CHAPTER 1

# PERSPECTIVES ON COLOUR SPACE

JAN J. KOENDERINK AND ANDREA J. VAN DOORN

## *Preface*

*We attempt to give a bird-eye's view of 'colorimetry', a field that has existed (as a science) at least since the mid-nineteenth century. The first quantitative empirical work due to Maxwell and Helmholtz, the first comprehensive theoretical approach was due to Graßmann. Of course, these scientists based their studies on earlier work, of which that due to Newton is perhaps the most influential. It is interesting that there have been parallel developments that can be traced from Goethe and Schopenhauer to Ostwald and (perhaps) Schrödinger, with a branch off due to Hering. Important modern developments are due to Cohen (1970s) and Schrödinger (1920s). It is interesting that these parallel developments have remained largely isolated, with—to our mind—detrimental effects on the field. Another aspect that has worked (strongly we believe) against the development of colorimetry as a science has been its extreme anthropocentric orientation. For instance, no one has developed tetrachromatic colour vision (as many animals have) to any serious extent, no one has explored the consequences of variations in the nature of the action spectra, and so forth. Here we attempt to present a balanced perspective on the field, although we foresee that many cognoscenti will consider this essay quite besides the point, and generally in bad taste. Our perspective is not that of the professional colorimetrist (we don't speak the proper CIE language), nor the experimental psychologist (we present no data on perception) or neurophysiologist (we don't deal with the neural substrate at all). Our interest is more that of the interested amateur with a general background in the exact sciences. Our ideal would be to write an authorative text on the essential structure of colorimetry of the type one reads as a student in physics curricula (for instance classical mechanics or electromagnetic theory: excellent texts abound). In our opinion no such a thing exists in colorimetry. Of course, the present essay represents only a preliminary, feeble attempt.*

*J. J. Koenderink and A. J. van Doorn*

## Introduction

In this chapter we will consider the following sundry questions from the field of 'colorimetry':

- Why are there so many 'colour solids' (A perusal of the literature [3] yields examples of cubes, spheres, pyramids, double pyramids, cones, double cones, colour trees, and others.) Is any one type to be preferred?

- Why are most colour order systems based on the 'colour circle', that is a *periodic* linear sequence [36,37], whereas the *spectral colours* are naturally ordered as a linear, open segment [31,32]? (See Fig. 1.1.)
- Why is the most basic dichotomy recognized by the artist, that of the 'warm' and 'cold' colour families, not reflected in the colour spaces of 'official' colorimetry?
- Are Newton's 'homogeneous lights' (the spectral colours) [31,32] to be considered especially basic? What is the position of other contenders, such as Goethe's 'boundary colours' (*Kantenfarben*) [44,45] or Ostwald's 'semichromes' (*Farbenhalb*)[34]? Is white to be considered a 'confused mixture' or the simplest colour imaginable?
- Apart from physiological considerations (fundamental response curves), the colour spaces of colorimetry are only affine, that is to say arbitrary linear deformations are to be considered irrelevant [6]. Is there any way to define a preferred ('canonical') basis? Does a 'natural' metric exist?
- Is there any principled way to mensurate the colour circle, or can this only be done by 'eye measure' (which places it outside the realm of colorimetry proper)?
- Are the colours of colorimetry a truthful (though limited) reflection of the physical structure of the radiation? That is, colour vision—in the approximation of colorimetry—simply a form of low-resolution spectroscopy, or does the observer's share go beyond this?

Apart from these key questions, we will have occasion to discuss several related, more technical questions. We assume that the reader has scant knowledge of the technical aspects of colorimetry. We provide an introduction that stresses the conceptual issues and will suppress technicalities, especially those involving extensive formal, mathematical notation. Because most of the formal structure of colorimetry is essentially of a geometrical nature, we will make up for the lack of formalism through illustrations that should yield an intuitive grasp of the structures involved.

## Colorimetry

### The world of colour

When we open our eyes we see the world around us, its geometrical layout (important for navigation) and various objects of possible importance to our continued existence. The objects have geometrical (size, shape) as well as material (colour, texture) attributes. These objects are involved in many processes, that is to say, our visual world is in a continual state of flux. If we are able to assume a painter's attitude, we may succeed (at least somewhat and for short periods) in perceiving the world as a two-dimensional array of coloured patches [39]. The 'colours' can be perceived (with effort) as essentially meaningless (i.e., not particularly favourable to effective optically guided behaviour), shapeless and textureless qualities, appearing in such and such a direction [12,39]. This is certainly not a natural state of affairs. In real life we deal with processes and objects, and 'colour' is used to label particular objects or classes of objects (red fireengines, blue-eyed blondes, ripe and unripe apples, etc.). The 'world of colour' is a mess that we won't be concerned with in

