to interpretation. In practice, products generated by this technique show substantial improvement over conventionally generated products and have been interpreted quite satisfactorily. Figure 5.3(c) (upper set) shows the result of treating the Landsat subscene of the Broken Hill area in this manner; the components were maximally expanded in UCS to keep less than 1% of the points in the image saturated. The increased colour contrast in this image, over that in the conventionally displayed image, can be seen clearly. The image also retains the essential character of the conventionally displayed false colour image, unlike the false colour composite of principal components in 5.3(b). There is also an additional advantage in that striping (a known artifact in Landsat MSS data) in the scene is less evident. The striping seen in 5.3(b) (upper) is most obvious in PC3, but also appears in PC2; as a result it appears as a blue-red colour and is visible because it has a significant lightness component. The human visual system is extremely sensitive in detecting even small variations in lightness. In 5.3(c) (upper), PC2 and PC3 are mapped into constant lightness directions, and as the visual system is less sensitive to small chromatic variations (see Chapter 7), the striping is much less evident.

When a scene has widely diverse characteristics resulting in a very large overall dynamic range, good colour space utilisation can be extremely important if local detail is to be well separated. This type of variation can be encountered in mosaics of Landsat scenes, where the area covered becomes so large that very different physical features appear on the image; the histograms of such a composite are generally strongly multi-modal. Figure 5.6 shows the one-dimensional channel histograms of a re-sampled mosaic of portions of 6 Landsat scenes covering the Simpson Desert area in central Australia. In this case, the double peak indicates two principal types of surface characteristics in the image; this is confirmed by a dumb-bell structure in the two-dimensional histograms. Table 5.2 shows the band correlations for this image; again, extremely high correlations between bands are evident.

The colour image of figure 5.4(a) was created using a conventional approach with MSS bands 4, 5 and 7 displayed on blue, green and red colour guns respectively. The high correlations prevent expansion for good chromatic contrast; it is not possible to further expand the chromatic component without also expanding, and hence saturating, the lightness. Figure 5.4(b) was created using the technique described in this section, with approximately 1% of the image points saturated. Again, the improvement in colour contrast is substantial, and hence data characteristics are more widely separated. The image also retains the essential character of the conventionally displayed false colour image. Figure 5.4(c) shows a set of 4 sub-sections from the images of 5.4(a) (upper set) and 5.4(b) (lower set), illustrating in detail the extent to which the full use of colour space (and inclusion of band 6 information) in 5.4(b) increases the degree to which information inherent in the data is rendered visible.

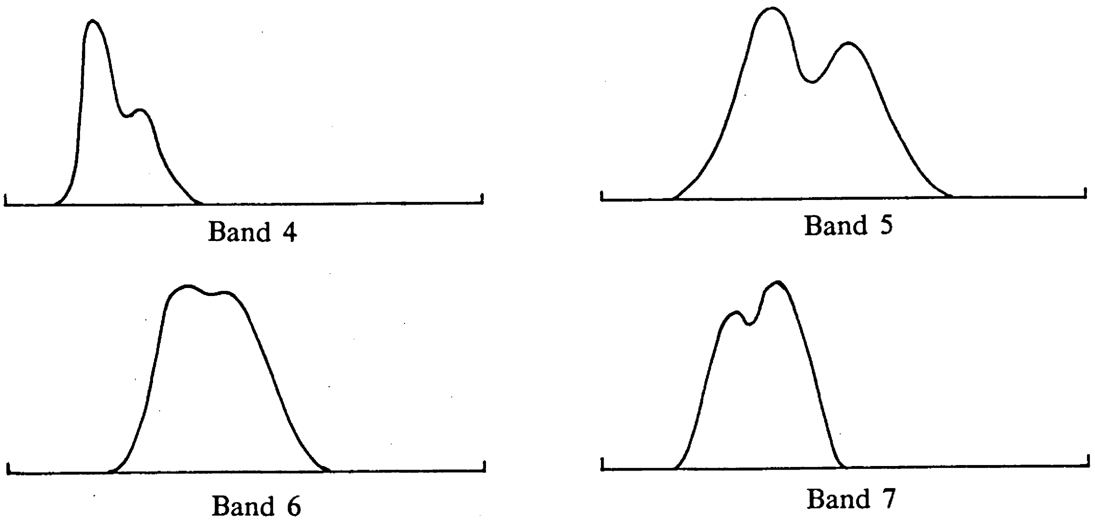


Figure 5.6 Histograms of a multi-modal correlated Landsat MSS mosaic of the Simpson Desert in central Australia. Two major types of physical characteristics give the histograms a bi-modal distribution; uniform dry flat vegetation results in highly correlated spectral bands.

	Band 4	Band 5	Band 6	Band 7
Mean	57.45	105.94	106.84	79.95
Standard deviation	14.69	26.47	21.73	17.02

	Correlation matrix			
	Band 4	Band 5	Band 6	Band 7
Band 4	1.00			
Band 5	0.88	1.00		
Band 6	0.79	0.95	1.00	
Band 7	0.73	0.90	0.97	1.00

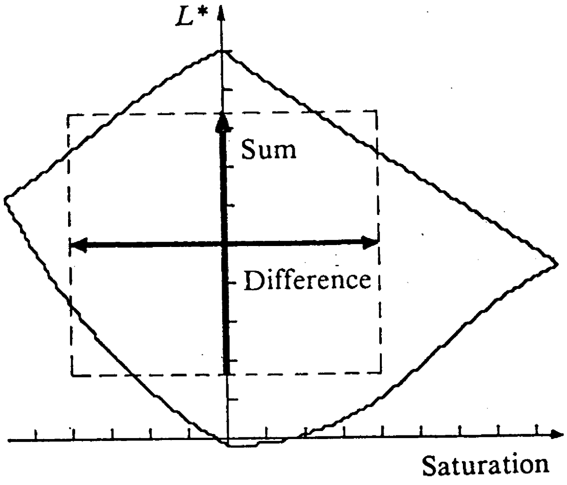
Table 5.2 Spectral band statistics and correlations for a mosaic of Landsat MSS scenes of the Simpson Desert area in central Australia.

5.4 Discussion

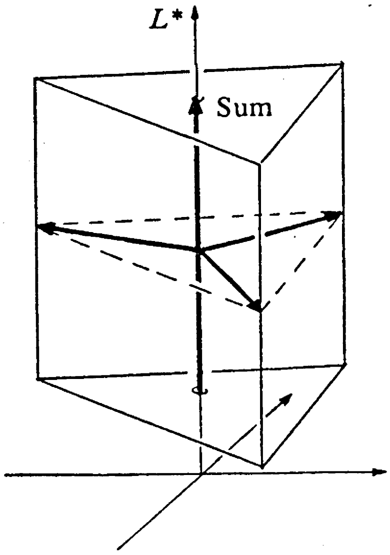
5.4.1 Application of techniques to remotely sensed data sets not derived from reflected radiation

The emphasis in this chapter has been on developing techniques for depicting covering characteristics on a surface, the structure of which is inherently embedded in the data. However, the methods of achieving colour expansion, uniform representation of data in colour, and alignment of chosen data variables in perceptually defined directions in UCS can also be applied to data which does not rely on surface topography depiction for comprehension. The lower set of images in figure 5.3(a), (b) and (c) show examples of mapping pairs of variables into planes in UCS, while figure 5.3(d) (lower) shows three variables mapped into the full colour volume. The data variables are derived from measured counts of uranium, thorium and potassium radioisotopes, recorded remotely in an aerial survey, and are consequently measures of emitted, rather than reflected, radiation. While these images are not directly interpretable as real-world images, they are nevertheless useful to a trained analyst, and benefit from application of the approach described in this chapter.

Relevant to interpretation of these data sets is not just absolute values, but rather relative proportions of the measured counts. Consequently variables mapped into colour space were, in 5.3(a), (b) and (c), the sum of two channel counts (into lightness), and the difference between the same two channels (into saturation). These mappings are shown schematically in figure 5.7(a). The chosen UCS cross-sections were red/magenta-cyan (uranium-thorium), green-blue/magenta (thorium-potassium), and blue/cyan-orange (potassium-uranium). In 5.3(d) three variables, being each radioisotope count, were mapped into saturation directions (from the grey axis) $2\pi/3$ apart (uranium-red; thorium-green; potassium-blue), while the overall sum of counts was mapped into lightness. Figure 5.7(b) shows this mapping schematically.



(a) Mapping two variables, the count sum and the count difference, into lightness and saturation in a constant-hue UCS cross-section.



(b) Mapping three variables into saturation in hue directions $2\pi/3$ apart, and the sum of the three into lightness.

Figure 5.7 Schematic of mapping proportional variables in UCS for maximal chromatic expansion

The resulting images are not readily understood as real-world scenes, but were designed to be meaningful to an experienced analyst. They are an example of how this approach; that is, controlled alignment of chosen variables, the use of UCS gamut shapes for data expansion, and the regular representation of numerical data differences by perceived colour differences; can be used to advantage in more general display problems.

5.4.2 Treatment of saturated pixels

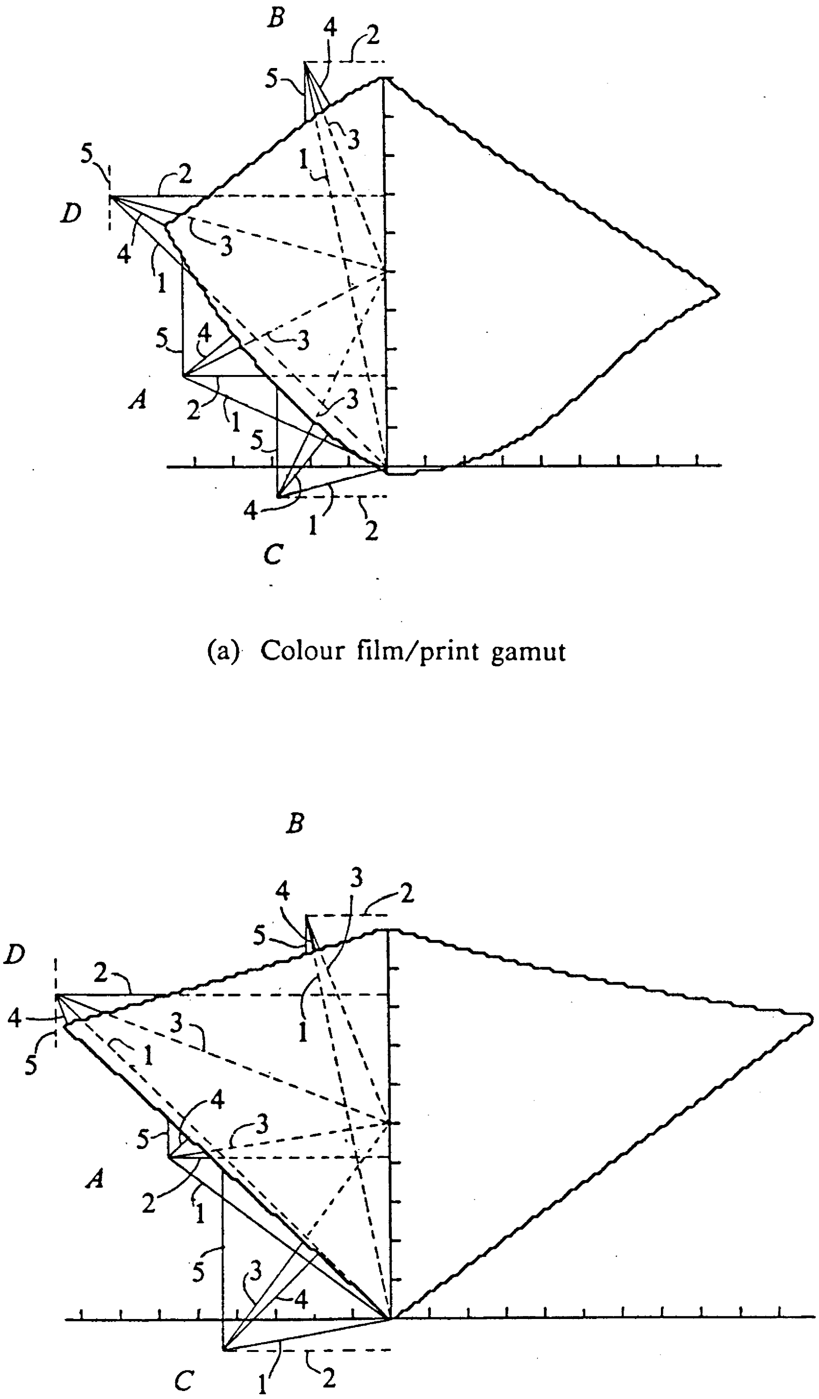
As mentioned previously, both imposing the linearity condition, and scaling to achieve reasonable colour volume utilisation, will generally mean that some proportion of the pixels in an image will be outside the range of achievable colours of the display device. In practice, in a well balanced image with good colour spread, 1-2% of the pixels might be saturated.

A saturated pixel has one or more of its gun-count space co-ordinates outside the realisable limits (fully off or fully on); consequently it is produced by application of a device model outside the range over which it is valid. For a model which involves polynomial approximation, such as the film/print model, the resulting gun-counts produced could be quite misleading, since in the region outside the gamut the polynomials are unconstrained. Hence the simplest approach to treating a saturated pixel, direct truncation to the gun-count range limits, can produce colours which would not be expected from the UCS value of the pixel; in particular the resulting hue can be markedly different from the UCS hue of the saturated pixel. Consequently a more sophisticated method of dealing with saturated pixels is required.

Implicit in imposing linearity in the mapping from informative data space to UCS is that some of the information carried by saturated pixels will be lost. (The alternative of maintaining *one-to-oneness* but relaxing linearity to allow, say, compression towards the volume edge, will also result in an information loss due to reduced resolution and discernibility of numerical data differences.) A method of minimising this information loss, while maintaining an aesthetically acceptable product in which the saturated points do not detract from data appreciation, is required. It could be argued that since the true numerical value of the saturated pixel cannot be reflected in colour, it is better to make that clear by designating to the pixel some colour significantly different from that of its neighbours, such as black or white. However, this could be detrimental to interpretation, and unsatisfactory on aesthetic grounds. At the risk of misleading, it seems attractive to maintain some of the information in the pixel, such as the point or locality of the colour volume closest to it.

In practice it was found to be most acceptable to project the pixel along some chosen path to the colour volume boundary, or to just inside it, thus giving it a colour in some way in keeping with its nature. The path chosen can depend on the data type; in general for reasons of visual consistency it was found to be desirable to keep the hue of the pixel unchanged, and to change either saturation or lightness or both. Ideally some physical rationale based on the nature of the informative data variables being displayed should be used to determine the direction of the projection. Consideration of gamut shapes can also affect the choice of method for returning saturated pixels to the gamut. In addition, the actual pixel value can govern the suitability of a particular return strategy. In figures 5.8 (a) and (b), sample gamut cross-sections are shown (for the colour film recorder and colour monitor gamuts), together with a sample set of saturated pixels. Lines show possible return paths for these saturated pixels, with line annotations corresponding to a set of possible return strategies which modify the pixel values as follows:

(1) return the pixel to the origin in UCS;
(2) return the pixel to the lightness axis at an unchanged lightness value;
(3) return the pixel to the mid-point of the lightness axis;
(4) return the pixel to the closest gamut point;
(5) return the pixel to the closest point of the same saturation and hue;
(6) assign some specified distinctive colour to the pixel.



(a) Colour film/print gamut

(b) Colour television monitor gamut

Figure 5.8 Schematic of strategies for the treatment of saturated pixels

As a general rule, unless strategy 6 is to be used, if the range of the mapping is a one- or two-dimensional subspace of UCS, saturated pixels should be returned into that subspace. This is to prevent isolated pixels, with spectral characteristics markedly different from those of their neighbours, having a distracting effect on appreciation of the data. For example, if a constant-hue plane is the range of a two-dimensional mapping, saturated pixels should be returned to that plane. Physical considerations will often suggest a particular strategy, but the practical problems of dealing with pixels which then have no obviously appropriate return point cannot be overlooked. Pixel *A* under strategy 1 is such a pixel, as are pixels *B* and *C* under strategy 2, and pixel *D* under strategy 5. Even strategy 3, which is essentially a trap-all strategy, can result in pixels in an image apparently dislocated in lightness; pixel *C* would appear overly light. Returning a pixel to its closest gamut boundary point (strategy 4) generally has much the same effect as strategy 3, due to the shape of most display device gamuts, but it can involve considerably more computation unless the gamut boundary can be parametrised.

From a computational point of view, each of the schemes is straightforward to implement if the gamut boundary can be parametrised. The equation of the line joining the saturated pixel to its chosen return point can be obtained, and the intersection of this line with the gamut boundary can then be calculated. For the colour film/print model, in which each gun-count value is expressed as a polynomial in UCS co-ordinates, this involves attempting to find solutions for zero or maximum gun-count values for each of the colour guns, since it will not in general be known which will be saturated for an arbitrary point. Of the six possible cases, only two real-valued solutions will exist, one being the required intersection, and the other being the intersection on the other side of the gamut. (For a return line passing through the UCS origin, these solutions may be co-incident.) The solution closest to the saturated point is that required.

For the colour monitor model, however, such an approach cannot be used because the individual gun counts are not separately defined in terms of UCS values; a set of possible solutions results. Consequently a more general method, independent of the display device modelling technique, is really required. One such method consists of stepping a pointer along the return line, in a direction determined by whether the pointer is inside or outside the gamut, in ever decreasing steps until the pointer is within the gamut, and close to the gamut edge to within any specified accuracy (the final step size following a sign change must be less than the required accuracy). The saturated pixel is then returned to this position. This method is computationally more expensive than finding an exact intersection point, as it can involve a substantial number of conversions calculations per pixel, but its generality makes it more attractive. For this reason it was implemented in the developed system.

Returning a pixel to the gamut with a high degree of accuracy can result in homogeneous patches of colour at a boundary point; this can be visually distracting, even though it might be an indication of saturation in that area. If it is to be avoided, pixels can be returned to the gamut boundary with a lower degree of accuracy, resulting in slight texturing in such an area. This also has the computational advantage of being much faster. In practice, an accuracy of better than 2 JND was generally found to be satisfactory from both visual and computational points of view.

As mentioned earlier, if the range of the transform mapping is a subspace of UCS, a saturated pixel should be returned to this subspace. In practice this can be achieved by projecting the return line to the subspace, and then returning along this projection.

5.4.3 Noise considerations

The techniques described in this chapter have not included any consideration of the level of noise in a data variable. If a data variable is particularly noisy, maximal expansion in UCS can result in chromatic speckle. This can occur when a statistically based transform, such as a principal components transform, is performed on a data set, and a channel with low statistical variance is displayed. Highly correlated Landsat MSS data sets often have very noisy third and fourth principal components. This is because while the transform is chosen to maximise signal variance, noise variance is generally uniform in the spectral space. Thus while the signal-to-noise ratio in the spectral bands will be very similar, in the transformed principal components it will decrease progressively with the order of the components. Expansion of the lower order components to increase the signal variance for display will also result in expansion of the noise component.

However, it can still be important to display a noisy component, as it might contain significant features in some, if limited, areas of the spatial extent. Chromatic speckle resulting from expansion of a noisy channel can cause appreciable visual distraction; consequently some method of dealing with it is required. In fact there are good reasons for suppressing non-informative high frequency chromatic signals, as described in detail in Chapter 7. Smoothing the noisy channel, to an extent which can either depend on the data variance or be interactively determined, results in effective suppression of this speckle. While some information will also be lost, useful information in such an image can only be extracted relative to the noise level, and in fact is generally of a lower frequency than would be appreciably affected by smoothing.

In fact smoothing was performed on the third principal component of the image of figure 5.3(c) (upper) to reduce chromatic speckle. Figure 5.3(d) shows images of an expanded sub-section of this image, enlarged for emphasis, illustrating the effect of smoothing PC2 and PC3 to varying degrees. In the upper left hand image the unsmoothed data was used, resulting in a speckled effect. In the upper right hand image, PC3 was smoothed using a $3x3$ box filter which replaced each pixel value with the average of the values in the $3x3$ neighbourhood centred on the pixel. (This is the smoothing applied to the full image of 5.3(c).) In the lower left hand image, PC2 and PC3 are smoothed using a $3x3$ box filter, and in the lower right hand image, PC2 with a $3x3$ box filter and PC3 with a $5x5$ box filter. Because of the histogram modification effect of spatial-frequency filtering in this manner, the range of colour space used is slightly different for smoothed and unsmoothed data; for the sub-sections shown, chromatic expansion to a uniform level of saturation in the overall image was performed. This histogram modification effect is discussed in Chapter 7.

5.5 Conclusions

This chapter has developed an approach to mapping remotely sensed data sets into UCS in a way which considers both the nature of the data and the perceptual attributes of the colour space. It considers the problem of spectral assignment for up to three channels of continuous real data which depict the covering of a surface whose structure is inherently embedded in the data, and preserved by appropriate choice of the mapping. It thus extends the approach to mapping discretely levelled data variables (on a flat surface) into UCS which was developed in Chapter 4. The method developed is directly realisable within the framework developed in Chapter 3.

The importance of extracting the informative aspects of the data for portrayal in chosen colour attributes is emphasised, and examples of suitable treatments for informative data variables generated by various methods are given.

The method of specifying alignment of informative data variables in UCS was developed from practical considerations; it satisfies the basic requirement of being able to specify data variables in data-dependent terms and their alignment in UCS in perceptual terms, allowing the required representation of informative data in intuitively appreciable colour ranges.

The approach to colour utilisation was also found to be applicable to remotely sensed data sets derived from emissive sources, and thus not inherently depicting surface structure, with the proviso that under such use, the analyst is not trying to interpret surface topography and coverings, but rather relies on experience to interpret the data.

Chapter 6
Synthesis of scenes with implied spatial three-dimensionality

6.1 The problem of displaying composite images of data variables with dissimilar spatial structure

Spatial superimposition of several image data variables with different spatial structure can be used to aid interpretation of multi-dimensional data. In geographic information systems, data variables may represent geological, geophysical, ecological, statistical or other measured or derived variables; very often one variable will describe the topography of the area in the form of height data. Generally the data variables are of a common geographic area, and their integration allows the spatial distribution of one variable to be directly related to that of another. A major goal of the composite image is that not only should the spatial correlations between the data variables be clearly displayed, but that it should also be possible to visually separate and fully appreciate each individual data variable. This separability is seldom achieved when two superimposed image data variables have arbitrary and largely dissimilar spatial structures. As described in Chapters 1 and 2, the conventional method of presenting each data variable on a primary colour, or combination of primary colours, of a display device results in an image which is difficult to interpret.

This situation is illustrated in the colour plates of figure 6.6. Figure 6.6(c) is an overlay of two image data variables, depicted in cyan and red colours respectively. The first data variable, shown in figure 6.6(a), is an image of surface height derived from a contour map. The second data variable, shown in figure 6.6(b), is an image of magnetic field strength over the same area derived from data acquired by an aerial survey. The composite image is difficult to interpret because it cannot be appreciated as a realistic scene.

One approach to this problem is to represent the data variables to be interpreted as naturally occurring properties or variables of a synthesised real-world scene. The visual system is experienced at interpreting surfaces and their coverings, and if these can be simulated in a sufficiently realistic manner, then the normal scene analysing capabilities of the visual system can be exploited. This should allow dissociation, and hence intuitive appreciation, of the individual data variable characteristics, while still allowing portrayal of correlated characteristics.

As seen from figure 6.6(c), it is extremely difficult to fully appreciate the characteristics of more than one data variable portrayed on a surface (in this case a flat one) if the variables have dissimilar spatial structures. To allow appreciation of another variable, some kind of further dimensionality must be introduced. This can take the form of an implied spatial three-dimensionality which results in a non-flat surface, the topography of which represents the additional variable in some chosen way. Alternatively a temporal variation can be used, though this might be less satisfactory from computational or presentation considerations.

Implied spatial three-dimensionality has been widely used in the form of shaded-relief, perspective, or stereoscopic viewing of geographic and other data. In this chapter an approach using a relief- or hill-shading method is developed. The same techniques can be applied to generate realistic perspective or stereoscopic views, with the penalty of greater computational expense that these display methods entail. An added advantage of a shaded relief representation is that the image is displayed on its original rectangular spatial grid. This means that direct overlays of, for example, line or area type data can be made, increasing its usefulness to portray geographic information. Further, direct measurements can easily be taken from such an image, making it more useful in an interactive environment.

Methods of data presentation which use the colouring of a three-dimensional surface have been developed (TASC, 1981; UNIRAS; Arvidson et al., 1982), but the colouring and illustration methods used have not resulted in an apparently realistic scene. Intuitive interpretation of natural scene variables, and hence associated data variables, does not then take place. Realistic colouring of surfaces, or portions of surfaces, with a single colour has been achieved by modelling the surface reflectance characteristics (Cook and Torrance, 1982; Warn, 1983). However, scenes in which the surface colour varies in a realistic and natural manner to depict a data variable with arbitrary two-dimensional variation have not been produced; it is this problem that is addressed in this chapter (see also Robertson, 1984; Robertson and O'Callaghan, 1985).

To synthesise realistic non-flat scenes, we first need to choose appropriate scene representations for the data variables. Various investigations into computational approaches to scene understanding (Horn, 1981; Marr, 1982; Pentland, 1984; and see Brady, 1982 for a summary) have suggested the causes of lightness and colour variations on surfaces, and ways of extracting (and hence, conversely, depicting) them. Secondly, the chosen representations must be rendered realistically, suggesting the application of recently developed techniques in the field of colour graphics (Phong, 1975; Blinn, 1977; Whitted, 1980; Cook and Torrance, 1982) to simulate colour reflectance from surfaces.

In this chapter we present an approach to representing physical image data as a realistic scene. One variable is represented as the topography of a surface, and another as a property which varies in colour on that surface in the form of a pigmenting colour, or above it in the form of a pigmented transparent sheet. Implementation of this approach is described, and its effectiveness and limitations are discussed.

6.2 Realistic depiction of a coloured surface

We consider first the display of two data variables in the form of a coloured surface. The success of the technique relies upon achieving a realistic representation of one data variable as the topography of the surface, and of the other as a surface property. The representations must be such that the effects of one variable on the perception of the other are present, so that the normal visual processes of compensating for such effects in realistic situations can take place.

Surface depiction can be achieved by means of lightness variations which are dependent on illumination and viewing angles relative to the surface normal (Horn, 1981). In most viewing situations, the reflected intensity reaching the viewer has two components; a specular component representing light reflected directly from the surface of the material, and a diffuse component arising from scattering on or below the surface of the material (Judd and Wyszecki, 1975; Cook and Torrance, 1982). Each illuminating source contributes to the reflected components, and in addition, an ambient illumination component arises from background scattering. The directional dependence of the reflected intensity is responsible for realising surface depiction; while it is the direction of the surface normal which determines the relative magnitude of the perceived reflected components, the visual system competently and intuitively extracts the apparent surface height, at least relative to its neighbourhood. A surface can be depicted with diffuse reflected components only, but inclusion of a specular component with an appropriate directional distribution lends greater realism, and can resolve ambiguous situations. Cook and Torrance suggest that while the spectral distribution of the diffuse reflected component varies little with the direction of illumination, this is not so for the specular component. Rather a colour shift takes place when angles of incidence and reflection approach grazing. This shift depends on the surface material, and is generally small for illumination/viewing angles of less than $\pi/2$, which occurs for overhead viewing and lighting of a scene.

The importance of characterising the type of surface is also stressed by Cook and Torrance. For example, plastic materials produce a component reflected from the substrate surface which has a spectral distribution very close to that of the illumination, and a substantial diffuse component due to scattering within the pigment. Metals, on the other hand, generally have a largely specular reflection since surface penetration is minimal, and the reflection is generally closer to the apparent colour of the metal than to that of the illumination. Choosing the appropriate specular directional distribution also affects the degree of realism achieved, and for this purpose Cook and Torrance follow Torrance and Sparrow (1967) in modelling surfaces as consisting of specularly reflecting microfacets with orientation distribution depending on surface roughness. Cook and Torrance's reflectance model achieves the realistic rendering of a wide range of types of three-dimensional surfaces to a degree not previously attained. Consequently it was used as a basis for this work, and has been extended to cover the requirement of displaying arbitrarily varying colours on a surface. In the following sections a computational approach to this problem is developed.

6.2.1 Surface depiction - reflectance model

The basic form of the reflectance model proposed by Cook and Torrance defines the reflected intensity I_r from a surface by

$$I_r = I_{ia}R_a + \sum_l \left\{ I_{il}(\mathbf{N}.\mathbf{L}_l)\, d\omega_{il}(dR_d + sR_s) \right\}$$

where the surface geometry is as indicated in figure 6.1. The first term results from ambient illumination, while the second term describes diffuse (d) and specular (s) reflected components of a number of light sources l. The dot product **N.L** modulates the reflected intensity according to the orientation of the surface, and $d\omega$ is the solid angle subtended by the source. I_i is the relative intensity of the impinging radiation as a function of wavelength. R_a is the ambient reflectance, R_d the diffuse reflectance (the normally measured surface reflectance function $r(\lambda)$), R_s the specular reflectance, and d and s are the diffuse and specular fractions of the reflected component ($d + s = 1$).

More specifically, the specular reflectance can be derived from the microfacet surface model (Torrance and Sparrow, 1967) and is given by

$$R_s = \frac{FDG}{\pi(\mathbf{N}.\mathbf{L})\,(\mathbf{N}.\mathbf{V})} .$$

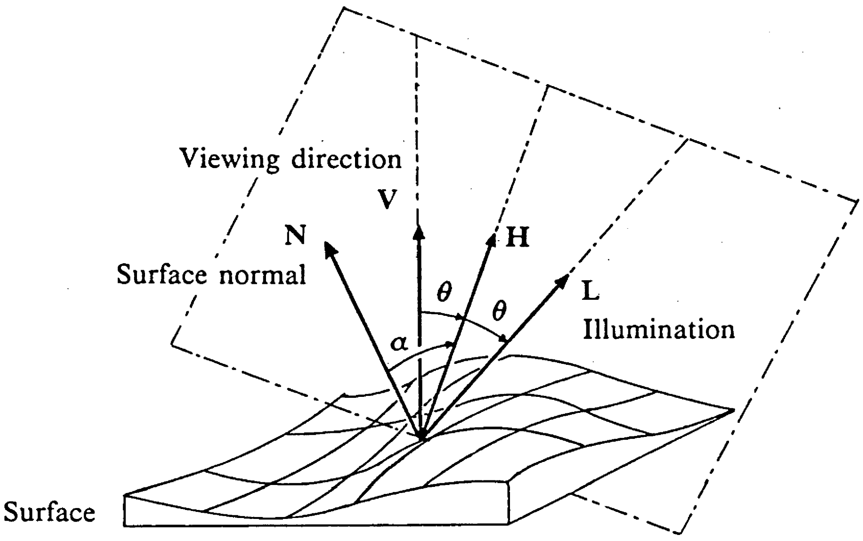


Figure 6.1 Geometry of surface reflectance model

G is a geometrical attenuation factor accounting for the mutual masking and shadowing of surface facets (Torrance and Sparrow, 1967; Blinn, 1977) and has the value

$$G = \min\left\{ 1, \frac{2(\mathbf{N.H})(\mathbf{N.V})}{(\mathbf{V.H})}, \frac{2(\mathbf{N.H})(\mathbf{N.L})}{(\mathbf{V.H})} \right\}.$$

D, the facet slope distribution function, describes the fraction of facets specularly reflecting (aligned normal to **H**, the directional bisector of viewing (**V**) and illumination (**L**) directions). Hence the value of D depends upon the angle between the surface normal and each facet normal, and upon the spread of the distribution. This spread is determined by the steepness of the facet slopes; smooth surfaces have facets more closely aligned with the surface while rough surfaces have more steeply sloped facets; and is characterised by the root-mean-square slope m. Section 6.4.2 further discusses facet slope distributions.

F describes the wavelength dependence of the specular reflectance from each individual microfacet, and can be derived from the Fresnel equation, which accounts for reflection from a smooth mirror-like surface, allowing for surface layer penetration of the impinging radiation. It depends on the incident illumination angle θ, the refractive index μ of the material, and its extinction coefficient. Cook and Torrance suggest an approximate method of evaluating F, relying on estimating the refractive index for an extinction coefficient of zero (non-metals have a zero extinction coefficient), which generates correct values at normal incidence and a good estimate of the incidence angle (θ) dependence of F. Specifically then,

$$F = \frac{(g-c)^2}{2(g+c)^2}\left\{ 1 + \frac{[(g+c)-1]^2}{[(g-c)+1]^2} \right\},$$

where $c=\cos\theta$, and $g^2 = \mu^2 + c^2 - 1$.

At normal incidence $\theta=0$, and F collapses to

$$F_0 = \left\{ \frac{\mu-1}{\mu+1} \right\}^2,$$

from which μ can be determined, since F_0 is the normally measured specular reflectance distribution. Substituting μ into F gives an estimate for F at incidence angles off-normal.

As θ approaches $\pi/2$, F approaches unity for all wavelengths, and this is the incidence-angle dependent spectral shift which takes place when the illumination comes close to grazing the material surface.

Under the proposed overhead viewing and side illumination of a scene, this spectral shift is small, making the evaluation of the θ-dependence of F less important for realism. However, for reasons described in later sections, wavelength-specific access to reflectance spectra is required, and possible extension of the developed technique to more general viewing conditions (such as perspective viewing), make it worth retaining this aspect of the model.

This, then, is the reflectance model proposed by Cook and Torrance. Its advantage over the earlier models of Phong (1976), Blinn (1977), and Whitted (1982) lies in the detailed modelling, and hence control over, the specular component of the reflected illumination. We now extend the reflectance model to allow incorporation of a model of surface colour pigmentation with spatially varying spectral distribution.

Cook and Torrance present their model for depicting a surface with a single colour, and propose an approximation for evaluating resulting surface colour co-ordinates which avoids lengthy wavelength-dependent calculations at every point in an image. Allowing a surface point to have an arbitrary spectral reflectance function, such as is required to depict variations in pigment colour and density, means that this approximation can no longer be used directly. Instead, we develop an approach which separately evaluates tristimulus components due first, to ambient and diffuse reflections, which are taken as having uniform directional distributions; and second, to specular reflection, which is dependent on the illumination angle (θ).

We can treat the ambient illumination as having a component from each source:

$$I_{ia} = \sum_l a_l I_{il} d\omega_{il} ,$$

where a_l is a fractional constant. Specifying wavelength dependence, the reflected intensity can be expressed as

$$I_r(\lambda) = \sum_l I_{il}(\lambda) d\omega_{il} \left[a_l R_a(\lambda) + (\mathbf{N.L}_l) \{ dR_d(\lambda) + sR_s(\lambda) \} \right] .$$

Cook and Torrance suggest that it is reasonable to take the ambient illumination as a hemispherical-surface-dependent factor of the diffuse illumination, or specifically, to use

$$R_a(\lambda) = \pi R_d(\lambda).$$

We then have

$$I_r(\lambda) = \sum_l I_{il}(\lambda) d\omega_{il} \{ W1_l R_d(\lambda) + W2_l F(\lambda) \} \qquad (1)$$

where

$$W1_l = \pi a_l + (\mathbf{N.L}_l) d \quad \text{and} \quad W2_l = \frac{sDG}{\pi(\mathbf{N.V})} .$$

$R_d(\lambda)$ is the diffuse bidirectional reflectance and is taken by Cook and Torrance as being invariant to direction. (In fact, this holds for viewing/illumination angles of up to about 70°.) It is thus approximated by the bidirectional reflectance under normal incidence, $r_d(\lambda)$. $R_s(\lambda)$ can be evaluated using the refractive index approximation at normal incidence, when $F(\lambda) = r_s(\lambda)$, $r_s(\lambda)$ being the reflectance function for normal reflectance from a polished surface (in this case applied to each individual microfacet). We note that $I_r(\lambda)$ depends upon θ (the illumination angle) and α (the angle between the surface normal and the viewing/illumination angle bisector). More specifically, both $W1$ and $W2$ are θ-dependent ($\cos\theta = \mathbf{N.L} = \mathbf{N.V}$), $W2$ is α-dependent (due to the distribution factor D), and $F(\lambda)$ is θ-dependent.

We can evaluate tristimulus values from

$$X = \int_{\lambda} \bar{x}(\lambda) I_i(\lambda) r(\lambda) \, \mathrm{d}(\lambda)$$

and similarly for Y and Z, where $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ are the CIE Standard Observer colour matching functions (see Appendix 1), $I_i(\lambda)$ is the impinging radiation, and $r(\lambda)$ is the surface albedo or reflectance function.

Hence the observed surface colour tristimulus values are given by

$$X_{\theta,\alpha} = \sum_{l} d\omega_{il} k \left\{ W1_{l,\theta} \int_{\lambda} \bar{x}(\lambda) I_{il}(\lambda) \, r_d(\lambda) \, \mathrm{d}\lambda + W2_{l,\theta,\alpha} \int_{\lambda} \bar{x}(\lambda) I_{il}(\lambda) F_{\theta}(\lambda) \, \mathrm{d}\lambda \right\}$$

and similarly for $Y_{\theta,\alpha}$ and $Z_{\theta,\alpha}$. The terms within the integrals correspond to diffuse and specular tristimulus components $X_{d,l}$ and $X_{s,l,\theta}$ respectively, so we have

$$X_{\theta,\alpha} = \sum_{l} d\omega_{il} k \{ W1_{l,\theta} X_{d,l} + W2_{l,\theta,\alpha} X_{s,l,\theta} \} \tag{2}$$

and similarly for $Y_{\theta,\alpha}$ and $Z_{\theta,\alpha}$.

The normalising constant k is taken as $100/Y_{\theta,0}max$ where

$$Y_{\theta,0}max = \sum_{l} d\omega_{il} \{ W1_{l,\theta}max \; Y_{d,l} + W2_{l,\theta,0}max \; Y_{s,l,\theta} \}$$

and $W1_{l,\theta}max$ and $W2_{l,\theta,0}max$ are the maximum values of $W1$ and $W2$ for each light source l, for $\alpha=0$. This results in an image normalised to the applied illumination which can then be realised using a display device model as described in Chapter 3. The colour at any point in the image depends upon diffuse and specular tristimulus components, the generation of which is treated in the following section.

6.2.2 Surface colouring - pigmentation model

Given then that any chosen colour can be realistically portrayed on a chosen surface type (provided its diffuse and specular spectral reflectance functions are known), a rationale for choosing colour ranges to represent the surface property is required. For a surface which is to appear as if painted by the surface property variable, a form of colouring which conforms with the general appearance of paint pigment is suitable. Pigmentation can be modelled as a stack of identical filters, leading to a power law of increasing density (Beer's Law; see Wyszecki and Stiles, 1967). Varying pigment density causes a variation in lightness and saturation, but hue remains very close to constant. The crucial aspect of a variation in pigment density, according to Rubin and Richards (1982), is that it corresponds to a smooth variation in surface reflectance function. Furthermore, reflectance functions are monotonically related to pigment density. Pigment mixtures can be treated by various methods, an example of which is the Kubelka-Munk analysis (see Judd and Wyszecki, 1975), but in general follow these same basic rules.

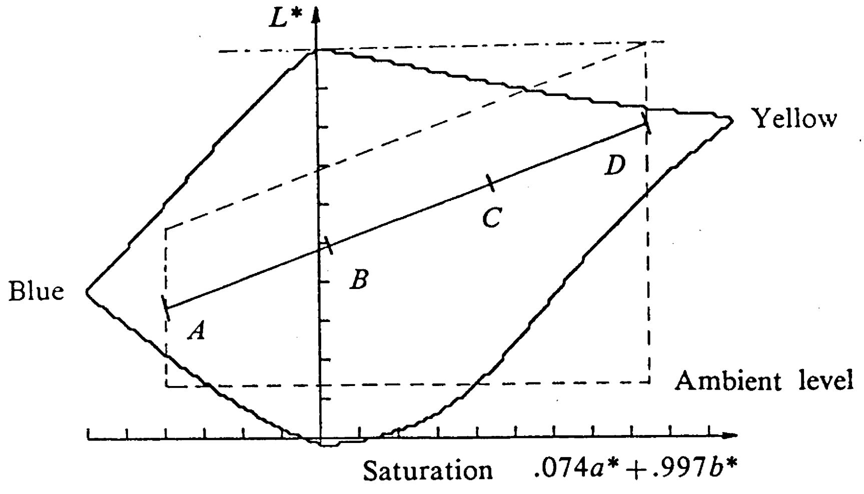
This simple pigmentation model was used to generate surface colour ranges from one specified colour to another, corresponding to gradually decreasing the density of one pigment and increasing that of the other. Single pigment ranges were also generated, but were found less successful for reasons which are discussed in section 6.4. In order to keep explicit control over lightness levels of the colour ranges, the spectral reflectance functions of the chosen enamel paint pigments were normalised to specified lightness values in a perceptually uniform colour space. This corresponds to varying the density of an added pigment of neutral (grey) hue. As described in Chapter 4, (see also Meyer and Greenberg, 1980; O'Callaghan et al., 1981; Robertson and O'Callaghan, 1982; Tajima 1983; Robertson, 1984), it was found advantageous to represent numerical variables by colours evenly spaced according to the metric of a uniform colour space (UCS). In addition, it was found that to indicate an increasing sense of magnitude, or value, an increase in lightness is important (Robertson and O'Callaghan, 1984 - see also Bertin, 1981).

Consequently, a path of regularly increasing lightness through an approximately constant-hue cross-section of UCS was used to generate the required colour range. Such a path which crosses the neutral-hue boundary, or grey axis, corresponds to a mixture of three pigments; a pigment of particular hue which has gradually increasing density, a pigment of approximately complementary hue with gradually decreasing density, and a neutral pigment to control lightness. The resulting colour progression satisfies the requirement of having a smooth change in reflectance function from one end point to the other. The base colour range so generated can then be modulated in lightness according to the surface orientation at any point. Despite this lightness modulation, perceived saturation still uniquely represents the surface property data value. The path in UCS of such a colour range is shown in figure 6.2(a), in a blue-yellow cross-section of the CIELAB film/print gamut. The path limits must be chosen so that excessive saturation does not result from lightness modulation. Surface orientation variations modulate lightness between zero and the base pigment lightness level; normalisation of the scene lightness (above the ambient component) to the maximum lightness results in a colour space utilisation bounded by the broken line shown in figure 6.2(a). Inclusion of a specular component results in a more complex gamut utilisation; this is treated in section 6.4.3.

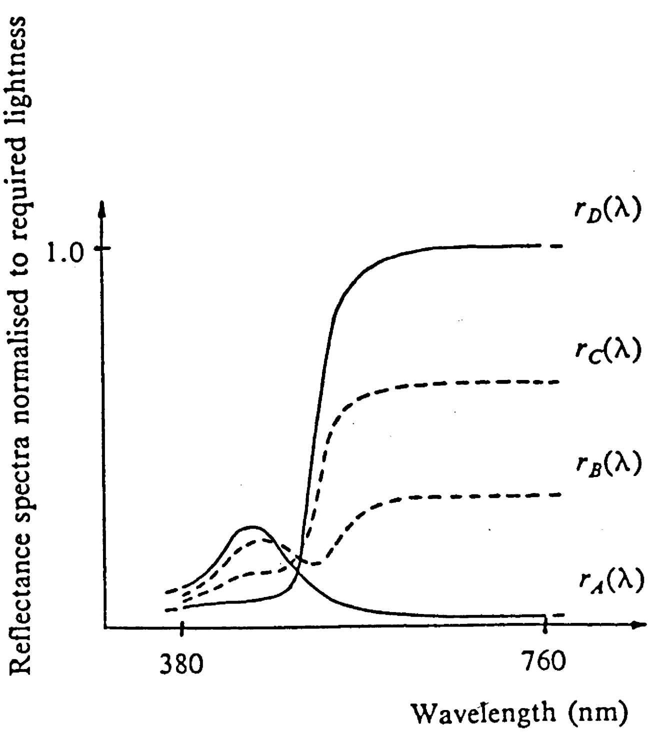
Pigmented materials have an appearance which can be characterised by the spectral distributions of reflected diffuse and specular illumination. These reflectance spectra, $r_d(\lambda)$ and $r_s(\lambda)$ respectively, are actually continuous functions of both wavelength and degree of pigmentation (and hence of the value of the second data variable, $D2$). We artificially create these reflectance functions using two specified extreme value reflectance functions. If low and high value reflectance spectra are $r_1(\lambda)$ and $r_n(\lambda)$ respectively, then for intermediate reflectance spectra we have

$$r_i(\lambda) = (1-i)r_1(\lambda) + ir_n(\lambda) \ , \quad i \in [0,1] .$$

We can then evaluate tristimulus values by weighting with the Standard Observer colour matching functions and the illumination spectral distribution. Normalisation to the required lightness level for a given reflectance spectrum (corresponding to the mixing in of a neutral



(a) Path choice to allow surface-orientation-dependent lightness variations without excessive saturation (boundaries shown in broken lines).



(b) Reflectance spectra at chosen points through the pigment range. End-point spectra are shown in bold.

Figure 6.2 Blue-yellow pigment range with increasing lightness in a constant-hue UCS cross-section.

pigment) can then be performed:

$$X_i = \frac{(Y_i norm) \int_{\lambda} \bar{x}(\lambda) I_i(\lambda)\, r_i(\lambda)\ \mathrm{d}(\lambda)}{\int_{\lambda} \bar{y}(\lambda) I_i(\lambda)\, r_i(\lambda)\ \mathrm{d}\lambda}$$

and similarly for Y_i and Z_i, where $Y_i norm$ is the Y tristimulus value normalising factor calculated from the required lightness value (Appendix 2 gives the relationship between the Y tristimulus value and lightness in each UCS). The two extreme value reflectance functions used were measured reflectance spectra for the appropriate material being simulated, and correspond to the colours and lightnesses representing the extreme values of $D2$ under illumination normal to the surface.

Evaluation of the reflected intensity given by equation 2 for every point in the image would be computationally expensive. Reflectance spectra are generally measured at specific wavelengths and interpolated between measured values. In addition, to generate reflectance spectra for arbitrarily levelled values between extreme spectra would require some form of parametrisation if a metric in UCS were to result. Instead, the reflectance spectra were generated at N discrete colour levels between extreme points, and were represented by samples at 20 nanometer (nm) intervals in wavelength between limits at 380 nm and 760 nm. This produces a set of $Nx20$ samples for which both diffuse and specular tristimulus value components can be pre-calculated. To improve on the coarse quantisation of $D2$ values into N steps, interpolation between pre-calculated tristimulus value components, in proportion to the data value position between the two discrete levels, was performed. The value of N was adjusted to make the error arising from the approximation less than one JND in UCS. In fact the value of N required depends on the length of path in colour space between end point reflectance spectra. For the longest path used, $N=20$ satisfied the required error condition quite adequately. This generates look-up tables of $20x3$ elements for each of the diffuse and specular tristimulus value components.

In order to give the colour sequences even spacing in UCS M spectra were generated ($M >> N$), and a set of N spectra selected as follows. First we calculate the cumulative piecewise linear path distance to sequence element i, D_i, in UCS from

$$D_i = \sum_{h=2}^{i} \left\{ \sum_{k=1}^{3} \left[(U_{k,h} - U_{k,h-1})^2 \right] \right\}^{\frac{1}{2}} \qquad \text{for } i = 2,\ldots,M$$

where $D_1 = 0$. $\{U_k,\ k=1,3\}$ are the three-dimensional UCS co-ordinates calculated from tristimulus values, which are produced by weighting each reflectance spectra with the CIE colour matching functions and the chosen illumination spectral distribution. The average path distance between each pair of N samples is taken as

$$d_{av} = D_M/(N-1)$$

For every j, $j = 1,\ldots,N$ we find an i such that

$$D_i < (j-1)d_{av} \leqslant D_{i+1}\ , \qquad i \in [1,M-1]$$

and set $r_j(\lambda) = r_i(\lambda)$.

This generates a set of reflectance spectra $r_j(\lambda)$ ($j = 1,\ldots,N$) whose colour co-ordinates are approximately evenly spaced in UCS. The closeness of the approximation depends on the value of M; in practice a value was chosen to keep spacing errors less than one JND. The path of a sequence of smoothly varying spectra in a constant-hue cross-section of the CIELAB film/print gamut (shown in figure 6.2(a)) has end-point reflectance spectra as shown in figure 6.2(b) (in bold), corresponding to colours blue and yellow. Intermediate spectra are shown in dotted lines. In each case, normalisation to the required lightness value has been performed. The pigment sequences formed in this manner, and used in the colour illustrations in this chapter, are shown in figure 6.9(e); the blue-yellow sequence described above is shown in range 1.

The overall process for generating an image of a coloured surface is outlined in figure 6.3, with pre-calculated and point-wise calculated sections delineated. For each point in the image, given the surface orientation as determined by the first data variable value

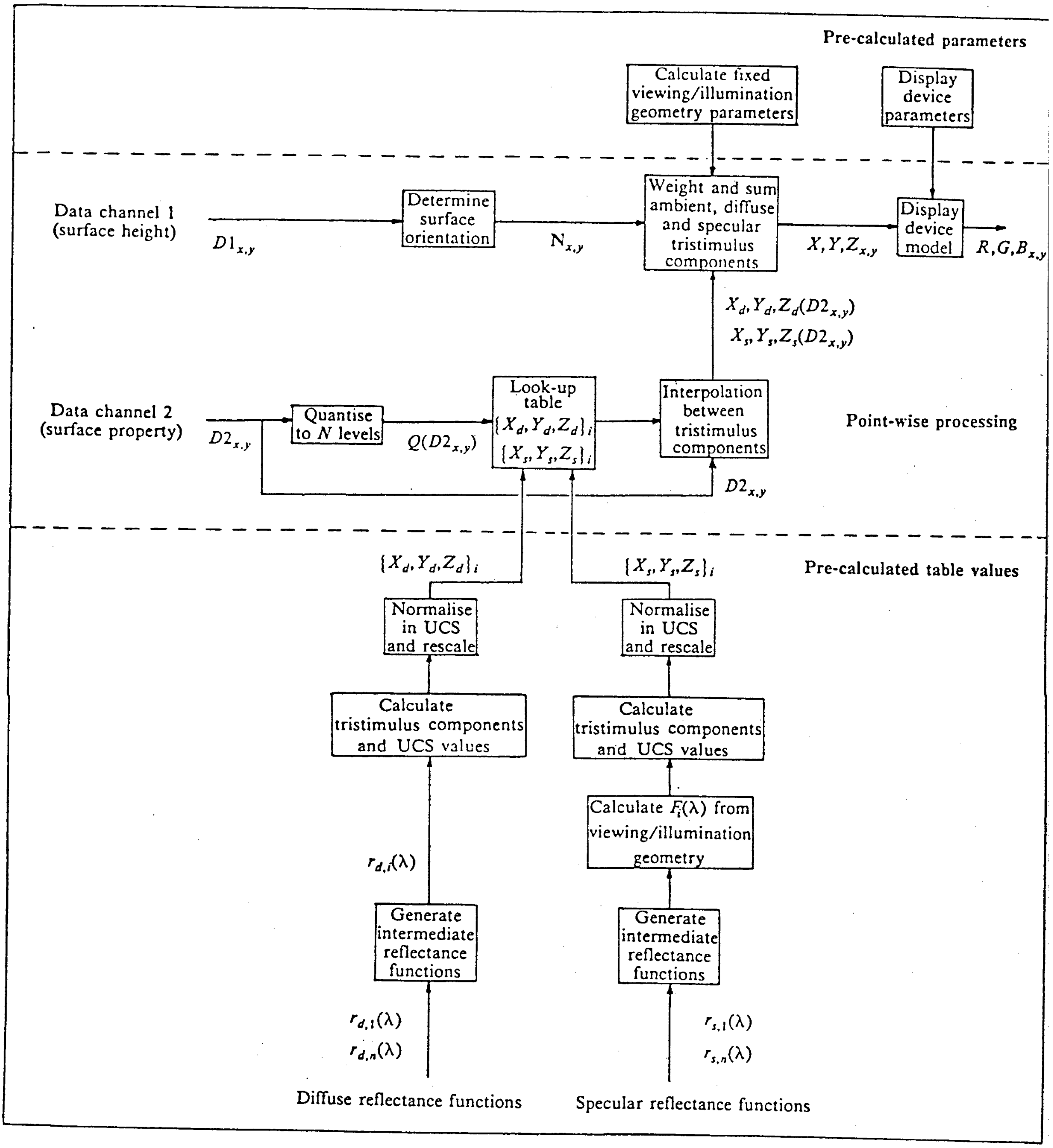

Pre-calculated parameters
Calculate fixed viewing/illumination geometry parameters
Display device parameters
Data channel 1 (surface height)
$D1_{x,y}$
Determine surface orientation
$N_{x,y}$
Weight and sum ambient, diffuse and specular tristimulus components
$X,Y,Z_{x,y}$
Display device model
$R,G,B_{x,y}$
$X_d,Y_d,Z_d(D2_{x,y})$
$X_s,Y_s,Z_s(D2_{x,y})$
Data channel 2 (surface property)
$D2_{x,y}$
Quantise to N levels
$Q(D2_{x,y})$
Look-up table $\{X_d,Y_d,Z_d\}_i$ $\{X_s,Y_s,Z_s\}_i$
Interpolation between tristimulus components
Point-wise processing
$D2_{x,y}$
$\{X_d,Y_d,Z_d\}_i$
$\{X_s,Y_s,Z_s\}_i$
Pre-calculated table values
Normalise in UCS and rescale
Normalise in UCS and rescale
Calculate tristimulus components and UCS values
Calculate tristimulus components and UCS values
Calculate $F_i(\lambda)$ from viewing/illumination geometry
$r_{d,i}(\lambda)$
Generate intermediate reflectance functions
Generate intermediate reflectance functions
$r_{d,1}(\lambda)$
$r_{d,n}(\lambda)$
$r_{s,1}(\lambda)$
$r_{s,n}(\lambda)$
Diffuse reflectance functions
Specular reflectance functions

Figure 6.3 Block diagram of the process of generating a coloured surface

$D1$, and the viewing and illumination geometry, the weighting factors $W1$ and $W2$ can be determined. The diffuse and specular tristimulus components are determined by the value of the second data variable $D2$, and equation 2 can then be evaluated to give image tristimulus values for any given scene geometry and data variable values.

6.3 Modelling an overlaid transparency

An alternative approach to portraying the second data variable as a pigmenting colour on the surface is to use it to pigment a transparent sheet laid over the surface. Although this is not a naturally occurring situation, it is one with which the visual system in general has some experience (Metelli, 1974; Marr, 1982). If the simulation is sufficiently realistic, the dissociated nature of the scene might assist the process of separating the two data variables. Again a simple dye colourant or pigmentation model can be used to generate the colouring effect of the transparent sheet with the spectral distributions being transmittance, rather than reflectance, functions. In fact such a transparency can be laid over a coloured surface, but the additional filtering stage makes it very difficult to retain the original surface colour information. The transparency facility can, however, be used to effect in overlaying specific localised information on a coloured surface, in the form of coloured arrows or patches, without completely destroying underlying information.

6.3.1 Ideal transparency model

Initially, we consider a simple ideal modelling of an overlaid transparency, the geometry of which is shown in figure 6.4. Filtering of the incident scene illumination by the transparency results in spatial confusion unless the lighting is from directly overhead. (Overhead lighting corresponds to a filtering of the impinging radiation and of the reflected signal, or simply to a transparency filter of double the density.) To avoid this spatial confusion, the transparency must be sufficiently above the surface to allow complete sidelighting. As seen from figure 6.4, this requires that $\tan\phi \leqslant \tan2\theta$. The pigmentation

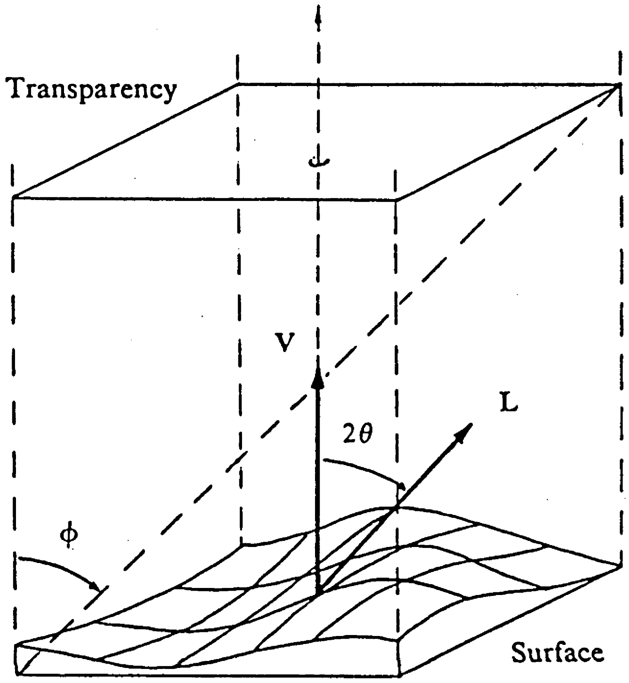


Figure 6.4 Geometry of overlaid transparency model

simulation used for surface colouring was used to colour the transparent sheet, a set of transmittance spectra $t_i(\lambda)$ being generated. Overlaying a transparency involves at each spatial point, multiplying at every wavelength the reflectance spectrum from the surface by the transmittance spectrum. This must be done before tristimulus values are calculated. Using equation 1, the filtered reflected intensity $I_{rt}(\lambda)$ at every point in an image is given by

$$I_{rt}(\lambda) \;=\; \sum_{l} I_{il}(\lambda)\, d\omega_{il} \{\, W1_l R_d(\lambda) + W2_l F(\lambda) \} \; t(\lambda) \qquad (3)$$

Evaluation of this viewed intensity can be simplified by pre-multiplying reflectance and transmittance spectra, again for each of the diffuse and specular reflected components, thus producing filtered tristimulus components X_{dt} and X_{st}, and similarly for Y and Z.

Allowing for the overlay of a variably pigmented transparent sheet on a variably pigmented coloured surface requires that the look-up tables of specular and diffuse tristimulus components be two-dimensional; one dimension is argumented by the surface colour variable and the other by the transparency colour variable. The look-up table is generated in the same manner as before, but with the reflectance and transmittance spectra being multiplied together to generate the resulting filtered reflected signal for each of the 20 surface colour levels and 20 transparency colour levels. This generates look-up tables of $20x20x3$ elements each for diffuse and specular tristimulus components, which is computationally quite manageable. Two-dimensional interpolation between discrete tristimulus component levels must then be performed to produce smoothly varying colours. Thus equation 3 reduces to the form of equation 2 for evaluation of image tristimulus values. That is,

$$X_{\theta,\alpha} = \sum_{l} d\omega_{il} k \{\, W1_{l,\theta} X_{dt,l} + W2_{l,\theta,\alpha} X_{st,l,\theta} \,\} \qquad (4)$$

and similarly for $Y_{\theta,\alpha}$ and $Z_{\theta,\alpha}$. Figure 6.5 outlines the process of generating an image of a surface with an overlaid transparent coloured sheet, again with pre-calculated and point-wise calculated sections delineated.

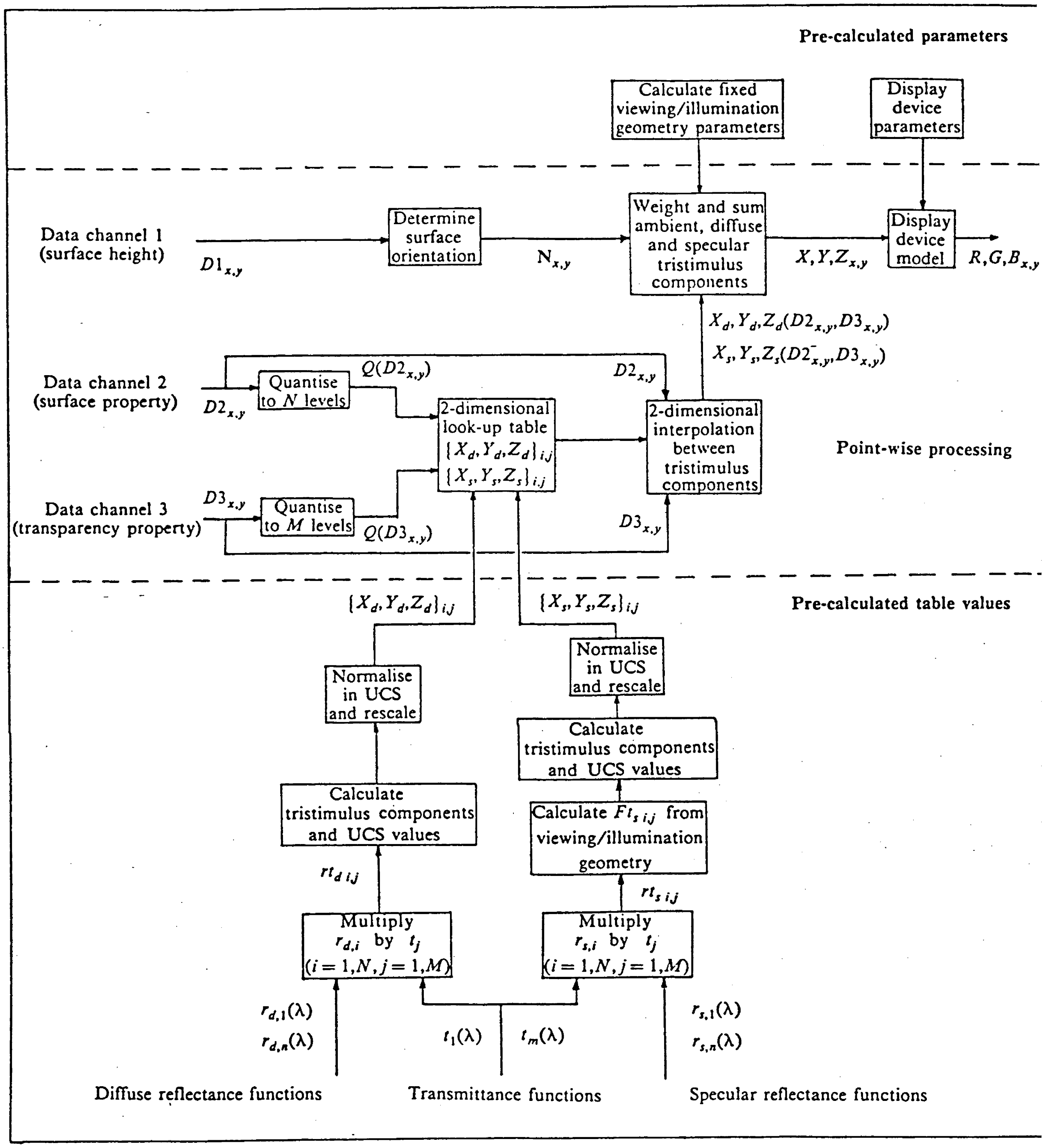

Pre-calculated parameters
Calculate fixed viewing/illumination geometry parameters
Display device parameters
Data channel 1 (surface height)
$D1_{x,y}$
Determine surface orientation
$N_{x,y}$
Weight and sum ambient, diffuse and specular tristimulus components
$X,Y,Z_{x,y}$
Display device model
$R,G,B_{x,y}$
$X_d, Y_d, Z_d(D2_{x,y}, D3_{x,y})$
$X_s, Y_s, Z_s(D2^{-}_{x,y}, D3_{x,y})$
Data channel 2 (surface property)
$D2_{x,y}$
Quantise to N levels
$Q(D2_{x,y})$
$D2_{x,y}$
2-dimensional look-up table $\{X_d, Y_d, Z_d\}_{i,j}$ $\{X_s, Y_s, Z_s\}_{i,j}$
2-dimensional interpolation between tristimulus components
Point-wise processing
Data channel 3 (transparency property)
$D3_{x,y}$
Quantise to M levels
$Q(D3_{x,y})$
$D3_{x,y}$
$\{X_d, Y_d, Z_d\}_{i,j}$
$\{X_s, Y_s, Z_s\}_{i,j}$
Pre-calculated table values
Normalise in UCS and rescale
Normalise in UCS and rescale
Calculate tristimulus components and UCS values
Calculate tristimulus components and UCS values
Calculate $Ft_{s\,i,j}$ from viewing/illumination geometry
$rt_{d\,i,j}$
$rt_{s\,i,j}$
Multiply $r_{d,i}$ by t_j $(i = 1,N, j = 1,M)$
Multiply $r_{s,i}$ by t_j $(i = 1,N, j = 1,M)$
$r_{d,1}(\lambda)$
$r_{d,n}(\lambda)$
$t_1(\lambda)$
$t_m(\lambda)$
$r_{s,1}(\lambda)$
$r_{s,n}(\lambda)$
Diffuse reflectance functions
Transmittance functions
Specular reflectance functions

Figure 6.5 Block diagram of the process of generating a coloured transparency overlaid on a coloured surface

6.3.2 Non-ideal transparency model

In an attempt to enhance the sense of realism in depicting an overlaid transparency, a more complex model was also developed. Rather than treating the transparency as having a substrate of ideal transmittance, it can be treated as a partially opaque surface. In this case, for a transparency factor τ, the viewed signal $I_v(\lambda)$ is given by

$$I_v(\lambda) = \tau I_{rt}(\lambda) + (1-\tau) J_r(\lambda),$$

where $I_{rt}(\lambda)$ is the previously derived signal reflected from the lower surface and filtered by an ideally pigmented transparency, and $J_r(\lambda)$ is the reflectance from the transparency itself of the form given by equation 2. Evaluation of $I_v(\lambda)$ requires evaluation of tristimulus values for each of $I_{rt}(\lambda)$ and $J_r(\lambda)$, which can then be weighted according to the transparency factor to generate the tristimulus values of the viewed signal.

Reflections from the internal surfaces of the transparent sheet also occur, both of the illumination component which can enter obliquely, and of the filtered reflected component which though normal to the transparency, could cause interference. These internal reflections give rise to diffusion within the sheet and have not been included in the model. Scattering of the viewed reflected component within the sheet has also not been modelled.

6.4 Application to real data sets and discussion of models

Figures 6.6(d), (e) and (f) shows examples of applying the reflectance and pigmentation models to the real data variables of figures 6.6(a) and (b). In each case the first data variable ($D1$) is in fact the topography of the area. The second ($D2$, depicting the variation in magnetic field strength) is displayed in a two pigment representation which ranges from dark blue (low strength) to light yellow (high strength). The pigments were simulated as enamel paint, and a single illumination source (CIE standard illuminant C) was used. Figure 6.6(d) shows a representation using a purely diffuse reflection, while figures 6.6(e) and (f) show inclusion of the specular component. In 6.6(f) a white (the colour of the illumination) specular component was used, with the resulting loss of colour (*D*2) information in specular highlights. In 6.6(e) a coloured specular component identical in spectral distribution to the diffuse component was used. The base pigment range (before lightness modulation) used in these figures is shown in figure 6.9(e)-1. The discrete elements are the colours obtained from tristimulus components taken directly from the sequence look-up table; data-value dependent interpolation between these elements gives smooth colour variation. The reflectance model scene parameters used for these images are summarised in table 6.1; the effects of varying these parameters are discussed in section 6.4.2.

6.4.1 The effects of data smoothness on surface depiction

If the data represented by the surface topography is quantised, as is often the case with digital image data, step changes in surface orientation result, producing an apparent terracing effect in the viewed image. This is shown in the upper half of figure 6.7(a). This terracing does not necessarily unduly detract from surface comprehension, but if the second data variable has a spatial variation which might in some way be confused with the terraced contours, correct data appreciation can suffer. The effect can also be distracting if a specular component is present. On first considerations it would seem better to have a height data image which is not quantised, but rather smooth enough that step changes in the first derivative are not visually resolvable. However, as can be seen from the lower half of figure 6.7(a) which shows the same data in real-valued smooth form, this can have undesirable and in fact unrealistic effects. Perfectly smooth curved surfaces are seldom encountered in natural scenes; usually some surface roughness or texture is present, and such variations help greatly in topography appreciation. In the developed model, using a perfectly smooth surface can result in large areas of specular highlight, effectively destroying surface colour in these areas if the specular component does not have the same spectral distribution as the diffuse component.

Table 6.1

Image		Representation												
		Surface representation parameters								Surface colouring			Transparency colouring	
		Reflection components					Geometry							
	Data	ratios		contrast range				roughness		Data	refl.fns		Data	tr.fns
		a/d	d/s	$\%a$	$\%d$	$\%s$	h	2θ	m		$r_d(\lambda)$	$r_s(\lambda)$		$t(\lambda)$
6.6(d)	D1	0.25	$(s=0)$	20	80	0	1	64°	-	D2	BY-1	-	-	-
6.6(e)	D1	0.25	1	12	44	44	1	64°	0.4	D2	BY-1	BY-1	-	-
6.6(f)	D1	0.25	1	12	44	44	1	64°	0.4	D2	BY-1	U	-	-
6.7(a)upper	D1	0.25	1	12	44	44	1	64°	0.4	D2	BY-1	U	-	-
6.7(a)lower	D1	0.25	1	12	44	44	1	64°	0.4	D2	BY-1	U	-	-
6.7(b)upper	D1	01	1	4	48	48	1	64°	0.4	D2	BY-1	U	-	-
6.7(b)lower	D1	0.4	1	16	42	42	1	64°	0.4	D2	BY-1	U	-	-
6.7(c)upper	D1	0.25	2.33	15	60	25	1	64°	0.4	D2	BY-1	U	-	-
6.7(c)lower	D1	0.25	0.43	7	28	65	1	64°	0.4	D2	BY-1	U	-	-
6.7(d)upper	D1	0.25	1	12	44	44	5	64°	0.4	D2	BY-1	U	-	-
6.7(d)lower	D1	0.25	1	12	44	44	0.2	64°	0.4	D2	BY-1	U	-	-
6.7(e)upper	D1	0.25	1	12	44	44	1	77°	0.4	D2	BY-1	U	-	-
6.7(e)lower	D1	0.25	1	12	44	44	1	38°	0.4	D2	BY-1	U	-	-
6.7(f)upper	D1	0.25	1	12	44	44	1	64°	0.2	D2	BY-1	U	-	-
6.7(f)lower	D1	0.25	1	12	44	44	1	64°	0.6	D2	BY-1	U	-	-
6.8(a)	D1	0.25	1	12	44	44	1	64°	0.4	D2	GY-2	U	-	-
6.8(b)	D1	0.25	1	12	44	44	1	64°	0.4	D2	BY-3	U	-	-
6.8(c)	D1	0.25	1	12	44	44	1	64°	0.4	Q(D2)	BY-1	U	-	-
6.8(d)	D2	0.25	1	12	44	44	1	64°	0.4	D1	RC-4	U	-	-
6.8(e)	D1	0.25	1	12	44	44	1	64°	0.4	D1	RC-4	U	-	-
6.8(f)	D2	0.25	1	12	44	44	1	64°	0.4	D2	BY-1	U	-	-
6.9(a)	D1	0.25	1	12	44	44	1	64°	0.4	-	G	U	D2	BY-1
6.9(b)	D1	0.25	1	12	44	44	1	64°	0.4	-	G	U	D2	BO-5
6.9(c)	D1	0.25	1	12	44	44	1	64°	0.4	-	G	U	Q(D2)	BO-5
6.9(d)	D1	0.25	1	12	44	44	1	64°	0.4	D2	BY-1	U	Graph	Red

Notes

1. For parameter definitions, see text and figures 6.1 and 6.4.
2. Pigment ranges specified (for reflectance and transmittance functions) correspond to those in figure 6.9(e).
3. A uniform specular reflectance function (U in table) corresponds to a perfectly reflecting surface, resulting in a specular component with the spectral distribution of the illumination.
4. Q(D2) implies that D2 has been quantised.

Table 6.1 Model parameters for the images in figures 6.6 to 6.9

Image descriptions for the following 4 colour plates

Figure 6.6 Application of the coloured surface model to real data.

(a) Data set $D1$, representing the surface height of an area.

(b) Data set $D2$, representing the magnetic field strength measured over the same area.

(c) Conventional colour composite of $D1$ (cyan) and $D2$ (red).

(d) Coloured surface representation of $D1$ (as surface topography) and $D2$ (as surface colour, in a blue-yellow pigment range) using a diffuse reflected component only.

(e) Coloured surface representation of $D1$ (as surface topography) and $D2$ (as surface colour, in a blue-yellow pigment range) using equally contributing diffuse and specular components. The specular component is of the colour of the surface material.

(f) Coloured surface representation of $D1$ (as surface topography) and $D2$ (as surface colour, in a blue-yellow pigment range) using equally contributing diffuse and specular components. The specular component is of the colour of the illumination.

Figure 6.7 Effects of coloured surface model parameter variations. Variations of model parameters to either side of those used in figure 6.6(f) are shown. Table 6.1 gives the actual parameter values used.

(a) The effects of height resolution in the data set represented by surface topography: terracing effects caused by quantisation (upper half) and glare resulting from highly smooth data (lower half).

(b) The effects of decreasing (upper) and increasing (lower) the ambient reflected component magnitude.

(c) The effects of increasing (upper) and decreasing (lower) the diffuse/specular component magnitude ratio.

(d) The effects of increasing (upper) and decreasing (lower) the surface height emphasis.

(e) The effects of increasing (upper) and decreasing (lower) the illumination/viewing angle.

(f) The effects of decreasing (upper) and increasing (lower) the surface micro-roughness.

Figure 6.8 Application of the coloured surface model to real data - pigmentation variations and application to non-topographic and single data variables.

(a) The effect of using a single hue pigment range.

(b) The effect of using an isoluminant hue range.

(c) The effect of quantising the surface colour variable (*D*2) to 4 levels.

(d) Coloured surface representation of *D*2 (as surface topography) and *D*1 (as surface colour, in a red-cyan pigment range).

(e) Coloured surface representation of *D*1 (as surface topography and surface colour, using a red-cyan pigment range).

(f) Coloured surface representation of *D*2 (as surface topography and surface colour, using a blue-yellow pigment range).

Figure 6.9 Application of transparency model to real data, and pigmentation colour ranges.

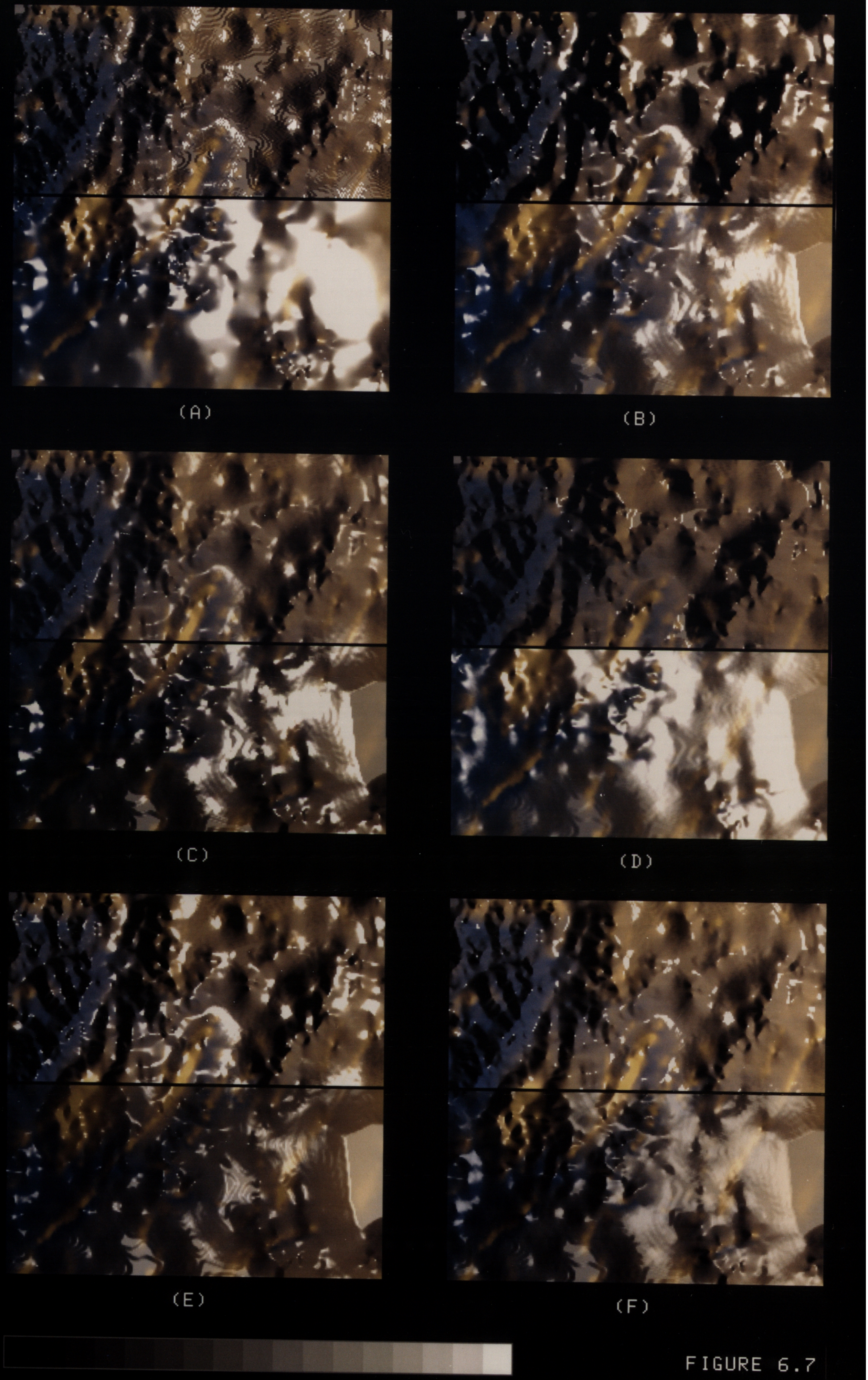
(a) Coloured transparency representation of *D*2 (in a blue-yellow pigment range) overlaid on a grey surface representing *D*1 (as surface topography).

(b) Coloured transparency representation of *D*2 (in a blue-orange pigment range; see text) overlaid on a grey surface representing *D*1 (as surface topography).

(c) Border-delineated coloured transparency representation of a 4-level quantisation of *D*2 (in a blue-orange pigment range) overlaid on a grey surface representing *D*1 (as surface topography).

(d) Low density transparent overlay (in red) of graphic data on a coloured surface representation of *D*1 (as surface topography) and *D*2 (as surface colour, using a blue-yellow pigment range).

(e) 20-level pigment ranges:
(1) blue-yellow with incorporated lightness increase;
(2) grey-yellow with incorporated lightness increase;
(3) blue-yellow isoluminant;
(4) red-cyan with incorporated lightness increase;
(5) blue-orange with incorporated lightness increase.



(A)

(B)

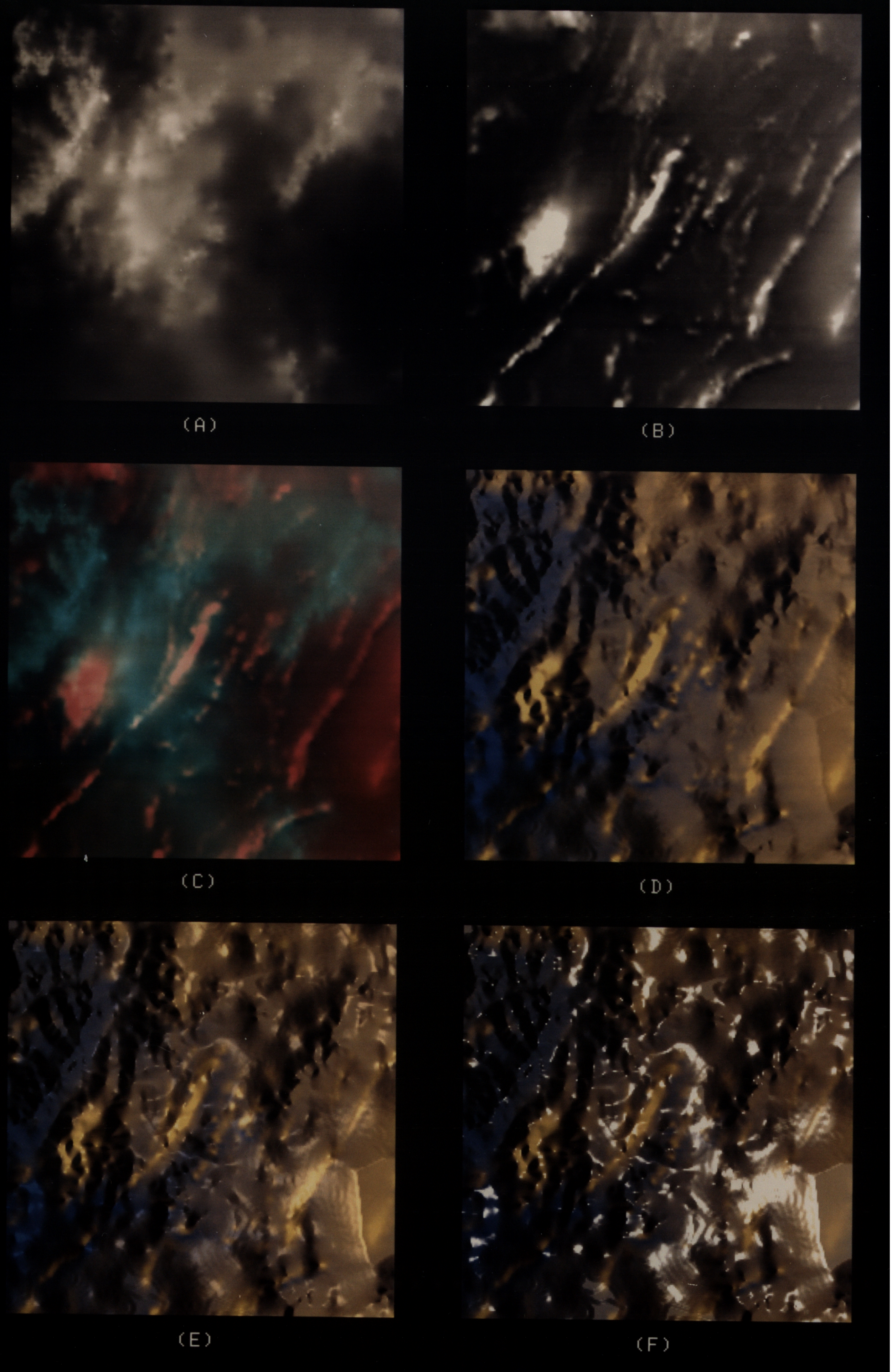
(C)

(D)

(E)

(F)

FIGURE 6.7



(A)

(B)

(C)

(D)

(E)

(F)

FIGURE 6.6

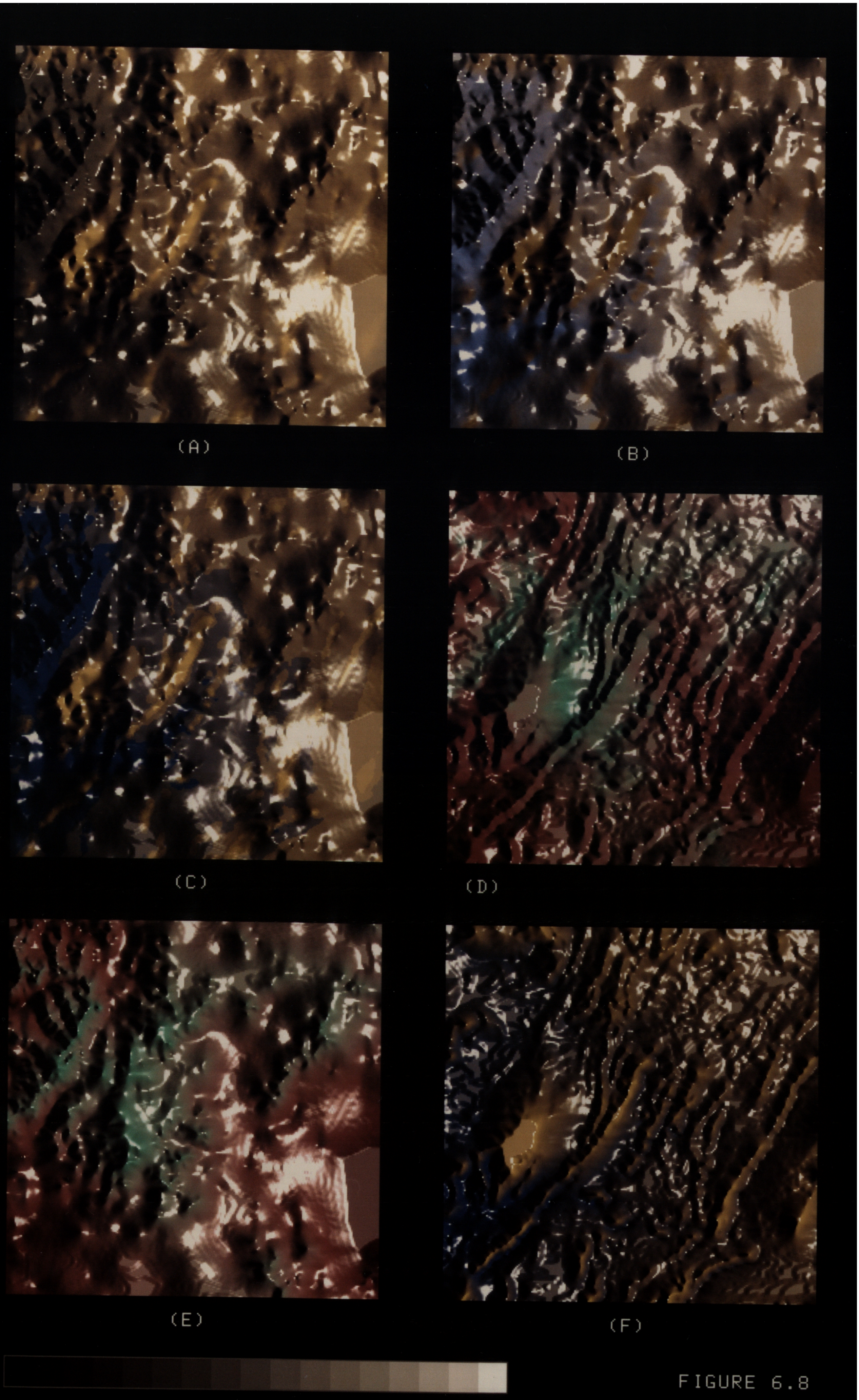


FIGURE 6.8

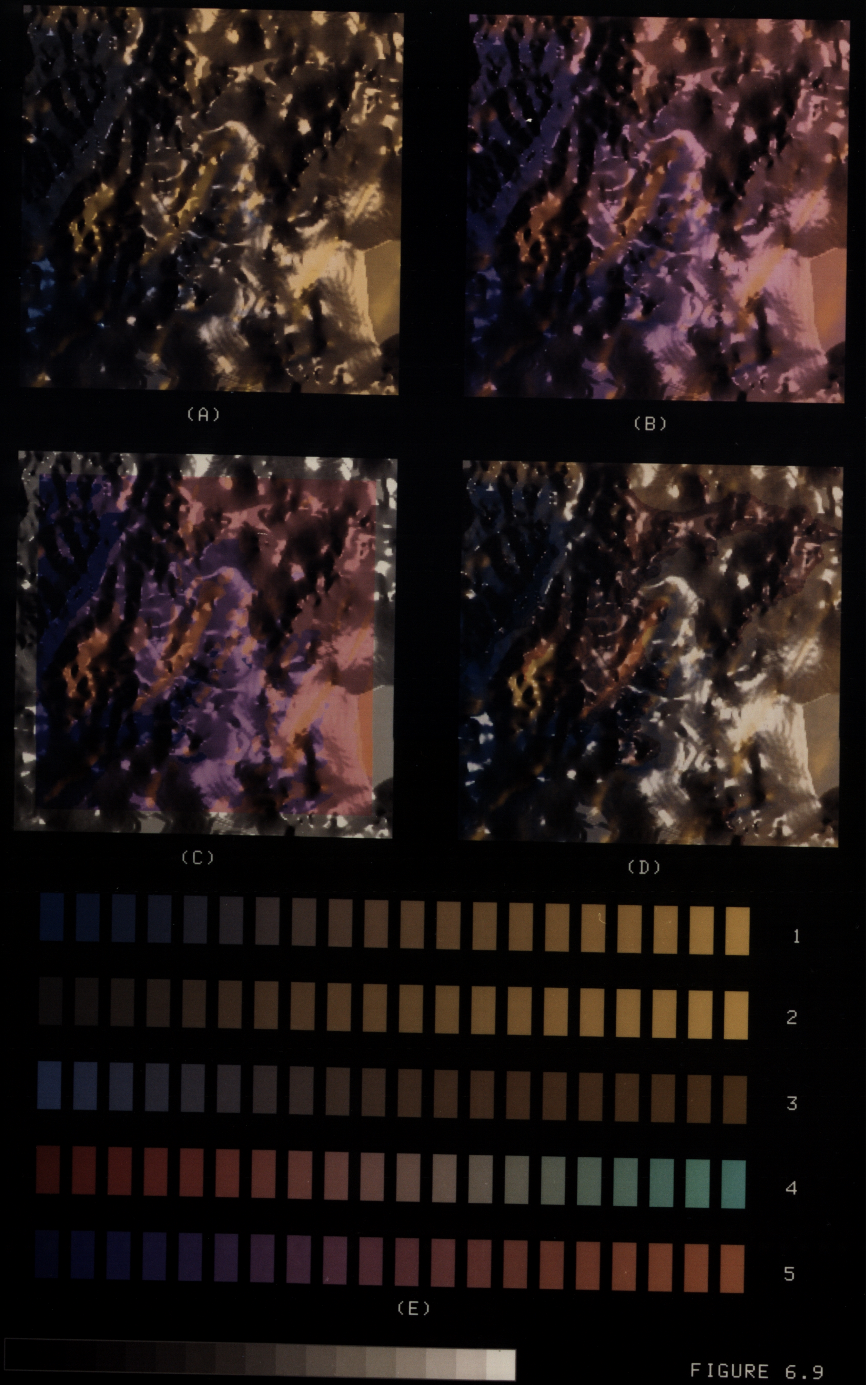


FIGURE 6.9

Two approaches to achieving an appropriate level of smoothness for the surface height data can be taken. The first, which has been used in the illustrations in this chapter, is to partially smooth the 8-bit quantised data (or the gradient if an intermediate gradient image is used). Various approaches to smoothing can be taken; most simply, it can be done by applying a simple 3×3 box filter, which replaces each pixel value with the average value of the $3x3$ area centred on the pixel, and representing the data in either 16-bit integer, or real-valued, form. This results in a surface which is sufficiently smooth to give reasonable surface representation while retaining enough local variation to break up specular patches. (Figures 6.7(d), (e) and (f), and subsequent figures in this chapter, were produced using this smoothed 8-bit data.) An alternative is to derive a smoother gradient estimate from the surface height data by using height information appropriately weighted from a larger sample area. The nature of the data can determine which of these approaches is most suitable.

6.4.2 Reflectance model - influence of scene parameters on the local and global comprehension of each data variable

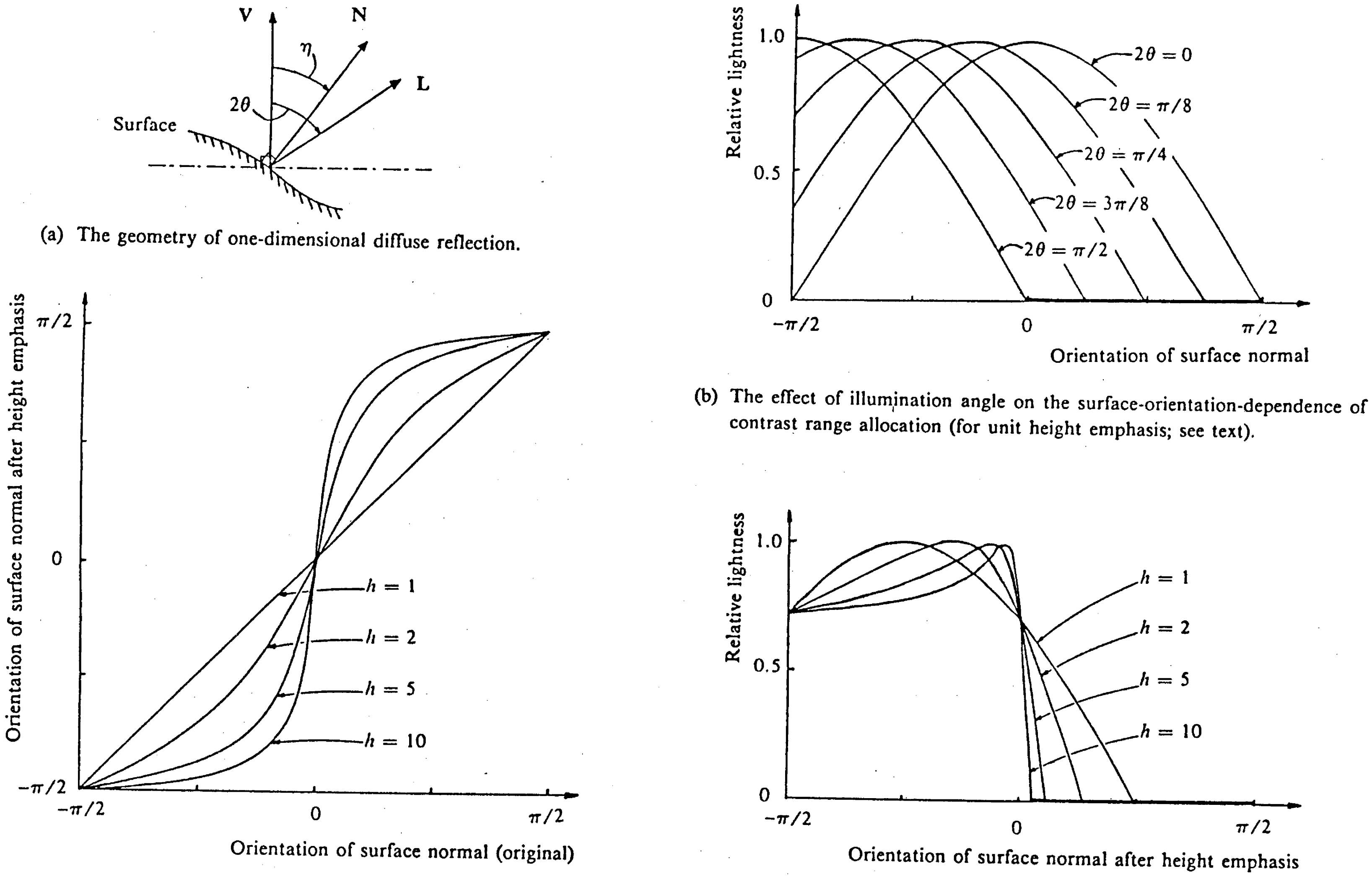
Realistic portrayal of a coloured surface depends on choosing the relative proportions of diffuse and specular reflected components appropriate for the surface material as characterised by its roughness, reflecting properties, and pigmentation. In the implemented model, each of these parameters can be varied to cater for variations in presentation. When choosing a material to simulate, it is important to ensure that the required balance of diffuse and specular components not only best depicts the surface form, or topography, but also leaves the colour of the surface at any point still apparent. Figures 6.7(b) to 6.7(f) show the effects of model parameter variations on a chosen standard image (6.6(f)). In each split image, one parameter only is varied to either side of the value chosen for the standard. These effects are discussed in the following sub-sections.

Roles of ambient, diffuse and specular reflected components

In general, appreciation of detail in the data variable represented by the surface topography depends on controlling the lightness contrast due to the diffuse reflected component from the surface. Introducing specularity can add realism and help in the global comprehension of the surface, but the diffuse component is still important in depicting local variations. Too much specularity causes a high contrast between specular highlights and their neighbourhoods, and hence reduces the contrast range available for the diffuse component. This is because the overall contrast range is limited, and depends on the particular display device being used. In addition, increasing the ambient component reduces the contrast range available to the diffuse and specular components since it has the effect of raising the base illumination level of the image. Too low an ambient component results in a loss of colour information in the darker areas, and hence loss of information of the second data variable. Figure 6.7(b) shows the effect of decreasing and increasing the ambient component (as a proportion of the diffuse component), while 6.7(c) shows the effect of varying the specular/diffuse component ratio. Table 6.1 gives the actual values used. It was found in practice that a diffuse component accounting for about 45% of the available contrast range gave satisfactory depiction of the surface topography, while an ambient component raising the background lightness level to about 10% of the maximum was required to render the surface colour always visible. The remaining 45% of the available contrast range is then available for the specular component.

Effects of height emphasis and illumination angle on topography depiction

The depiction of the topography by the diffuse component is also dependent on the illumination angle θ and the height emphasis h used. Neglecting the effect of shadowing, figure 6.10(b) shows in one dimension the effect of varying illumination angle for a fixed (unit) height emphasis. The geometry of this one-dimensional projection is shown in figure 6.10(a).



(a) The geometry of one-dimensional diffuse reflection.

(b) The effect of illumination angle on the surface-orientation-dependence of contrast range allocation (for unit height emphasis; see text).

(c) The effect of height emphasis on surface orientation (for an illumination angle of 45°).

(d) The resulting effect of height emphasis on the surface-orientation-dependence of contrast range allocation (for an illumination angle of 45°).

Figure 6.10 The effects of height emphasis and illumination angle on topography depiction.

The relationship between perceived reflected intensity (or surface lightness) and surface normal orientation η is given, for an ideal Lambertian surface, by

$$I = k\cos(2\theta - \eta) \quad for\ |\theta| < \pi/2\ and\ |2\theta - \eta| < \pi/2,$$

and $I = 0$ for surface orientations or viewing/illumination geometry not satisfying these conditions (k is a normalising constant). The illumination angle θ is taken as half that between the illumination and viewing directions, consistent with the geometry of figure 6.1. Surface orientation η is given with respect to viewing direction. Height emphasis h is taken as the absolute value of surface gradient, which depends on the data value range. For practical purposes this value range was normalised to the square root of picture area as measured in resolution elements. As can be seen from figure 6.10(b), the choice of illumination angle affects the portion of the surface on which surface topography will be portrayed. For the best appreciation of detail in local variations at any chosen surface orientation, the illumination angle producing curves with the greatest gradient at that point is most appropriate. Thus if it is important to render a particular aspect of the data, such as ridge peaks and valley bottoms, most clearly, the illumination angle corresponding to curves with the greatest lightness range in the required area can be chosen. On the other hand, if all surface detail is to be portrayed, direct overhead illumination ($\theta = 0$) must be used. Global surface comprehension is enhanced by increasing the illumination angle.

The effect of varying height emphasis on surface orientation is shown in figure 6.10(c). Height emphasis results in a change of surface orientation given by

$$\eta' = \tan^{-1}(h\tan\eta),$$

where η' is the modified surface orientation. In figure 6.10(c), the surface orientation transfer functions resulting from height-emphasis factors of 2, 5 and 10 are shown, each for an illumination angle of 45°.

Thus the modified perceived lightness I' is given by

$$I' = k\cos\{2\theta - \tan^{-1}(h\tan\eta)\},$$

and the result of this, for each height emphasis factor, is shown in figure 6.10(d), again for a fixed illumination angle of 45°. Fractional height emphasis corresponds to a reflection about the unit height-emphasis axis. The penalty of increasing the height emphasis is that when non-overhead illumination is used, a greater range of surface orientations is left dark; the advantage seems to be that global, or larger scale, surface structure is more easily appreciated. A given lightness does not necessarily give a unique surface orientation, but it appears that the visual system is able to use local smoothness and continuity properties to extract shape from lightness variations (see also Pentland, 1984). This might well impose smoothness and continuity constraints on a data variable if it is to be represented by surface topography. It should also be pointed out that, as shown by Horn (1981), a different surface reflectance model might result in a better depiction of surface topography at any given detail level. In figure 6.7(d) the effect of varying height emphasis is shown, while 6.7(e) shows the effect of varying the illumination angle. As before, exact parameter values are given in table 6.1.

Effects of surface micro-roughness and specular distribution shape

The specular directional distribution can be characterised by a distribution function which depends on the root-mean-square slope (m) of the microfacets of which the surface is modelled (Torrance and Sparrow, 1967; Blinn, 1977). Cook and Torrance suggest that the Beckmann distribution is suitable for the modelling of both metallic and non-metallic surfaces, and that it is reasonable to ignore the wavelength dependence of this distribution function. This distribution function has the form

$$D = \frac{e^{-[(\tan\alpha)/m]^2}}{m^2\cos^4\alpha},$$

where α is as before the angle between the surface normal and the viewing/illumination angle bisector. Smooth surfaces, with small mean facet-slope angles, produce highly directional distributions, while rough surfaces, with large values of m, produce directional distributions closer to a diffuse distribution. The effect of varying surface roughness is shown in figure 6.7(f); values of $m=0.3$ and $m=0.7$ were used in the upper and lower images respectively.

Using a high roughness value can have the disadvantage that while for a narrow distribution some saturation of specular highlights can be tolerated, for a wide distribution such saturation results in a loss of detail in a larger area. This means that the dynamic range of the diffuse component must be reduced to avoid such saturation. As pointed out by Cook and Torrance, realistic surfaces can be modelled by a combination of several surface roughnesses; this appears to be an unnecessary refinement to the current use of the model.

Subsequent suitability of materials for surface simulation

In general, a suitable surface material has a substantial diffuse reflected component, some specularity, and some ambient illumination. Highly specular metallic surfaces are unsuitable for this type of data representation since they have a very low diffuse reflected component. Smooth plastics have a highly directional specular component (which is close to the colour of the illumination) reflected from the substrate surface; they also have a substantial diffuse component which results from scattering within the pigment of illumination penetrating the surface. In rougher plastic or plastic-like materials, the specular component can be wider, and the colour of it can be closer to that caused by the pigment. In fact, provided that the relative proportions of the ambient, diffuse and specular components required to satisfactorily depict each data variable are maintained, and the surface roughness parameter *m* is kept within the range of about 0.2 to 0.8, the actual material simulated is not crucial to data appreciation. For given specific data variables, possibly with particular spatial frequency content, surface roughness and specular reflectance properties can be chosen to avoid possible masking of data variations.

6.4.3 Surface colour pigmentation model

Effects of pigmentation model on surface property depiction

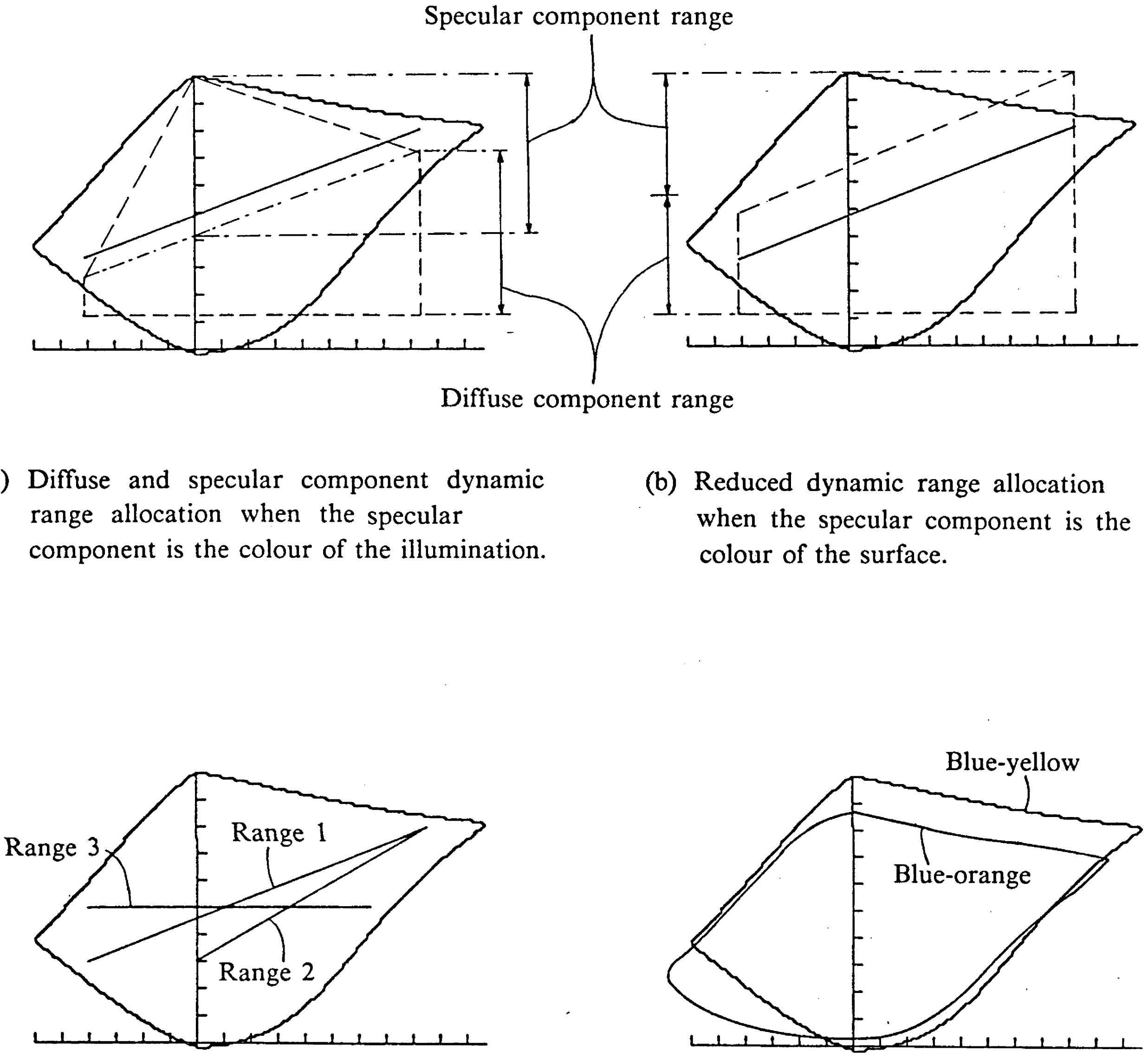
Global appreciation of the characteristics of the data variable represented by surface colour is substantially enhanced by using a colour range encompassing two major hues, rather than one. This technique effectively gives a rough dichotomy into "highs" and "lows", with the saturation providing localised relative information. Gradients in the data are also emphasised by this technique; this can be particularly useful in the interpretation of geophysical data. For comparison, figure 6.8(a) shows an image produced using a colour range of a single hue; the pigment sequence used is shown in figure 6.9(e)-2.

As mentioned earlier, incorporating a lightness difference between the two hues helps to emphasise rapidly (spatially) changing variations in colour. It is well known (Evans, 1974; Frome et al., 1981; Wolfe and Owens, 1981) that chromatic borders, which would occur under conditions of steep gradient in the surface property and smooth surface topography, are difficult to perceive distinctly if isoluminant; the lightness difference overcomes this problem. It is possible that a lightness difference used in this way could be misinterpreted as being caused by surface topography variations, leading to confusion of the two data variables. In general, the smoothness and spatial continuity of each variable should prevent such misinterpretation, but in cases of localised high spatial correlation between data variables, this could occur. In such cases either an isoluminant pigment variation can be used, or some additional perceptual cue introduced to aid interpretation. Figure 6.8(b) shows the result of using an isoluminant pigment variation; this pigment sequence is shown in figure 6.9(e)-3. Again the reference image (in which the pigment range incorporates a lightness variation) is that of figure 6.6(f).

Coarse quantisation can assist in portraying the actual scalar values of the data variable being represented by surface colour. This is analagous to step changes in pigment (mixture) density, and involves a loss of scalar resolution; however, if interpretation of the actual data values at any point, rather than the trends or variations in relative value, is required, quantisation might improve the display. The quantisation also effectively contours the data, and this can aid the process of separating the two variables. Figure 6.8(c) shows an image produced from a 4-level quantisation of the surface property data variable; again, figure 6.6(f) provides the reference for comparison.

Colour gamut considerations

If a material with a coloured specular component is simulated, surface colour in specular highlights is not lost. However, there is also a practical disadvantage to using a surface in which specular highlights are coloured. When choosing an appropriate pigment lightness range, the path extent in UCS was determined by the lightness modulation range required for surface topography depiction by the diffuse component. When the specular highlights are the colour of the illumination, or the reference white, effective utilisation of the available colour gamut can be made; this is illustrated in figure 6.11(a). It should be noted that the upper utilisation boundary lines, which in the illustration are shown as straight, are not always straight; rather they depend on the relationship between saturation and tristimulus values, which depends on the UCS used and also on the hue angle (chromatic axes are differently defined in the CIE spaces). In the illustration the ratio between the largest diffuse component dynamic range and the specular component dynamic range is unity, in keeping with the standard image of figure 6.6(f). Overall lightness normalisation results in the schematically shown allocation of ranges. On the other hand, if the specular highlights are the colour of the surface, the available dynamic range at the lightest pigment range point must be shared between the diffuse and specular components. Figure 6.11(b) illustrates the gamut shape utilisation in this case; even allowing some specular component saturation, the dynamic range of each component is less than three quarters of the range when white specular highlights are used. Comparison of figures 6.6(e) and 6.6(f) shows this difference in both dynamic range and apparent specularity; the specular and diffuse component ratios were the same for these images, but in 6.6(e) the specular component is partly saturated, and is consequently less well distinguished from the diffuse component.



(a) Diffuse and specular component dynamic range allocation when the specular component is the colour of the illumination.

(b) Reduced dynamic range allocation when the specular component is the colour of the surface.

(c) The effect of gamut shape on pigment range choice.

(d) Comparison of the cross-sections including the blue-yellow and blue-orange pigment ranges, showing the reduced lightness range available when a pigment range not incorporating the lightness axis is used.

Figure 6.11 Gamut shape and dynamic range considerations in the choice of pigment colour ranges

Effective utilisation of any particular colour gamut can lead to greater spectral variation, and hence data resolution, in an image. For example, the blue-yellow (1), the grey-yellow (2), and the isoluminant blue-yellow (3) pigment ranges are shown in the blue-yellow gamut cross-section in figure 6.11(c); clearly range 2 is not suited to the chosen gamut cross-section, while range 3 does not include highly saturated yellows. In fact the blue-yellow gamut cross-section was chosen because it allows effective utilisation for a pigment range from a dark colour to a light complementary colour. For an isoluminant pigment range, a different gamut cross-section would be more suitable.

Control over pigmentation colours

The method of generating a smoothly varying colour range between two chosen end points uses a path in UCS determined by the spectra of those end points. However, it can be desirable that a path should pass through a specific colour; often that colour is a neutral grey. If the two chosen end points are not approximately complementary colours (that is, on opposite sides of the grey axis in the UCS used) this can be achieved by using a piecewise path through UCS, breaking the path at the chosen point through which it is to pass. Provided that the hue angle change is not too sharp, sequence progression remains intuitive. Alternatively, the end points can be modified to cause the path to pass through the required grey.

In practice it was found useful to have two facilities for modifying the reflectance spectra of colours. One is to increase the saturation of a colour, keeping its hue constant, while the other is to change the hue of a colour by combining specified proportions of two spectra. Provided an initial base set of spectra are available, the two techniques can be used to produce most required colours. The essential feature of a pigmenting process, that of having a smooth spectral reflectance distribution, must be maintained. Increasing the saturation can be performed by emphasising the reflectance distribution values with respect to some chosen level at all wavelengths; values below the reference level are reduced in proportion to the modulus of the difference, while those above it are similarly increased.

Hue modification is similarly straightforward, with chosen fractions of each of two spectra being simply added together to form the new colour. Simple iteration can be used to cause a path to pass as close to any point as might be required.

An alternative approach to modifying existing reflectance spectra is to artificially generate spectra from tristimulus values. This can be done by using some numerical means of generating smooth spectra which, when weighted with the illumination distribution and Standard Observer matching functions and integrated, generate the required values. In general, such a process will generate a spectral distribution for a specified colour different from that which would be measured from a sample of that colour; the colours produced would be metameric to a human observer. Consequently slightly different sequence paths will be generated between end points derived from measured, or thus constructed, spectra. In practice the method of using a base set of measured paint spectra, and combining or modifying these spectra as described in the previous paragraph, was found to involve substantially less computation, and was hence used in the generation of pigmenting sequences. The simple saturation modification technique was useful to most effectively utilise the available colour gamut of any particular display device.

6.4.4 Application of surface representation to non-topographic data variables

While in the examples so far shown the relief-depicted data variable has been terrain height, the approach can also be used to display other kinds of data variables. Figure 6.8(d) shows an image made using the magnetic field strength data to form the surface topography, and the height data to colour the surface. A red-cyan pigment range was used in this representation; this pigment range is shown in figure 6.9(e)-4.

The technique of representation in the form of a coloured surface can also be used to advantage when displaying only one image data variable, representing it by both the surface topography and the surface colour. A shaded-relief image gives the viewer information about the topographic form, or surface gradient, at any point in the image, from which the visual system deduces the surface height relative to height in the neighbourhood of the point. Absolute height, or actual scalar value, information can be provided by colouring the surface, using a suitably chosen colour range, according to its value. Again a colour range encompassing a change in hue emphasises steep gradient. The overall result is a better and more intuitive understanding of the scalar and spatial properties of the data variable, in both local and global senses. Thus all three of Bertin's levels for appreciating a data variable are achieved (see Chapter 4). Figures 6.8(e) and (f) show examples of images representing a single data variable (the height data and the magnetic field strength data respectively) in this way. Red-cyan and blue-yellow colour ranges respectively represent the scalar value range.

6.4.5 Overlaid transparency model

Model parameter requirements for data depiction

For simulating an overlaid transparency a slightly different set of considerations apply. Local topography depiction of the underlying surface still depends on the diffuse reflected component, but the specular highlights are filtered by the overlaid transparency, and so no longer appear the colour of the illumination. This means that information from the second data variable is not lost in areas of specular highlighting. Hence if required, a material with a wide specular directional distribution can be used. The earlier mentioned disadvantage of reduced diffuse component contrast due to saturation of specular highlights also applies to an overlaid transparency situation. Figure 6.9(a) shows an example of a coloured transparency overlaid on the surface depicting data variable $D1$. Data variable $D2$ determines the pigmentation of the transparency, again using the blue-yellow colour range.

Achieving the impression of transparency - perceptual considerations

The blue-yellow colour range passes through a neutral (grey) colour, and one implication of this is that saturated specular highlights, which in a chromatic region do not seriously detract from the impression of transparency, become white in a grey region (mid-values of *D*2) of the image. Because of the importance of white as a visual reference in a presented image (being representative of the illumination), the result is that the transparency loses its realism when specular highlights in grey regions occur; these areas no longer appear to be filtered by the overlaid transparency. Reducing the dynamic range of the reflected components removes this effect, but also results in an image of low contrast. An alternative is to use a colour pigment range which does not pass through grey, such as the blue-orange range shown in figure 6.9(e)-5; the image of figure 6.9(b) shows the result of using such a pigment range. Depending on the display device being used, such a colour range can result in reduced lightness resolution. This is the case for the colour film/print gamut; figure 6.11(d) shows a comparison between the cross-sections used in the blue-yellow (1) and blue-orange (2) colour ranges; the blue-orange range has a lightness range about 0.8 of that of the blue-yellow range. A path which does not pass through the grey axis also incorporates a hue change which may be less amenable to intuitive interpretation than a path which varies only in saturation and lightness (see Chapter 4).

Additional perceptual cues can also help to emphasise the transparency effect. Figural delineation of the transparency border, for example by making it distinct from that of the underlying surface, enhances the impression of an overlaid transparency (Beck and Prazdny, 1983). Contouring, or coarsely quantising, the second data variable can also aid the process of dissociating the transparency, as an artificially overlaid structure, from the surface below it. Figure 6.9(c) shows the combined effects of coarse quantisation (to 4 levels) and figural delineation on the overlaid transparency simulation.

Extension of the transparency model to include transparency opacity (section 6.3.2) did not appreciably enhance the achieved sense of realism; rather the result was to make the underlying topography less clear. This might be due to the lack of a spatially adjacent reference to indicate the nature of the image without a transparency imposed. Because of the loss of topographic detail entailed, and because figural delineation appeared to be a more effective way of enhancing the impression of transparency, this form of extended modelling was not pursued.

Representation of additional variables - graphic overlay

The use of an overlaid transparency suggests that it might be possible to introduce a third data variable by overlaying it as a pigmented transparent sheet over an already coloured surface. In practice this generates a colouring effect which is confused and very difficult to separate, posing essentially the same problem as that of simple spatial superimposition in two colours. However, such a technique can be used to overlay a third data set of limited spatial variation, with certain restrictions on the size of contiguous pigmented area, and pigment density, on the transparency. For example, isolated features such as text or arrows, perhaps depicting some relevant directional variable, can be overlaid in a pigment of low density and of a colour contrasting with that of the surface below. This is essentially a graphic overlay process, but one which preserves the underlying information to an extent dependent on the density of the transparency pigmentation and on the continuity of the structure in the underlying data variable. Figure 6.9(d) shows an example of such an overlay; in this case the higher-valued areas of data variable $D1$ have been overlaid in red on the blue-yellow surface of figure 6.6(f).

Displaying three full variables with different spatial structures in one image would seem to require an alternative approach, or incorporation of an additional set of perceptual cues. One naturally occurring variable which might be so exploited is that of texture variation. Whilst texture variation is usually closely associated with material type, and hence hue attributes, it might be possible to use a correlated variation in texture and hue to depict additional information. However, it seems likely that it would require the use of additional techniques such as viewing in perspective or stereo allowing a more comprehensive use of implied three-dimensionality, or incorporation of a temporal variation, to satisfactorily depict more than two variables with full spatial variation. Such approaches result in an image less amenable to direct use. Not only do they have the disadvantage of being costly in computation time, but also they are visually more complex; the simple overhead viewing of data on a rectangular grid from which direct measurements can be taken, or line overlays can be made, is lost.

6.4.6 Limitations and possible extensions of models

Data smoothness requirements

Requiring that one data variable can represent the height of a realistic surface restricts the types of data which can be displayed. While the surface topography representation does not preclude the use of a surface to represent a variable which is not naturally seen in the form of a surface, creating a realistic surface imposes continuity and smoothness constraints. Arbitrary data variables may not be sufficiently smoothly varying to satisfy these constraints. Thus if neither of two data variables to be superimposed satisfies the requirements for surface representation, some form of pre-processing would be necessary for the technique to be applicable. In fact, as described in section 6.4.1, such pre-processing may be necessary even for data which does actually represent surface height.

Dangers of misrepresentation due to processing artifacts

One problem with any technique which modifies the original data in any way is the danger of introducing artifacts of the processing technique which are seen as real data characteristics in the resulting image. A relief-shading model, for example, corresponds to a directionally dependent emphasis in the spatial frequency domain, and if the model is in some way deficient for computational or conceptual reasons, the real-world observed effect of relief-shading will not be simulated. Instead, the erroneous modelling process might cause the data to be incorrectly interpreted. This danger is particularly pertinent when spectral shifts are introduced in a scene in which the colour attributes represent the data values. However, due to the chosen overhead viewing geometry of the surface representation used, such spectral shifts are small.

In that respect, the reflectance model used to make a coloured surface look realistic is possibly unnecessarily sophisticated, the spectral shifts predicted by the Fresnel equation being significant only when the illumination and viewing vectors come close to grazing the surface. However in practice, as pointed out earlier, bypassing the calculation of this spectral shift produces a relatively insignificant saving in computation time, since it is performed in a pre-calculated section of the processing.

Potential for enhancing realism by extending models

There is some potential for extending and improving upon these approaches. First, inclusion of additional effects which enhance the sense of realism, such as shadowing (Crow, 1977; Williams, 1978), could be made. Second, the use of models of surface roughness or texture (Blinn and Newell, 1976; Blinn, 1978,1982; Schweitzer, 1983; Haruyama and Barsky, 1984), either to provide a more absolute height indication (for example by reducing texture cell size with height), or to assist in depiction of the second data variable by introducing a correlated variation of texture with colour, might improve the representation. Third, allowing the ambient illumination component to have a spectral distribution different from that of the illumination source, such as occurs in sun-lit blue sky scenes, might enhance the realism of scenes intended to look natural. Implementation would be straightforward, requiring only an additional tristimulus component look-up table for the ambient component, rather than including it with the diffuse reflected component.

Specific drawbacks encountered in the presented scenes can also be overcome by using the full potential of the developed models. For example, the loss of surface topography detail in areas orientated away from the illumination source can be treated by applying more sophisticated lighting conditions such as using two sources, or by modelling the ambient component more realistically to give it a non-uniform directional distribution. The use of reflectance spectra and surface roughness values appropriate for simulating natural materials might also be advantageous for representing some types of data. Cook and Torrance (1982) or Wyszecki and Stiles (1967) should be consulted for further details on the reflectance characteristics of natural materials.

6.5 Conclusions

In judging the success or otherwise of the developed technique, several factors must be considered. The first is whether the image produced can be intuitively appreciated by the visual system. If it can, the question is then whether the chosen natural scene properties can be appreciated separately, and further, whether the representation for each data variable in the form of its chosen natural property is suitable.

The images of coloured surfaces produced are clearly recognisable as such, the sense of realism being enhanced by the application of colour graphics methods. Thus the first criterion is satisfied. It also appears that both the topography of the surface, and its colouring, are clearly portrayed, and can be appreciated separately. What is less easy to judge is whether each of these natural scene properties adequately portrays the data variables.

The scalar value of the first data variable is represented as the height of the surface above some base level, but it is in fact the surface gradient that is depicted by the relief-shading technique. This corresponds to the first derivative of the scalar variable. Height relative to its neighbourhood is deduced by the visual system, but absolute height, which represents the data variable value, is not necessarily extracted.

The scalar value of the second data variable is directly represented by colour over a fixed range. This is a representation similar to that of a pseudo-colour univariate display, the salient aspects of which have been treated in Chapter 4. Choosing a colour range to model pigmentation should assist the intuitive process of relating the colour to the data variable scalar value. The results obtained indicate that this relation can in fact be achieved.

We initially posed the problem of achieving individual and joint appreciation of two superimposed data variables with arbitrary spatial variation. With the proviso that the local relative, rather than absolute, value of the first data variable is extracted, the coloured surface approach appears to provide a solution to this problem. Use of the reflectance model makes a significant difference to the realism of surface representation. Overlaying a coloured transparency on the surface achieves the required representation in a similar manner, though the initial step of visual comprehension of the scene is perhaps less intuitively attained.

With the aforementioned smoothness constraints on the data variable to represent the surface topography then, the developed techniques seem to be well suited to the display of two data variables with arbitrary spatial variation, with potential for inclusion of additional data with limited spatial variation. No restrictions on the smoothness or nature of the variable representing surface or transparency colour are necessary; both gradual and abrupt spatial variations are clearly depicted, and the visual system seems well able to interpret the chosen form of colouring if the underlying topography is clearly portrayed. The improvement in data variable comprehension achieved using the developed approach, over that achieved using conventional superimposition-type display methods, appears to be substantial.

Chapter 7
The effects of high spatial frequencies on perceived colour

This chapter considers the presence of high spatial frequencies in an image, and their effect on perceived colour. The chromatic spatial resolution of the human visual system is not as high as the achromatic spatial resolution, and possible ramifications of this in the perception of detail-rich imagery such as that from Landsat data are investigated. Also considered is whether knowledge of this effect can be used to develop techniques which under normal image viewing conditions cause the image information to be perceived more accurately. Our aim in this work is twofold: first, to establish whether the different contrast sensitivity characteristics of the visual chromatic and achromatic mechanisms might cause detrimental effects in image comprehension; second, to investigate whether such effects can be avoided by processing within the developed display framework.

7.1 Colour and spatial-frequency dependence of contrast sensitivity

This section looks at the spatial frequency characteristics of the low level visual pathway, and at techniques in image processing which have been based on, or have taken into account, these characteristics. We also specify a representation of opponent channels in the display framework, and in following sections investigate whether this representation provides an appropriate access to low visual levels.

7.1.1 Colour-opponent channels and their spatial frequency characteristics

In Chapter 2 we discussed the scope and limitations of various models of the human visual system, and also pointed out the dangers of attempting to be too specific in attributing high level (perceived) effects to low level (physiological) mechanisms. However, evidence that the visual signal exists in opponent form between the retina and LGN is substantial (see Gouras and Zrenner, 1981). The spatial frequency characteristics of the retinal colour mechanisms have been investigated by Green (1968) and Kelly (1974). Kelly used intense adaptation to supersaturate two of the three receptor mechanisms, and was thus able to measure the spatial and temporal response characteristics of the third. These results are explained in terms of cellular inhibition; it is clear that the role of inhibition is complex, and as suggested by Barlow (1981), probably still not well understood.

The limits in spatial resolution can be considered by looking at the contrast sensitivity functions of the opponent level signals, or opponent channels. Mullen (1985) has made a thorough and comprehensive investigation, which includes correction for optical aberrations, of these sensitivity functions. These results show that the presence of luminance components in earlier experiments (van der Horst and Bouman, 1969; Granger and Heurtley, 1973; Kelly, 1983) can seriously affect the measured contrast sensitivities, and that luminance matching of grating constituent components is dependent on spatial frequency. In fact the difficulty of estimating true chromatic drop-off is pointed out in these earlier studies; both van der Horst and Bouman, and Granger and Heurtley, describe the problem of dissociating a luminance component from a test chromatic grating.

The principal departures of Mullen's results from those performed earlier are that the red-green and blue-yellow mechanisms are shown to have very similar contrast sensitivity functions which drop off at lower spatial frequencies than previously estimated, and that both chromatic mechanisms have no appreciable low frequency drop-off. To fully dissociate the luminance component from the chromatic gratings, Mullen used an intensity match criterion which maximally separated the contrast sensitivities of the chromatic and a monochromatic luminance grating. This match was found to vary with spatial frequency, and did not necessarily correspond to a luminance match between the grating colours. Wide visual fields were used to allow estimates of low frequency sensitivities. The general shapes of the contrast sensitivity functions obtained by Mullen are shown in figure 7.1; the achromatic sensitivity function is shown in bold, the red-green chromatic function in broken line, and the blue-yellow chromatic function in dotted line. The significance of these shapes is discussed in later sections.

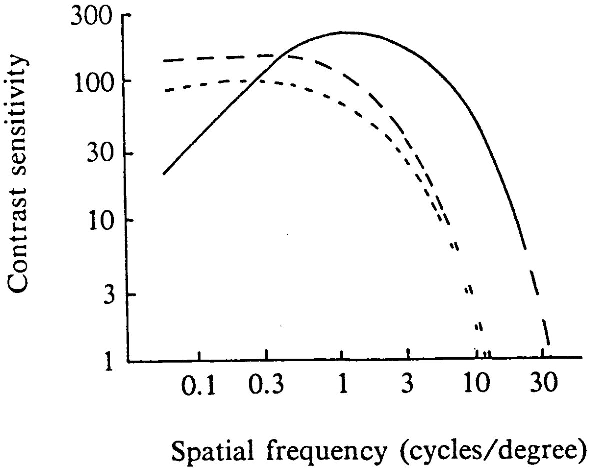


Figure 7.1 Contrast sensitivity functions for the achromatic and chromatic channels of the human visual system (from Mullen, 1985). The achromatic sensitivity function is shown in bold, the red-green chromatic channel in broken line, and the blue-yellow chromatic channel in dotted line. Note that the general form only of these functions is shown; the figure is not reproduced exactly. The original work should be consulted for full details.

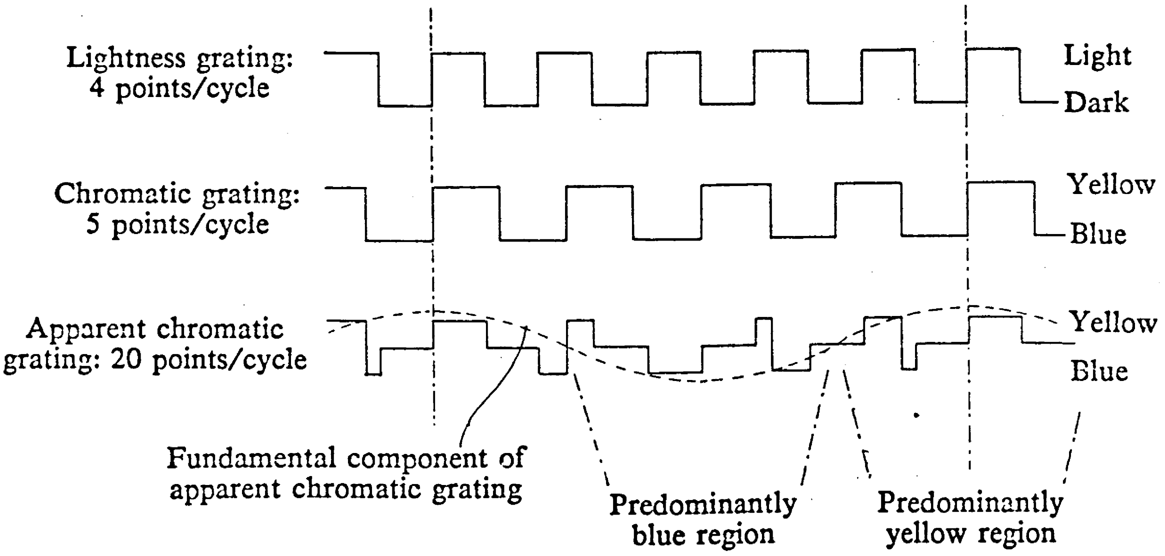


Figure 7.2 Schematic illustration of a grating interference effect.

It is not necessarily clear that the limiting factors in the contrast sensitivity of the visual system do actually occur at opponent channel stage. Recent neurophysiological investigations have commented on the difficulty of finding clear cut opponent type stages in the retinal layers, throwing doubts on the specificity of cone-to-opponent type formulations, such as in Faugeras' model. It could well be that the spatial response is limited at all stages of the process, the supposed opponent channel responses being a combination of filters at various stages. At the perceptual level the nett filtering effect could be seen as bandwidth limiting of the opponent channels due to the possibly more straightforward nature of the relationship between chromatic perceptual attributes (hue and saturation) and chromatic opponent channels, than that between signals which can be isolated at lower levels of the visual pathway. What is probably most important for the purposes of image processing is that the overall filtering effect can be specified in terms of its effect on perceptual attributes, so that its ramifications in the perception of these attributes in an image can be predicted.

7.1.2 Image processing techniques based on the opponent channel spatial frequency characteristics

Given that the achromatic and chromatic opponent mechanisms in the visual system have different contrast-sensitivity/spatial-frequency characteristics, the question of how this might be relevant to the display of information in colour arises.

In an image processing context, Faugeras has shown that potential for enhancing subjective image quality exists. Specifically, he used linear combinations of cone outputs to create opponent channels, one achromatic and two chromatic, treating these channels as forming a perceptual space. He derived the spatial frequency characteristics of these opponent channels from measurements of grating illusion (simultaneous contrast) cancelling effects at a number of spatial frequencies. He then performed image processing on these channels. In particular, he attentuated the slowly varying component as representing illumination properties and accentuated the rapidly varying component, which he treated as representing reflectance properties. This was designed to increase the saturation of smaller objects, and remove or reduce overall colour casts on an image. He also used the model as a perceptual rationale for image compression.

Hall and Andrews (1978) considered requirements for image coding bandwidth based on a visual system model (Hall and Hall, 1977) which similarly isolates an achromatic and two chromatic channels in its perceptual stage, using Green's (1968) results to specify their spatial frequency characteristics.

Faugeras did not consider the consequencies of his observed differential loss in visual resolution of the chromatic channels at high spatial frequencies, or the effect of this on image interpretation (as pointed out earlier, Mullen's (1985) results suggest that this differential loss is an experimental artifact). Neither Faugeras nor Hall and Andrews considered the possible disadvantages of having spatial frequency content above the level of maximum visual resolution in artificially created imagery. We consider these disadvantages in section 7.2, and in section 7.3 look at whether factors leading to such effects can be isolated within the developed framework.

7.1.3 Representation of opponent level signals within the display framework

In keeping with the emphasis in this work on the characterisation of visual effects at a perceptual level, we realise a representation of opponent channels directly within the UCS central to the display framework. The achromatic channel can be aligned with the UCS lightness axis, and the chromatic channels represented by lines of constant hue orthogonal to this axis. Saturation is hence interpreted as chromatic contrast. We do not necessarily expect these UCS axes to represent opponent channels accurately, but rather we use them initially as an approximate representation for a controlled means of investigation. Since we are looking at these phenomena qualitatively, the exactness of representation is unlikely to be critical. What is important is that the achromatic channel is isolated, and that a variation in differential chromatic channel response will give a variation in hue. The ramifications of Mullen's results are that we need only separate the achromatic and chromatic components; if this is performed sufficiently accurately, chromatic signals of all hues should have similar resolution limits. The accuracy with which this separation might be performed is considered in section 7.3.

7.2 The effects of differential achromatic/chromatic contrast sensitivities on image interpretation

In an image viewing situation, the perceived spatial resolution will depend on the data and its representation; if an image is produced with an actual spatial resolution twice that of an otherwise identical image, it does not mean that it will be viewed twice as closely. Rather viewing distance will depend upon the subject matter and size of the image. For example, a full Landsat scene may well be sized to subtend the same viewing angle as a small portion of such a scene; the full scene will then contain more high spatial frequency information, provided that the resolution is not limited by display processes. If spatial frequencies extend into the region where high frequency chromatic information is lost, the image will be perceived differently. Consequently if information is to be perceived as accurately as possible, factors which depend on viewing distance and hence perceived spatial frequency, such as contrast sensitivity, should be considered.

7.2.1 The achromatic appearance of high frequency chromatic gratings

The achromatic appearance of high spatial frequency chromatic gratings is a well-known effect (for example, see Evans, 1974; van der Horst and Bouman, 1969; De Valois and Switkes, 1983) which can be observed simply by generating gratings of varying spatial frequency and observing the point at which chromaticness disappears. It seems probable from Mullen's results that this effect is due to a lightness component in the chromatic grating which is masked at lower frequencies, but becomes the observed grating at higher frequencies with the loss of chromaticness. (De Valois and Switkes (1983) have reported on this masking effect; threshold level achromatic gratings are effectively masked by chromatic gratings, but in the reverse sense the masking is much less pronounced, if at all operative; hence up to the point of loss of chromaticness, the achromatic component would be masked by the chromatic component.) Because of the obvious difficulty in removing residual lightness components from a chromatic signal (in section 7.3 we investigate the feasibility of doing this in the developed framework), it is unlikely that a data variable displayed in a chromatic channel will ever be entirely free from such small lightness variations; the commonly noted achromatic appearance of high frequency chromatic gratings attests to this. It is thus worth looking at the effect of this phenomenon on the perception of detail in imagery with significant high spatial frequency content.

We can consider a simplified case where two data variables are displayed in a single image, one as a lightness variation and one as a constant-hue chromatic variation. If variations on the chromatic channel become achromatic at high spatial frequencies, the result of viewing this image will be that at high spatial frequencies, the detail in the two images will become indistinguishable. Detail on the lightness channel will be confounded by the achromatic-appearing data on the chromatic channel, resulting in a loss of information which would otherwise be appreciable on the achromatic channel.

In a more general case, such as that of Landsat data where three channels are usually assigned to the three colour guns of a display device, the same effect will take place. In this case, it is not the individual data channel comprehension which will be affected, but rather the combination of channels which is displayed on each of the perceptual channels normally used to interpret the images. Nevertheless, the effect will still be a confounding one on the achromatic signal. Of course, if the spatial frequencies of the data remain below the levels at which the visual system starts to lose chromatic resolution, the effect will not be misleading. However, in imagery such as that from satellites, images often contain detail above chromatic resolution cut-offs. For example, a full Landsat scene enlarged to 1m square, and viewed at a distance of 1m, results in a viewed spatial resolution of approximately 75 pixels/degree (assuming that resolution is not limited by display processes). At a viewing distance of 2m, the same image has a viewed spatial resolution of approximately 140 pixels/degree. According to Mullen, the effective cut-off in chromatic resolution occurs well below this level, at about 13 cycles/degree. Even if we consider the highest frequency components present in the scene to correspond to 2 pixels/cycle (the spatial-frequency content of images depends both on the information being displayed, and on the display process; it thus cannot necessarily be simply related to the spatial-frequency content of the information (see Pratt, 1978)), the viewed spatial resolution will be well above the limits for visual chromatic resolution.

7.2.2 Interference effects

In the course of experimenting with filtered high frequency data sets, an interference effect became apparent. It was discovered that low frequency chromatic fringes could be produced by the interference of high frequency achromatic and chromatic patterns. This low frequency fringing is entirely misleading about the actual spatial nature of the chromatic pattern. In real data displays this effect can arise when composite colour imagery is produced from several highly spatially correlated images (Landsat imagery again being a good example). In practice it was found to occur in areas where vegetation, and hence chromatic, striations are slightly out of frequency synchronisation with lightness striations due to slope or shading factors; the result is a small difference in the spatial fequency of the two patterns, causing interference.

These effects, the achromatic appearance of high spatial frequency chromatic gratings and the specious chromatic fringing, together suggest the value of providing within the framework access to some approximation to the achromatic and chromatic channels; this would allow spatial-frequency dependent processing such as filtering. The following section describes an informal investigation into these high spatial frequency effects, using test gratings and imagery, in an effort to discover how detrimental or misleading they might be in practice, and whether it is worth trying to avoid them. The final section discusses ways to reduce or avoid such effects.

7.3 Subjective experiments within the framework to investigate high frequency effects

7.3.1 Investigation of the achromatic appearance of high frequency chromatic gratings

As noted earlier, gratings which are supposedly isoluminant generally still appear visible as achromatic gratings after chromaticness is lost. The question is whether the perceived grating effect, above the point at which the chromatic content is lost, is due to a lightness component in the chromatic grating which is there due to lightness mismatching of the grating's constituent elements.

This could be investigated using the technique of flicker photometry, in which spatially adjacent samples are presented alternately at an appropriate rate; isoluminant samples appear identical under such conditions. This process makes use of the fact that the transient response of the visual chromatic mechanisms is slower than that of the achromatic mechanism (Kelly, 1974), and can hence be used to detect a lightness difference between compared samples. However, it is not clear that isoluminancy under temporal variation will necessarily correspond to isoluminancy under steady viewing conditions. Not enough is known of the temporal characteristics of possible interactions (or inter-dependencies) of achromatic and chromatic channels to assume this; slightly different mechanisms might well be the limiting factors in time-varying and time-invariant perception. In this work a time-invariant approach was taken, though careful comparison of the results of the two processes might well provide useful information on the spatio-temporal nature of possible inter-channel dependencies. (In fact, it is not possible to use a time-varying technique on all display devices.)

In the investigation performed, an attempt was made to null the perceived achromatic variation in the chromatic grating by adding a small lightness component. Initial efforts with sinusoid gratings achieved partial nulling; the location of the nulling in the grating cycle could be changed by phase shifting the lightness grating with respect to the chromatic grating. Presuming that this indicated a mismatch between the metrics of the chromatic and achromatic gratings (saturation and lightness respectively in UCS), the same experiments were performed with square wave gratings. This resulted in almost complete disappearance of the grating effect, suggesting that the gratings no longer contained a significant achromatic component. The amplitudes of the lightness component required to null the achromatic appearance of the chromatic gratings were small (less than 2 JND), though distinctly perceived as a grating when viewed alone. Presence of the nulling grating did not noticeably affect the point at which chromatic content was no longer visible in a square wave grating. These experiments were performed using a colour television monitor due to the need to interactively adjust the amplitude of the nulling grating by very small amounts (down to 0.2 JND, for a square wave grating with a saturation of 80 JND). Reproduction of the effect on a photographic print product is not feasible because process variations in a printed photographic product are large enough to introduce extraneous components.

Although substantiation of this result would require performing more carefully controlled experiments, it does suggest, in keeping with Mullen's results, that the perceived achromatic nature of high frequency chromatic gratings is due to a residual achromatic component, which is visible as a grating because of the extreme sensitivity of the visual system to small lightness variations. This means that it should be possible to trace a path in UCS corresponding to chromatic channels on which high frequency chromatic gratings do not appear achromatic. If lightness nulling data were obtained for a range of hues and saturations, and for a range of spatial frequencies, spatial-frequency-specific constant lightness surfaces in the UCS of the display-device model could be isolated.

In the practical display of data, however, the accuracy necessary to realise a purely chromatic representation, and the spatial-frequency dependence of such a realisation, suggest that it would be entirely unrealistic to expect to be able to place data in genuinely isoluminant chromatic ranges. Thus the high spatial frequency content of any data variable displayed on a practicably realisable chromatic channel is likely to appear achromatic.

7.3.2 The use of test gratings to investigate interference effects

High frequency gratings differing in frequency by a small amount, placed one each in achromatic and chromatic channels, produce clear interference. This is shown graphically in figure 7.2, using square wave gratings for ease of illustration; in the experiment performed, the achromatic grating had a lightness range of 40 JND, a mid-lightness level of half the maximum lightness (100 JND range), and a spatial period of 4 resolution elements. The blue-yellow chromatic grating had a saturation range of 80 JND and a spatial period of 5 resolution elements. The achromatic and chromatic gratings were mapped into UCS in the lightness and blue-yellow directions; in the resulting image, a low frequency chromatic variation with a spatial period of 20 resolution elements was clearly visible. Heuristically, the darker areas of the lightness variation mask the chromatic signal, while the lighter areas do not. Despite the chromatic grating being above the resolution limit for seeing colour, the interference beats are apparent as lower frequency chromatic (and possibly lightness) fringes. The low frequency lightness variations are less apparent, and proper experimental conditions would be required to establish their existence; it is possible that they are masked by the chromatic variations (De Valois and Switkes, 1983). Lowering both grating frequencies renders the chromatic fringing less obvious; possibly at low enough frequencies the visual system is used to interpreting such variation as due to the illumination and not due to the chromatic properties of the subject being viewed, and consequently less notice is taken of it. It is also possible that it is necessary to have a certain number of cycles of a pattern within a field of view if its presence is to be clearly detected.

It is very unlikely that in an image the visual system would reconstruct the higher frequency variations from the perceived low frequency fringing; rather it would be likely to interpret the data, incorrectly, as having a lower frequency chromatic component. This fringing then, could cause a distinct erroneous and detrimental effect in image interpretation.

7.3.3 Effects of differential contrast sensitivity on real imagery

Investigations into the effects of differential channel contrast sensitivity on real imagery were somewhat inconclusive, due to the subjective nature of judging such imagery, and the lack of any means of defining a "correct" interpretation. However, the effects seen in the test grating experiments; the achromatic appearance of a chromatic image of high spatial frequency content only, and the chromatic fringing caused by interference of spatially correlated high frequency achromatic and chromatic patterns; also appeared with real imagery. The addition of both correlated and uncorrelated high frequency chromatic image components to an achromatic image results in an obfuscation of detail which is evident when compared with the original achromatic image, but not necessarily evident otherwise. The effect of adding this high frequency chromatic component is almost indistinguishable from that of adding the same data as an achromatic component.

Removing the high frequency chromatic component from an image does not visibly detract from the composite image, but can have the effect of increasing apparent colour saturation. This is because removing the high frequency component modifies the data distribution shape (histogram); this generally results in better colour saturation if the data is stretched between fixed limits in colour space (depending on the data distribution and display device gamut shape). Consequently the effects of removing high frequency chromatic components are difficult to judge due to this rather more dominant histogram modification effect. A compensation for this effect can be made; standard histogram modification techniques can be used to equalise the histograms of the unfiltered and filtered images; but such processes affect the spatial frequency content of the data being presented and may hence introduce other artifacts.

7.4 Minimising the detrimental effects of differential contrast sensitivities - discussion on appropriate techniques

The experiments described in the previous section substantiate the suggestion that in artificially produced images, the effects of the differences in visual system contrast sensitivity to achromatic and chromatic high spatial frequency stimuli should be avoided.

Increasing saturation increases the apparent contrast sensitivity in the chromatic channels, and one way to compensate for the lower contrast sensitivities of the chromatic channels would be to enhance the signals by the inverse of their sensitivity functions. However, this form of compensation is unprofitable: the contrast drops so quickly from the peak that to make an appreciable effect, enhancement would have to be large, reducing the range of colour space which could be used for non-enhanced (low spatial frequency) variations.

A more practical approach is to ensure that no high frequency chromatic signals are present. This can be performed simply by filtering the UCS chromatic components to the level suggested from Mullen's results.

Choosing to filter the opponent channel representations in UCS does not preclude treatment of false-colour composites in which the data is placed directly on display device colour guns; provided that the relationship between UCS and colour gun-count co-ordinates is invertible (such as in the colour monitor model), the required filtering can be applied. If it is not invertible (such as in the photographic process model), either the data must first be placed in UCS and after filtering, transformed into gun-counts, or a numerical approximation to the filtering on gun-counts must be derived.

In the experimental work described in this chapter, filtering was performed in the frequency domain by taking the Fourier transform of the images and inversely transforming after filtering. This is computationally expensive; if a modulation transfer function of the required filter is estimated from Mullen's results, it can then be approximated by a simple box filter, which is substantially faster to compute.

Performing this filtering presupposes that images are to be examined at a fixed viewing distance, since the level of detail perceivable is dependent on viewing distance. In practice the closest expected viewing distance can be taken to determine filtering levels. A rationale for viewing distance might well be that at which the finest detail will just be resolved by the achromatic channel. This approach might be most suitable for Landsat data, which is often analysed at several resolutions.

7.5 Summary

The work described in this chapter was designed to investigate two factors relevant to the display process: first, whether there are good reasons for considering the requirement in a display system for processing at a visual level earlier in the pathway than the level at which we recognise perceptual signals; second, whether the chosen framework allows for a representation of such a level (the opponent level) at which spatial-frequency-dependent processing can profitably be performed.

The achromatic appearance of high spatial frequency chromatic variations, and the interference effect giving rise to specious chromatic fringing, suggest the value of providing access for spatial-frequency domain processing at a low visual level.

Imagery represented in UCS, and realised in the developed framework, can be appropriately filtered to remove the detrimental effects of the difference between achromatic and chromatic channel contrast sensitivities. Such filtering should remove confounding noise from the perceived achromatic information, while not detracting from the perceived chromatic information (in fact due to the distribution modification effect, apparent chromatic saturation can be improved with no penalty of additional image pixel over-saturation). Imagery not directly displayed in UCS can still be filtered to advantage using the developed display framework.

The experiments performed in the course of this investigation were designed to demonstrate the occurrence of various visual effects within the developed framework, rather than to contribute to current knowledge of the human visual system. However, more careful and rigorous controlling of the experiments performed might well provide useful information in this field, and consequently is an aspect of the work which appears in itself worth pursuing.

This work has extended the consideration of the spatial frequency characteristics of the human visual system, first treated by Faugeras and Hall and Andrews, in the field of image processing. We have concentrated on considering the effects of the differential contrast sensitivity of the visual system achromatic and chromatic mechanisms which might be detrimental to correct interpretation of colour image data, and on ways of avoiding these effects; this was not considered in the earlier works. These issues have also not previously been raised in the context of the interpretation of remotely sensed multi-spectral data; we believe that they could be significant to correct interpretation of such data.

It should be pointed out that the isolated opponent channel representations can be used in the manner suggested by Faugeras for subjective image quality improvement. In fact, Faugeras suggests that a comparison of such results with those obtained using his visual system model would be valuable; such a comparison would require experiments of a more objective and controlled nature than have been performed in this work. Similarly, the opponent channel representations could be used as a basis for image compression, such as performed by Hall and Andrews, who also suggest a comparison of such results with those obtained using their visual model.

Chapter 8
Conclusions

In Chapters 1 and 2 of this thesis we proposed a display approach which first, finds an appropriate structural representation for the data to allow the presented image to be recognisable as a realistic scene; second, represents the data variables by perceptual spectral attributes appropriate to their nature and structural representation; and third, performs appropriate spatial or spectral enhancements or compensations based on the characteristics of low-level visual mechanisms.

This approach is aimed not only at finding satisfactory representations for arbitrary data set types which under conventional display methods are difficult to appreciate, but also at using an understanding of the human visual system to rationalise the ad-hoc uses of colour in image data display and improve the effectiveness with which the intrinsic information in the data is communicated to an observer.

Realisation of the proposed approach required the development of a computational framework which we based on a perceptually uniform colour space. In this chapter we look at the overall effectiveness of the approach and its implementational framework, discussing its limitations and possible extensions.

8.1 The effectiveness of the developed approach

The first stage of the proposed approach involves choosing a realistic structural representation for a data display. In Chapter 6 we presented a method of spatially superimposing two data variables with arbitrary spatial variation. As summarised more fully in the concluding section of that chapter, the result is a display in which the coloured surfaces produced are clearly recognisable as such (figures 6.6 to 6.9), indicating that for the types of data considered, the first-stated requirement for data appreciation can be achieved by introducing an implied spatial three-dimensionality in this manner. The method involves extending scene synthesis techniques to incorporate a model of colour pigmentation, and cater for arbitrary variations in colour. The application of this method is not restricted to the particular display problem for which it was developed; it can also substantially enhance both local and global appreciation of the characteristics of a single data variable (such as in figures 6.8(e) and(f)).

We also showed, in Chapter 5, how appropriate colour assignment to a data set with embedded structural information can allow the scene structure to be intuitively appreciated.

Thus we can conclude that, apart from data-dependent limitations (discussed in the following section), the developed methods can achieve the required structural representation of data in the form of a realistic scene. The shaded-relief method of implying spatial three-dimensionality was used because of its relatively low computational cost, and because it maintains a rectangular grid representation allowing direct overlays of ancilliary data; other methods, such as stereoscopic or perspective viewing, could equally well be implemented in the developed framework.

The second stage of the display approach is that of spectral assignment to data variables. From the results of generating univariate and bivariate colour map sequences (figures 4.2 and 4.3), we can conclude that theoretical schemes specifying colour sequences in perceptual terms can be realised in practice. The use of such schemes allows the intuitive comprehension of coloured maps to an extent not achieved by ad-hoc colouring methods.

Similarly, in the display of multi-spectral data, spectral assignment in perceptual terms can be achieved, rendering the images intuitively interpretable. Products depicting informative data variables in chosen attributes, and uniform gradations, of colour (such as those in figures 5.2 and 5.3) have been used, and strongly preferred, by analysts carrying out ecological and geophysical studies; the approach is consequently well justified for such data sets.

It is also clear from the illustrations in Chapter 5 (figures 5.3 and 5.4) that the problem of achieving colour contrast in correlated multi-spectral data sets can be successfully solved by using a statistical decorrelating technique, and mapping the decorrelated components orthogonally in a perceptually uniform colour space, perceptual or data-dependent constraints determining spectral representation.

The use of perceptual spectral attributes to represent the physical properties of a scene; the use of lightness variations to depict surface topography, and the use of correlated saturation and lightness variations to depict pigment density; have also been successfully realised within the developed framework, in so far as this can be subjectively judged from the apparent realism of the scenes produced.

Thus we conclude that the framework gives sufficient control over specification of, and for realisation of, spectral perceptual attributes to allow their effective use for any required spectral assignment to displayed data variables. The use of perceptual attributes for representing data variables allows intuitive appreciation of the data characteristics to an extent not achieved by arbitrary or ad-hoc spectral assignment.

The third stage of the display approach involves the role of display enhancements, or compensations for low-level visual characteristics. In Chapter 7 we suggested that the perception of high-spatial-frequency information in detail-rich imagery (such as Landsat MSS data) can suffer detrimental effects due to the different spatial resolutions of the human visual system achromatic and chromatic mechanisms. We showed that such effects can be avoided by appropriately filtering the high spatial frequency chromatic information.

These results indicate that within the framework, it is feasible to perform spatial frequency processing corresponding to a low (opponent channel) visual level to enhance appreciation of image detail by minimising possible detrimental effects. These results also indicated that as suggested in Chapter 2, such operations make little difference to appreciation of the main characteristics of data variables in a display, but can affect interpretation at a detailed level.

We have shown that each stage of the proposed approach can be realised within the developed framework. Our results suggest that a framework based on a perceptually uniform colour space, allowing spatial and spectral access at various levels, is a suitable one in which to realise image processing techniques based on visual system considerations. Using an empirically derived uniform colour space appears to allow the representation of data variables by appropriately chosen perceptual attributes which can be intuitively interpreted in an image, while using a space with a metric representing noticeable colour differences appears to allow scalar interpretation of data values. These results suggest that the main attributes of colour space uniformity can be realised by the display device modelling techniques used. It is also clear that representation of display device gamuts in a perceptual space allows improved control over achieved colour contrast and saturation, leading to full use of the colour gamut of any particular modelled display device. Perceptual representation of display device gamut shapes also contributes to an understanding of the physical limitations of display devices, and the range of colours they can produce; appreciation of the limitations in reproducing colour representations on different types of display devices results.

There remains the question of whether the increased computational complexity involved in producing displays based on this approach is worth the gain in intuitive appreciation of the data. This might depend on the complexity of a display, and the spatial distribution and physical meaning of data variables being displayed. We have shown examples for which conventional display methods simply do not provide satisfactory data presentations, and also examples where the presentation is significantly improved by application of the proposed approach. The importance of being able to specify the required colour representation of informative data variables in perceptual terms cannot be over-emphasised; a better understanding of the representation of the data, and a more intuitive appreciation of its characteristics, invariably results. We believe that with the decreasing cost of computation, and the extent to which data analysts are now using remotely sensed and integrated data sets, the increased complexity of the proposed approach is justifiable.

Overall, these results lead to the general conclusions that use of the developed approach can result in a substantial improvement over conventional display methods, both in high level scene comprehension of complex composite images, and in lower level spectral interpretation of all types of imagery. The framework developed to realise the approach is sufficiently flexible and analytically tractable to allow the required spatial and spectral operations to be performed within it, and building it on three-dimensional spectral and spatial spaces is satisfactory for the types of data display considered.

8.2 Limitations

Structural representation in the form of a realistic scene cannot necessarily be achieved with all types of data. The depiction of a realistic surface enforces constraints on data smoothness and continuity, either limiting its application to data satisfying these constraints, or making some form of pre-processing necessary. In addition, the results obtained suggest that modelling an overlaid transparency may require some additional perceptual cues to assist in achieving realism.

Data variables used to judge the effectiveness of the developed approach were for the most part derived from real-world physical measurements; they consequently have an inherent physical structure. It is not clear that the techniques developed in this work would necessarily be effective for data without such structure, particularly since perceptual attributes used for variable representation have an intrinsic physical basis. Thus application to data which, for example, has arbitrary rather than geographic spatial attributes, might require some form of pre-processing or further consideration of the data nature.

This work has not made a detailed investigation into the extent to which observed limitations in uniformity achieved (for example, in the colour map sequences) are caused by variations in the actual uniformity of the CIE spaces, inaccurate display device modelling, or inconsistencies in the production of the final prints. Such a study would be valuable, though possibly extensive to perform. In addition, detailed or objective comparisons between the UCS representations on different types of display device were not made; this limits the extent to which the display-device-independence of UCS-specified colouring can be verified.

Achieving a display which accurately depicts any desired colours is limited not only by the accuracy and consistency of display device or process models, but also by the effects of the two-dimensional spatial distribution of the displayed data. In particular, this work has not considered the effects of induction on perceived colours, and the extent to which data values could be incorrectly judged under varying conditions of examination. While existing induction models could be realised within the developed framework, the difficulty of appropriately compensating for induction effects under varying display conditions suggests caution with incorporating such models.

Judgement of the results of this work in general was in part subjective; this limits the extent to which the relative advantages of the developed techniques can be gauged. It should also be recognised that the experiments on the effects of high-spatial-frequency signal content on image interpretation were investigatory only, and performed in an informal and non-rigorous manner; it is quite possible that visual factors other than those considered might have influenced the results, or that more rigorous or controlled experiments with imagery might suggest their re-appraisal. They should thus be considered in a role of confirmation that working within the developed framework leads to results consistent with those expected from consideration of low-level visual processes, and indicative that spatial access to the framework, and hence to a displayed product, at a low visual level is feasible. It should also be noted that the application of spatial frequency filtering introduces viewing distance restrictions.

8.3 Directions for future work

A number of areas for future investigation are evident from the results of this work. There is considerable potential for modelling and including additional effects which enhance the sense of realism in presented displays. This would be advantageous to reinforce both the structural representation of a scene, and the representations within a scene of individual data variables. For example, surface texture could be exploited as either a reinforcing, or a separating, variable; shadowing could also be included using existing models. As discussed in Chapter 6, specific drawbacks encountered in the presented scenes can also be overcome by utilising the full potential of the models used in this work.

As an alternative form of implying spatial three-dimensionality, displaying a synthesised colour scene in perspective, or as a stereoscopic pair, might further enhance intuitive appreciation of scene characteristics. A perspective view between a (possibly coloured) surface and an overlaid opaque layer, perhaps with non-obtrusive graphical spatial tie-lines, could also provide a method of introducing an additional variable into a display. Modelling haze, the effect of atmospheric scattering, in a spectral (distance-dependent modification of colour) or spatial (distance-dependent modification of spatial frequency) sense might also considerably enhance the sense of realism and depth achieved in a perspective view. This could be implemented directly within the developed framework. Although a perspective display can involve a partial loss of information, presenting either spatially or temporally distinct multiple views would overcome that problem.

In fact the exploitation of temporal variations, either to reinforce spatial representations, or to represent a separate data variable, could add significantly to the information conveying capabilities of a display process. This would require the development of computational techniques to increase the speed of scene synthesis, suggesting an evaluation of the degree of spectral smoothness required to maintain the realism of a synthesised scene, and the effective loss in realism and appreciation of data characteristics under more severe spectral quantisation.

At lower levels of representation, there is potential for improving the uniformity of the perceptual space used for any particular display device. This could incorporate not only developing more sophisticated device models, but also compensating for some known anomalies in existing models.

Induction affects the perceived colour in a display, and a study of the relative sizes of judgement errors and colour space non-uniformities would be valuable in considering the inclusion of an induction model in the framework. The choice of colour ranges which minimise induction effects would also be worth serious investigation.

In general, the development of objective testing techniques would be of value in judging the relative merits of display techniques. There is also considerable potential for performing the experiments described in Chapter 7 in a more rigorous manner; this would be advantageous in the context both of display effects, and of investigating visual system properties.

One final area which could merit investigation is in the design and use of specialised hardware for a colour image processing system which could perform some of the colour transformations in real time; for example, real time addressing of a colour display device such as a colour television monitor in UCS terms would be of considerable advantage. This would be particularly useful if temporal variations were incorporated.

This work has illustrated the value of a display approach which gives analysts the opportunity to display specific informative data variables in appropriate perceptually-defined colour representations. More generally, in the analysis of image data from remotely-sensed or otherwise derived data sets, it is crucial that the information directly of interest to the analyst can be intuitively extracted. This requires the specification of these informative aspects in the data domain, and their representation in perceptual terms at spatial and spectral resolutions appropriate to what is known about the data. Future work on display processes should hence be directed towards both the extraction from a data set of its informative aspects, and the use of display processes which can depict these aspects explicitly, thus allowing their intuitive appreciation.

Appendices

Appendix 1 The specification and measurement of colour

A1.1 The specification of colour

The response of the human visual system to a colour stimulus can be specified in terms of three co-ordinates. This stems from the experimentally established trichromatic generalisation, which states that any colour can be matched by a combination of three appropriately chosen primaries, subject to the condition that one or two of the primaries may have to be added to the colour being matched. The chosen primaries need only satisfy the condition that none can be matched by a mixture of the other two. In stronger form, the generalisation also states that linearity (proportionality and additivity) holds over a wide range of observing conditions. There is no physically realisable set of primaries which will match all colour stimuli in a strictly additive (positive amounts) manner.

The Commission Internationale d'Eclairage (CIE) recommends that colour stimuli be specified in terms of spectral tristimulus values $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$, which are also termed the colour matching functions of the CIE 1931 Standard Colorimetric Observer. These matching functions are linear transformations of matching functions experimentally established on the basis of colour matching with a 2° visual field. The functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ are positively valued at all wavelengths, and are not themselves physically realisable. The function $\bar{y}(\lambda)$ is also the standard luminous efficiency function of a normal observer, defining the CIE (1924) Standard Observer for Photometry.

An alternative specification is in terms of the CIE 1964 Standard Colorimetric Observer. In this case a 10° visual field was used for matching, and the colour matching functions generated are termed $\bar{x}_{10}(\lambda)$, $\bar{y}_{10}(\lambda)$, $\bar{z}_{10}(\lambda)$. The CIE recommends the use of the CIE 1931 matching functions for visual fields of angular subtense between 1° and 4° and the CIE 1964 matching functions for fields of angular subtense greater than 4°.

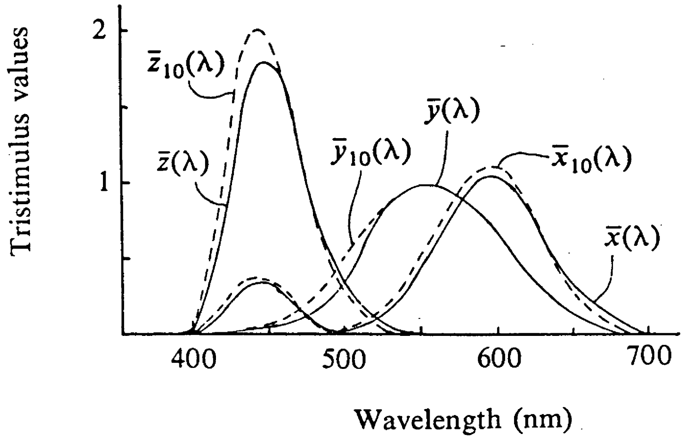


Figure A1.1 CIE colour matching functions. The matching functions for the CIE 1931 Standard Observer are shown in bold line, and the matching functions for the CIE 1964 Supplementary Observer are shown in broken line.

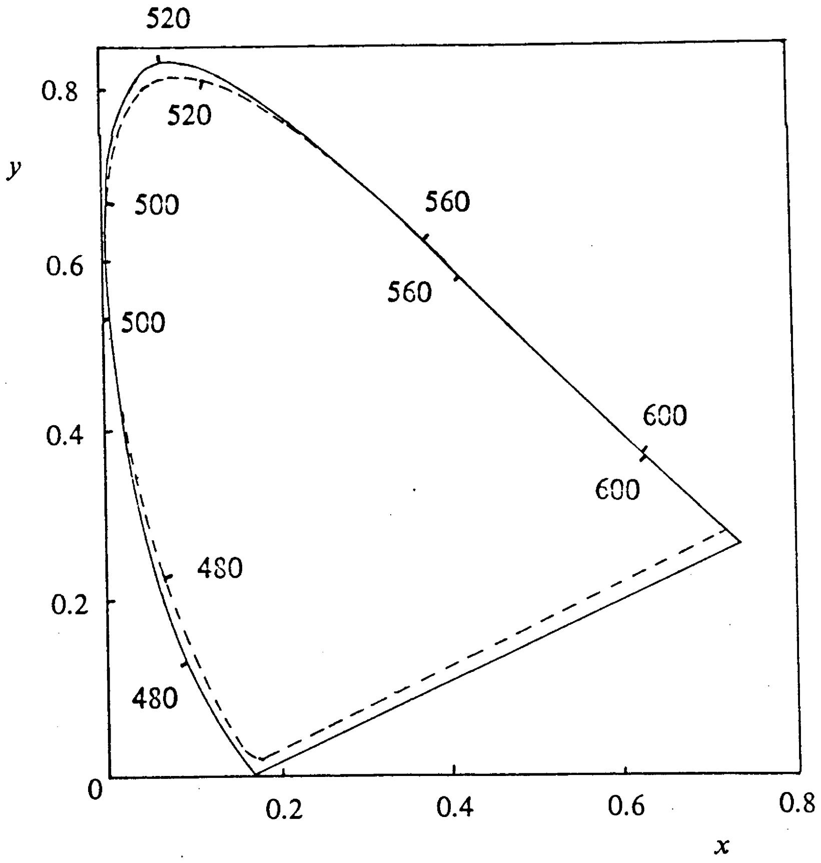


Figure A1.2 CIE Chromaticity diagram. The locus of spectrum colours for the CIE 1931 Standard Observer is shown in bold line, and the equivalent locus for the CIE 1964 Supplementary Observer is shown in broken line.

Figure A1.1 shows the CIE 1931 and 1964 colour matching functions in bold and broken lines respectively.

A colour stimulus is precisely defined in terms of its spectral energy distribution $\{E(\lambda)\}$. The human visual system response to this colour stimulus can then be defined in terms of tristimulus values, given by

$$X = \int_{\lambda} E(\lambda)\,\bar{x}(\lambda)\,\mathrm{d}\lambda,$$

and similarly for Y and Z. The computational evaluation of tristimulus values is usually performed by summation (weighted ordinate method), giving approximate values.

Colour stimuli with identical tristimulus values, but different spectral energy distributions, are termed metameric.

It is convenient to project the three-dimensional tristimulus colour space into a two-dimensional chromaticity space. This projection is performed by dividing each tristimulus value by the sum of the tristimulus values:

$$x = \frac{X}{X+Y+Z} \qquad y = \frac{Y}{X+Y+Z} \qquad z = \frac{Z}{X+Y+Z},$$

where x, y, z are the chromaticity co-ordinates of a stimulus with tristimulus values X, Y, Z. Since $x+y+z = 1$, chromaticity co-ordinates are generally specified by x and y only; the commonly used xy chromaticity diagram provides a convenient graphical representation of the chromaticity of a stimulus. Figure A1.2 shows the locus of spectrum colours in the chromaticity diagram for the CIE 1931 Standard Observer (in bold line), and for the CIE 1964 Supplementary Observer (in broken line). It should be noted that the chromaticity co-ordinates do not uniquely specify the colour of a stimulus, but rather its chromatic attributes. Unique specification requires also specification of its achromatic attributes, and consequently three-dimensional specification in terms of tristimulus values or other derived three-dimensional co-ordinates (see Appendix 2).

When viewing a stimulus by reflected illumination, the spectral energy distribution is given by the product of the spectral energy distribution of the incident radiation and the surface albedo, or spectral reflectance function.

A1.2 The measurement of colour

The human response to a colour stimulus can be predicted by measuring the tristimulus values of the colour under appropriate illumination. Two methods of performing this measurement are commonly used.

Most accurate (potentially) is to use a spectrophotometer to measure the emitted, reflected, or transmitted spectral energy distribution, and then to evaluate tristimulus values after weighting with the chosen colour matching functions. Accuracy depends on the bandwidth and reproducibility of the spectrophotometer, and on the method of calculating the tristimulus values (see Judd and Wyszecki, 1975; MacAdam, 1981; Stearns, 1981a,b).

Less accurate, but more convenient, is to use a tricolorimeter. A tricolorimeter has filters whose spectral characteristics approximate those of the colour matching functions; a direct measurement of tristimulus values results. For non-emitting colours, these values are valid when samples are viewed under the illumination used by the instrument (some tricolorimeters allow a choice of illumination).

It should also be mentioned that colour measurement can be made by visual comparison of a colour with a set of samples. The difficulties of controlling the illumination and surround effects, and the inconsistencies between individual observers, make this an unsatisfactory method for most applications.

A standard illuminant (of known and reproducible spectral energy characteristics) is generally used in the measurement of colour co-ordinates; tables of colour matching functions weighted with standard illuminants are widely available (see, for example, Wyszecki and Stiles, 1967). Detailed treatment of colour measurement under a variety of conditions can be found in Judd and Wyszecki (1975) or MacAdam (1981).

Appendix 2 Analytical formulations of CIELAB and CIELUV uniform colour spaces

The CIELAB space is produced by plotting in rectangular co-ordinates the quantities L^*, a^*, b^* defined by

$$L^* = 116(Y/Y_n)^{\frac{1}{3}} - 16, \qquad a^* = 500\{(X/X_n)^{\frac{1}{3}} - (Y/Y_n)^{\frac{1}{3}}\},$$

$$b^* = 200\{(Y/Y_n)^{\frac{1}{3}} - (Z/Z_n)^{\frac{1}{3}}\},$$

for $X/X_n,\ Y/Y_n,\ Z/Z_n\ > 0.01$.

The CIELUV space is produced by plotting in rectangular co-ordinates the quantities L^*, u^*, v^* defined by

$$L^* = 116(Y/Y_n)^{\frac{1}{3}} - 16, \qquad u^* = 13L^*(u' - u_n'), \qquad v^* = 13L^*(v' - v_n'),$$

for $Y/Y_n > 0.01$; and with

$$u' = \frac{4X}{X + 15Y + 3Z}, \qquad v' = \frac{9Y}{X + 15Y + 3Z},$$

$$u_n' = \frac{4X_n}{X_n + 15Y_n + 3Z_n}, \qquad v_n' = \frac{9Y_n}{X_n + 15Y_n + 3Z_n}.$$

In each case, the tristimulus values X, Y, Z can represent either the CIE 1931 Standard Colorimetric Observer and Co-ordinate System, or the CIE 1964 Supplementary Standard Observer and Co-ordinate System. (For further details of these systems, see Appendix 1, or Wyszecki and Stiles (1967).) The tristimulus values X_n, Y_n, Z_n define the colour of the nominally white object-colour stimulus.

For $X/X_n,\ Y/Y_n,\ Z/Z_n\ < 0.01$, slightly modified formulations apply, and are given in the Appendix of CIE (1978).

Appendix 3 Summary of developed software

The approach to colour image display, and the framework for its realisation, have been implemented in a comprehensive software package within CSIRONET (formerly the CSIRO Division of Computing Research). The software is written in Fortran 77, and is built on a general purpose image handling package (DISIMP - see references). Standard image processing utilities (SLIP - see references) were used to perform processing of a general nature on images.

The principal modules of the developed software are summarised below; the overall structure is outlined in figure A3.1.

Colour transformation

Subroutines to allow transformation between sets of spectral co-ordinates as follows:

HSL perceptual attributes (cylindrical polar co-ordinates);

UCS (CIELAB and CIELUV) co-ordinates (rectangular);

Tristimulus values;

Chromaticity co-ordinates (to chromaticity only; non-reversible);

RGB display device gun-counts (using display device model parameters).

Utility to interactively access these transformations.

Geometrical transformation

Subroutines to allow the generation of linear transformations between two finite-dimensional (up to 10-dimensional) spaces.

Utility, also available in subroutine form for operation under program control, to generate the required transformation from given input alignment specification.

Processing modules specific to image data types

(1) Scene synthesis

Utilities to perform:

colour surface representation;

overlaid colour transparency (on coloured surface) representation.

Subroutines to realise:

reflectance modelling;

pigmentation modelling;

spectral reflectance function creation, modification and manipulation.

(2) Treatment of multi-spectral data

Utilities for the placement of data variables in HSL/UCS:

informative data variables in specified alignments;

statistically decorrelated data variables in calculated alignments.

Subroutines to generate alignments if not otherwise specified.

(3) Map sequence generation

Utilities for sequence generation, and sequence and map display:

univariate sequence generation (interactively adaptive);

bivariate sequence generation (interactively adaptive);

sequence and map display;

graphical representation of sequence path saturation limits.

(4) Spatial frequency processing

Utilities to perform:

frequency domain analysis and filtering;

generation of images of one or two orthogonal or superimposed spatial frequency gratings, each of any specified spatial frequency and spectral characteristics.

Ancilliary operations

Utilities to generate display device model parameters for use by colour transformation routines:

generalised display device numerical model;

colour monitor physical model.

Utilities to display any defined colour gamut cross-sections:

graphical display;

image display;

graphical or image display with superimposed data scatter-plots.

Subroutines to monitor and record image saturation statistics.

Subroutines to deal with saturated pixels.

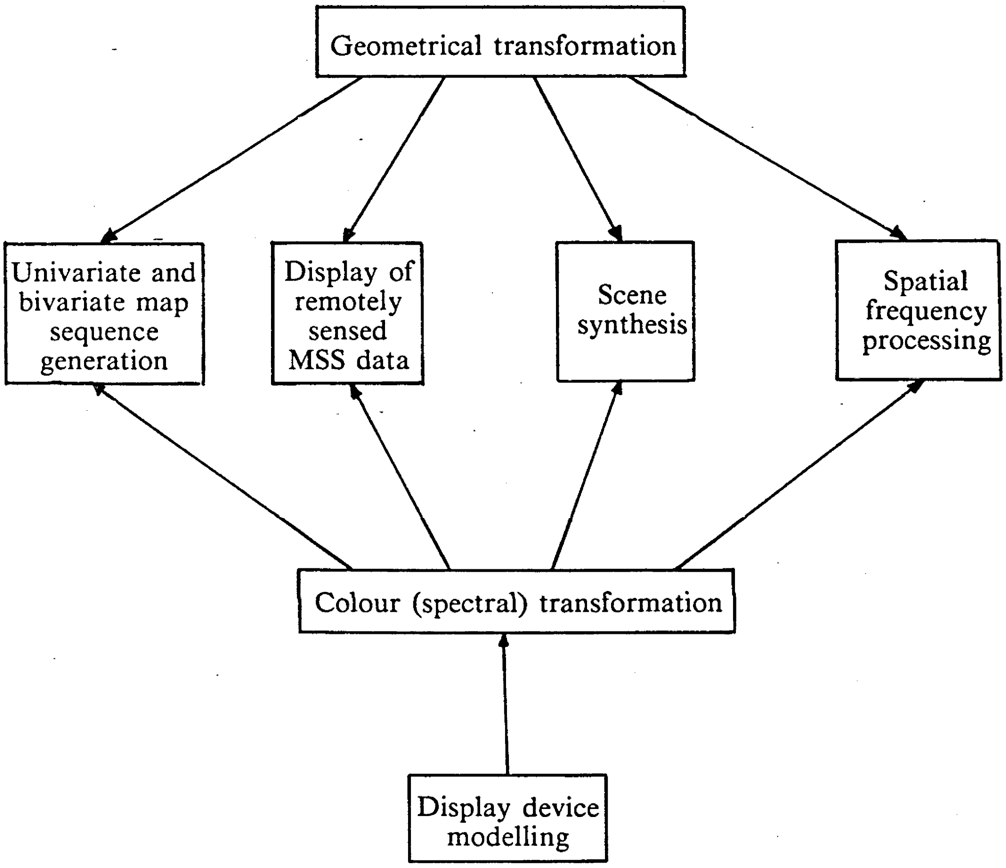

Geometrical transformation
Univariate and bivariate map sequence generation
Display of remotely sensed MSS data
Scene synthesis
Spatial frequency processing
Colour (spectral) transformation
Display device modelling

Figure A3.1 Overall structure of developed software.

References

Andrews, H.C. and Hall, C.F. (1978)
Digital Color Image Compression in a Perceptual Space
Ph.D. thesis (Hall); Univ. of Southern California IPI, USCIPI Rep. No.790.

Arvidson, R.E., Guinness, E.A., Strebeck, J.W., Davies, G.F., Schulz, K.J. (1982)
Image processing applied to gravity and topography data covering the continental U.S.
EOS : Trans. American Geophysical Union, Vol.63, No.18, May, pp.261-265.

Balon, R.J. and Cicone, R.C. (1979)
Uniform Color Space Analysis of LACIE Image Products
NASA Tech. Mem. 79-10221, (ERIM 132400-10-R), May.

Barlow, H.B. (1981)
Critical limiting factors in the design of the eye and visual cortex
Proc. R. Soc. Lond. B 212, pp.1-34. (The Ferrier Lecture, May, 1980.)

Barnard, S.T. (1985)
Choosing a Basis for Perceptual Space
Computer Vision, Graphics and Image Processing, Vol.29, pp.87-99.

Bartleson, C.J. (1978)
Comparison of Chromatic-Adaptation Transforms
Color Research and Application, Vol.3, No.3, pp.129-136.

Bartleson, C.J. (1979a)
Changes in Color Appearance with Variations in Chromatic Adaptation
Color Research and Application, Vol.4, No.3, pp.119-138.

Bartleson, C.J. (1979b)
Predicting Corresponding Colors with Changes in Adaptation
Color Research and Application, Vol.4, No.3, pp.143-155.

Beck, J. and Prazdny, K. (1983)
The perception of transparency with achromatic colors
Univ. of Maryland Computer Science Tech. Rep. TR-1240, Jan.

Bertin, J. (1981)
Graphics and graphic information processing
Translation of *La graphique et le traitement graphique de l'information* (1977)
Translated by W.J. Berg and P. Scott,
Walter de Gruyter, Berlin, New York.

Blinn, J.F. and Newell, M.E. (1976)
Texture and Reflection in Computer Generated Images
Comm. ACM, Vol.19, No.10, pp.542-547.

Blinn, J.F. (1977)
Models of Light Reflection for Computer Synthesized Pictures
ACM Computer Graphics (Siggraph) Vol.11, No.2, pp.192-198.

Blinn, J.F. (1978)
Simulation of Wrinkled Surfaces
ACM Computer Graphics (Siggraph), Vol.12, No.3, pp.286-292.

Blinn. J.F. (1982)
Light Reflection Functions for Simulation of Clouds and Dusty Surfaces
ACM Computer Graphics (Siggraph), Vol.16, No.3, pp.21-29.

Brady, M. (1982)
Computational Approaches to Image Understanding
ACM Computing Surveys, Vol.14, No.1, pp.3-71.

Buchanan, M.D. (1979)
Effective utilization of color in multidimensional data presentation
SPIE, Vol.199 Advances in Display Technology, pp.9-18.

Buchanan, M.D. and Pendergass, R. (1980)
Digital Image Processing:
Can Intensity, Hue, and Saturation Replace Red, Green, and Blue?
Electro-Optical Systems Design, March, pp.29-36.

Buchsbaum, G. (1981)
The Retina as a Two-Dimensional Detector Array in the Context of Color Vision Theories and Signal Detection Theory
Proc. IEEE, Vol.69, No.7, pp.772-786.

Campbell, F.W. and Robson, J.G. (1968)
Application of Fourier analysis to the visibility of gratings
J. Physiology, Vol.197, pp.551-556.

CIE (1978)
Recommendations on uniform color spaces - color difference equations, psychometric color terms
CIE Publication No.15(E-1.3.1)/(TC-1.3), Supplement No.2, pp.9-12.

Colwell, R.N. (1983)
(ed.) *Manual of Remote Sensing*, 2nd Edition.
American Society of Photogrammetry, Sheridan Press.

Cook, R.L. and Torrance, K.E. (1982)
A Reflectance Model for Computer Graphics
ACM Trans. on Graphics, Vol.1, No.1, pp.7-24.

Cornsweet, T.N. (1971)
Visual Perception, 2nd Edition
Academic Press, New York, London.

Cowan, W.B. (1983)
An inexpensive scheme for calibration of a color monitor in terms of CIE standard co-ordinates
ACM Computer Graphics (Siggraph), Vol.17, No.3, pp.315-321.

Crow, F.C. (1977)
Shadow Algorithms for Computer Graphics
ACM Computer Graphics (Siggraph), Vol.11, No.2, pp.242-248.

Draper, N.R. and Smith, H. (1964)
Applied Regression Analysis
John Wiley and Sons, Inc., New York, London, Sydney.

Daily, M. (1983)
Hue-Saturation-Intensity Split-Spectrum Processing of Seasat Radar Imagery
Photogrammetric Engineering and Remote Sensing, Vol.49, No.3, pp.349-355.

De Valois, K.K. and Switkes, E. (1983)
Simultaneous masking interactions between chromatic and luminance gratings
Journal of the Optical Society of America, Vol.73, No.1, pp.11-18.

DISIMP - Device Independent Software for Image Processing (1985)
CSIRONET Reference Manual No.33, Edition 2.1, March 1985
CSIRONET Image Systems Section, GPO Box 1800, Canberra, ACT 2601, Australia.

Egusa, H. (1983)
Effects of brightness, hue, and saturation on perceived depth
between adjacent regions in the visual field
Perception, Vol.12, pp.167-175.

Elachi, C. (1982)
Radar Images of the Earth from Space
Scientific American, Dec., pp.54-61.

Evans, R.M. (1974)
The Perception of Color
John Wiley & Sons, Inc., New York, London, Sydney, Toronto.

Faugeras, O.D. (1976)
Digital Color Image Processing and Psychophysics Within the Framework
of a Human Visual Model
Ph.D. thesis; Univ. of Utah Computer Science Dept. Tech. Rep. UTEC-CSc-77-029, June.

Faugeras, O.D. (1979)
Digital Color Image Processing Within the Framework of a Human Visual Model
IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.27, No.4, pp.380-393.

Fienberg, S.E. (1979)
Graphical Methods in Statistics
The American Statistician, Vol.33, No.4, pp.165-178.

Friele, L.F.C. (1979)
Color Metrics: Facts and Formulae
Color Research and Application, Vol.4, No.4, pp.194-199.

Frome, F.S, Buck, S.L. and Boynton, R.M. (1981)
Visibility of borders: separate and combined effects of color differences, luminance contrast, and luminance level
Journal of the Optical Society of America, Vol.71, No.2, pp.145-150.

Gagalowicz, A. (1982)
Masking Effects in the Discrimination of Color Texture Fields
Proc. IEEE Conf. Pattern Recognition and Image Processing (PRIP'82), pp.237-243.

Georgeson, M.A. (1980)
Spatial frequency analysis in early visual processing
Phil. Trans. R. Soc. Lond., B 290, pp.11-22.

Gerbrands, J.J. (1981)
On the relationships between SVD, KLT and PCA
Pattern Recognition, Vol.14, Nos.1-6, pp.375-381.

Gillespie, A.R. (1980)
Digital techniques of image enhancement
in *Remote Sensing in Geology*, Siegal, B.S. and Gillespie, A.R. (eds), pp.139-226.
John Wiley and Sons, Inc., New York.

Ginsberg, A.P. (1980)
Specifying relevant spatial information for image evaluation and display design: an explanation of how we see certain objects
Proc. SID, Vol.21, No.3, pp.219-227.

Gouras, P. and Zrenner, E. (1981)
Color Vision: A Review from a Neurophysiological Perspective
in *Progress in Sensory Physiology*, Volume I, Ottoson, D. (ed.-in-chief)
Springer-Verlag, Berlin.

Graetz, R.D., Gentle, M.R., Pech, R.P. and O'Callaghan, J.F. (1982)
The development of a land-based resource information system (LIBRIS) and its application to the assessment and moitoring of Australian arid rangelands
Int. Symp. on Remote Sensing of Environment: Remote Sensing of Arid and Semi-Arid Lands, Cairo, Egypt, Jan., pp.257-275.

Granger, E.M. and Heurtley, J.C. (1973)
Visual chromaticity-modulation transfer function
Journal of the Optical Society of America, Vol.63, No.9, pp.1173-1174.

Green, D.G. (1968)
The Contrast Sensitivity of the Colour Mechanisms of the Human Eye
J. Physiology, Vol.196, pp.415-429.

Grotch, S.L. (1983)
Three-Dimensional and Stereoscopic Graphics for Scientific Data Display and Analysis
IEEE Computer Graphics and Applications, pp.31-42, Nov.

Guth, S.L. (1972)
A New Color Model
in *Color Metrics*, Vos, J.J., Friele, L.F.C., Walraven, P.L. (eds)
Soesterberg, Holland, AIC.

Guth, S.L., Massof, R.W. and Benzschawel, T. (1980)
Vector model for normal and dichromatic color vision
Journal of the Optical Society of America, Vol.70, No.2, pp.197-212.

Haber, R.N. and Hershenson, M. (1973)
The Psychology of Visual Perception
Holt, Rinehart and Winston Inc., New York.

Haber, R.N. and Wilkinson, L. (1982)
Perceptual Components of Computer Displays
IEEE Computer Graphics and Applications pp.23-35, May.

Hall, C.F. and Hall, E.L. (1977)
A Nonlinear Model for the Spatial Characteristics of the Human Visual System
IEEE Trans. on Systems, Man, and Cybernetics, Vol.7, No.3, pp.161-170.

Hall, C.F. and Andrews, H.C. (1978)
Digital color image compression in a perceptual space
Proc. SPIE, Vol.149 Applications of Digital Image Processing, pp.182-188.

Haralick, R.M., Watson, L.T. and Laffey, T.J. (1983)
The Topographic Primal Sketch
Int. J. Robotics Research, Vol.2, No.1, pp.50-72.

Haralick, R.M. and Shapiro, L.G. (1985)
Image Segmentation Techniques
Computer Vision, Graphics and Image Processing, Vol.29, pp.100-132.

Haruyama, S. and Barsky, B.A. (1984)
Using Stochastic Modeling for Texture Generation
IEEE Computer Graphics and Applications pp.7-19, March.

Horn, B.K.P. (1974)
Determining Lightness from an Image
Computer Graphics and Image Processing, Vol.3, pp.277-299.

Horn, B.K.P. (1977)
Understanding Image Intensities
Artificial Intelligence, Vol.8, pp.201-231.

Horn, B.K.P. (1981)
Hill Shading and the Reflectance Map
Proc. IEEE, Vol.69, No.1, pp.14-47.

Horn, B.K.P. (1984)
Exact Reproduction of Colored Images
Computer Vision, Graphics and Image Processing, Vol.26, pp.135-167.

Huang (1975)
(ed.) *Picture Processing and Digital Filtering*
Springer-Verlag, Berlin.

Hunt, R.W.G. (1975)
The Reproduction of Colour in Photography, Printing and Television, 3rd Edition
Fountain Press, England.

Hunt, R.W.G. (1977)
The Specification of Colour Appearance. I. Concepts and Terms
Colour Research and Applications, Vol.2, No.2, pp.55-68.
The Specification of Colour Appearance. II. Effects of Changes in Viewing Conditions
Color Research and Application, Vol.2, No.3, pp.109-120.

Hunt, R.W.G. (1978)
Colour Terminology
Color Research and Application, Vol.3, No.2, pp.79-88.

Hutchinson, M.F. (1984)
A summary of some surface fitting and contouring programs for noisy data
CSIRO Div. of Mathematics and Statistics Cons. Rep. ACT 84/6.
G.P.O. Box 1965, Canberra, ACT 2601, Australia.

Ingling, C.R. and Tsou, B. (1977)
Orthogonal combinations of the three visual channels
Vision Research, Vol.17, pp.1075-1082.

Juday, R.D., Johnson, F., Abotteen, R.A. and Pore, M.D. (1978)
Generation of Uniform Chromaticity Scale Imagery From Landsat Data
Proc. LACIE Symp, NASA Johnson Space Flight Center Doc. JSC-16015, pp.899-910.

Juday, R.D. (1978)
Colorimetric Consideration of Transparencies for a Typical LACIE Scene
Proc. LACIE Symp., NASA Johnson Space Flight Center Doc. JSC-16015, pp.887-897.

Juday, R.D. (1979)
Colorimetric Principles as Applied to Multichannel Imagery
NASA Tech. Mem. 58215, July.

Judd, D.B. and Wyszecki, G. (1975)
Color in Business, Science, and Industry, 3rd Edition
John Wiley and Sons, Inc., New York, London, Sydney, Toronto.

Kanatani, K. (1984)
Detection of Surface Orientation and Motion from Texture by a Stereological Technique
Artificial Intelligence, Vol.23, pp.213-237.

Kaneko, T. (1978)
Color Composite Pictures from Principal Axis Components of Multi-spectral Scanner Data
IBM J. of Res. and Dev., Vol. 22, pp. 386-392.

Kelly, D.H. (1974)
Spatio-temporal frequency characteristics of color-vision mechanisms
Journal of the Optical Society of America, Vol.64, No.7, pp.983-990.

Kelly, D.H. (1983)
Spatiotemporal variation of chromatic and achromatic contrast thresholds
Journal of the Optical Society of America, Vol.73, pp.742-750.

Kender, J.R. (1976)
Saturation, Hue, and Normalized Color: Calculation, Digitization Effects, and Use
Carnegie-Mellon Univ. Comp. Science Dept. Report, Nov.

Land, E.H. (1977)
The Retinex Theory of Color Vision
Scientific American 237, No.6, pp.108-128.

MacAdam, D.L. (1981)
Color Measurement: Theme and Variations
Springer-Verlag, Berlin.

Marr, D. (1982)
Vision: A Computational Investigation into the Human Representation and Processing of Visual Information
W.H. Freeman and Co., San Fransisco.

Massof, R.W. and Bird, J.F. (1978)
A general zone theory of color and brightness vision. I. Basic formulation
Journal of the Optical Society of America, Vol.68, No.11, pp.1465-1470.
A general zone theory of color and brightness vision. II. The space-time field
Journal of the Optical Society of America, Vol.68, No.11, pp.1471-1481.

McDonnell, M.J. (1980)
Color Image Calibration
Univ. of Maryland Computer Science Center Tech. Rep. 964, Oct.

McDonnell, M.J. and Fowler, A.D.W. (1981)
Colour Image Calibration
Proc. Landsat-81, 2nd Australian Remote Sensing Conf., Laut, P. (ed), pp.6.7.1-6.7.5.
Canberra, Australia.

McLaren, K. (1980)
CIELAB Hue-Angle Anomalies at Low Tristimulus Ratios
Color Research and Application, Vol.5, No.3, pp.139-144.

Metelli, F. (1974)
The Perception of Transparency
Scientific American, 230 No.4 pp.90-98.

Meyer, G.W. and Greenberg, D.P. (1980)
Perceptual Color Spaces for Computer Graphics
ACM Computer Graphics (Siggraph), Vol.14, pp.254-261.

Mullen, K.T. (1985)
The Contrast Sensitivity of Human Colour Vision to Red-green and Blue-Yellow Chromatic Gratings
J. Physiology, Vol.359, pp.381-400.

Murch, G.M. (1984)
Physiological Principles for the Effective Use of Colour
IEEE Computer Graphics and Applications, Nov., pp.49-54.

Nayatani, Y., Takahama, K. and Sobagaki, H. (1981)
Formulation of a Nonlinear Model of Chromatic Adaptation
Color Research and Application, Vol.6, No.3, pp.161-171.

Neal, C.B. (1973)
Television Colorimetry for Receiver Engineers
IEEE Trans. on Broadcast and Television Receivers, pp.149-162, Aug.

Nemcsics, A. (1980)
The Coloroid Color System
Color Research and Application, Vol.5, No.2, pp.113-120.

Nickerson, D. (1981)
OSA Uniform Color Scale Samples: A Unique Set
Color Research and Application, Vol.6, No.1, pp.7-52.

O'Callaghan, J.F. (1979)
Colour Image Processing of Landsat Imagery
Proc. Landsat-79, 1st Australian Remote Sensing Conf.,
Green, A.A., Huntington, J.F. and Cook, R.W. (eds), May, pp.498-504.
Sydney, Australia.

O'Callaghan, J.F. and Simons, L.W. (1983)
Colourmap: An Interactive Colour Mapping System
Proc. 1st Australasian Conf. on Computer Graphics
Inst. of Engineers, Sydney, Australia.

O'Callaghan, J.F., Robertson, P.K. and Fraser, D. (1981)
Colour image display - it's not that simple
Proc. Landsat-81: 2nd Australian Remote Sensing Conf., Laut, P. (ed), pp.6.8.1-6.8.5.
Canberra, Australia.

Ohta, N. (1977)
Correspondance Between CIELAB and CIELUV Color Differences
Color Research and Application, Vol.2, No.4, pp.178-182.

Olsen, J.M. (1981)
Spectrally encoded two-variable maps
Annals. Assoc. American Geographers, Vol.17, No.2, June, pp.259-276.

Oyama, T., Mitsuboshi, M. and Kamoshita, T. (1980)
Wavelength-specific brightness contrast as a function of surround luminance
Vision Research, Vol.20, pp.127-136.

Paulus, W. and Kröger-Paulus, A. (1983)
A new concept of retinal colour coding
Vision Research, Vol.23, No.5, pp.529-540.

Pearlman, W.A. (1978)
A visual model and a new distortion measure in the context of image processing
Journal of the Optical Society of America, Vol.68, No.3, pp.374-386.

Pentland, A.P. (1984)
Local Shading Analysis
IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.6, No.2, pp.170-187.

Phong, B.T. (1975)
Illumination for computer generated pictures
Comm. ACM, Vol.18, No.6, pp.311-317.

Pointer, M.R. (1980)
The Gamut of Real Surface Colours
Color Research and Application, Vol.5, No.3, pp.145-155.

Pointer, M.R. (1981)
A Comparison of the CIE 1976 Colour Spaces
Color Research and Application, Vol.6, No.2, pp.108-118.

Pratt, W.K. (1978)
Digital Image Processing
John Wiley and Sons, Inc., New York, Chichester, Brisbane, Toronto.

Raines, G.L. (1977)
Digital color analysis of color-ratio composite Landsat scenes
Proc. Int. Symp. on Remote Sensing of Environment, Michigan
ERIM, Vol.11, pp.1463-1472.

Richter, K. (1980)
Cube-Root Color Spaces and Chromatic Adaptation
Color Research and Application, Vol.5, No.1, pp.25-43.

Rich, D.C. (1980)
A Color Metric from Opponent-Color Visual Channels
Color Research and Application, Vol.5, No.2, pp.76-80.

Robertson, P.K. and O'Callaghan, J.F. (1982)
Colour Image Display of Multi-dimensional Data in Uniform Colour Space
Proc. Radio Research Board Seminar on Image Processing
Univ. of NSW, Australia, Jun., pp.8.1-8.6.

Robertson, P.K. (1984)
The use of coloured surfaces in multi-channel data presentation
Proc. Landsat-84, 3rd Australasian Remote Sensing Conf., Walker, E. (ed), May, pp.509-511.
Queensland, Australia.

Robertson, P.K. and O'Callaghan, J.F. (1984)
The Generation of Color Sequences for Univariate and Bivariate Mapping,
Accepted for publication, IEEE Computer Graphics and Applications

Robertson, P.K. and O'Callaghan, J.F. (1985)
The application of scene synthesis techniques to the display of
multi-dimensional image data
Accepted for publication, ACM Transactions on Graphics.

Robson, J.G. (1980)
Neural Images: The Physiological Basis of Spatial Vision
in *Visual Coding and Adaptability*, Harris, C.S. (ed.)
Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Rosenfeld, A. and Kak, A.C. (1982)
Digital Image Processing (2nd Edition)
Academic Press, Inc., New York, London.

Rosenfeld, A. (1984)
Image analysis: problems, progress and prospects
Pattern Recognition Vol.17, No.1, pp.3-12.

Rubin, J.M. and Richards, W.A. (1982)
Color Vision and Image Intensities: When are Changes Material?
Biological Cybernetics, Vol.45, pp.215-226. (MIT AI Memo No.631, May, 1981.)

Santisteban, A. (1983)
The perceptual color space of digital image display terminals
IBM J. Res. & Dev., Vol.27, No.2, pp.127-132.

Schweitzer, D. (1983)
Artificial Texturing: An aid to Surface Visualisation
ACM Computer Graphics (Siggraph), Vol.17, No.3, pp23-29, July.

SLIP - Software for Satellite Image Processing (1985)
CSIRONET Reference Manual No.32, Edition 3.1, March 1985
CSIRONET Image Systems Section, GPO Box 1800, Canberra, ACT 2601, Australia.

Stearns, E.I. (1981a)
Influence of Spectrophotometer Slits on Tristimulus Calculations
Color Research and Application, Vol.6, No.2, pp.78-84.

Stearns, E.I. (1981b)
A New Look at the Calculation of Tristimulus Values
Color Research and Application, Vol.6, No.4, pp.203-205.

Stenius, A.S. (1978)
A Study in Black: The CIELAB and CIELUV L* Function for Very Low Values of Y
Color Research and Application, Vol.3, No.3, pp.109-113.

Taenzer, D. (1976)
Physiology and Psychology of Color Vision - a review
MIT AI Memo 369.

Tajima, J. (1983)
Uniform Color Scale Applications to Computer Graphics
Computer Vision, Graphics, and Image Processing, Vol.22, No.1, pp.305-325.

Takahashi, S. and Ejima, Y. (1983)
Chromatic induction as a function of wavelength of inducing stimulus
Journal of the Optical Society of America, Vol.73, No.2, pp.190-207.

TASC IPL Software (1981)
Advanced Digital Image Analysis for Geophysical Exploration
Optronics Journal, Nov.

Taylor, M.M., (1974)
Principal Components Color Display of ERTS Imagery
Proc. Canadian Symp. on Remote Sensing, Guelph, pp.1877-1887.

Tenenbaum, J.M., Garvey, T.D., Weyl, S. and Wolf, H.C. (1974)
An Interactive Facility for Scene Analysis Research
SRI AI Center Tech. Note 87, Jan.

Torrance, K.E. and Sparrow, E.M. (1967)
Theory for Off-Specular Reflection From Roughened Surfaces
Journal of the Optical Society of America, Vol.57, No.9, pp.1105-1114.

Troscianko, T.S. (1977)
Effect of Subtense and Surround Luminance on the Perception of a Colored Field
Color Research and Application, Vol.2, No.4, pp.153-159.

Trumbo, B.E.A. (1981)
Theory for coloring bivariate statistical maps
The American Statistician, Vol.35, No.4, pp.220-226, Nov.

Tsotos, J.K. (1984)
Knowledge and the Visual Process: Content, Form and Use
Pattern Recognition, Vol.17, No.1, pp.13-27.

UNIRAS software
European Software Contractors A/S, Denmark.

van der Horst, G.J.C. and Bouman, M.A. (1969)
Spatiotemporal Chromaticity Discrimination
Journal of the Optical Society of America, Vol.59, No.11, pp.1482-1488.

Vos, J.J. and Walraven, P.L. (1971)
On the derivation of the foveal receptor primaries
Vision Research, Vol.11, pp.799-818.

Vos, J.J. (1979)
Line Elements and Physiological Models of Color Vision
Color Research and Application, Vol.4, No.4, pp.208-216.

Wainer, H. and Francolini, C.M. (1980)
An empirical enquiry concerning human understanding of two-variable color maps
The American Statistician, Vol.34, No.2, pp.81-93, May.

Wallis, R.H. (1975)
Film recording of digital color images
Ph.D. thesis; Univ. of Southern Calif. IPI Tech. Rep. USCIPI No.570.

Walraven, J. (1980)
Perceived colour under conditions of chromatic adaptation:
evidence for gain control by π mechanisms
Vision Research, Vol.21, pp.611-620.

Walraven, J. and Werner, J.S. (1982)
Chromatic Adaptation and π Mechanisms
Color Research and Application, Vol.7, No.1, pp.50-52.

Ware, C. and Cowan, W.B. (1982)
Changes in perceived color due to chromatic interactions
Vision Research, Vol.22, pp.1353-1362.

Warn, D.R. (1983)
Lighting controls for synthetic images
ACM Computer Graphics (Siggraph), Vol.17, No.3, July, pp.13-21.

Wasserman, G.S. (1979)
The Physiology of Color Vision
Color Research and Application, Vol.4, No.2, pp.57-65.

Werner, J.S. and Walraven, J. (1982)
Effect of chromatic adaptation on the achromatic locus:
the role of contrast, luminance and background color
Vision Research, Vol.22, pp.929-943.

Whitted, T. (1980)
An Improved Model for Shaded Display
Comm. ACM, Vol.23, No.6, pp.343-349.

Williams, L. (1978)
Casting curved shadows on curved surfaces
ACM Computer Graphics (Siggraph), Vol.12, No.3, pp.270-274.

Wilson, H.R. and Bergen, J.R. (1979)
A four mechanism model for threshold spatial vision
Vision Research, Vol.19, pp.19-32.

Wolfe, J.M. and Owens, D.A. (1981)
Is accomodation colorblind?
Perception, Vol.10, pp.53-62.

Wolfe, J.M. (1983)
Hidden Visual Processes
Scientific American, pp.72-85, Feb.

Wright, W.D. (1981)
Why and How Chromatic Adaptation Has Been Studied
Color Research and Application, Vol.6, No.3, pp.147-152.

Wyszecki, G. and Stiles, W.S. (1967)
Color Science
John Wiley and Sons, Inc., New York, London, Sydney.