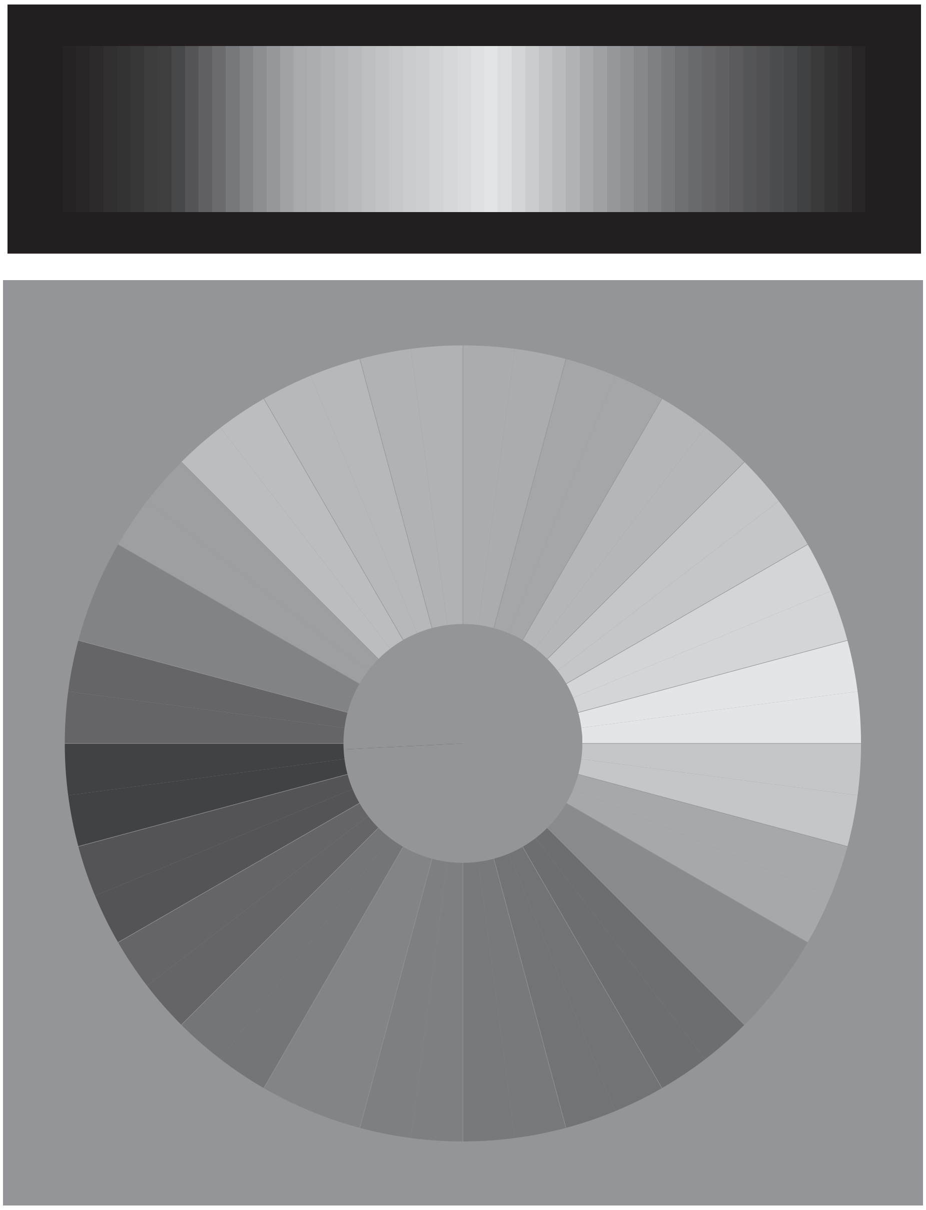

**Figure 1.1** The spectrum and the colour circle. The spectrum is an open, linear segment. Notice how the colours merge into black at each side. The spectrum is not complete as the purples are missing. The colour circle is complete by construction. It is a closed, continuous (thus periodic) arrangement. All hues are as colourful and bright as the printing permits (See colour Plate 1 in the centre of this book). The relation between spectrum and colour circle is a major topic of this chapter.*

*Colour versions of all figures in this chapter can be found in the website: http://www.phys.uu.nl/~wwwfm/Navigation/Resframes.html.

this paper. Instead we will consider the austere discipline of colorimetry[1] which applies a thoroughly Procrustean method [47] to obtain something essentially simple and elegant (though perhaps not very interesting).

### The physics of radiation

The physics of vision is rather simple, at least it is in its essential traits. The typical situation for the terrestrial observer is as follows: the observer and the objects are immersed in a transparent medium. A source of radiation (the sun) irradiates the scene. Radiometric interactions redistribute the radiation over the scene in complicated ways; moreover, the radiation interacts with the materials (surfaces of objects, the medium). Eventually we may pool all this in a single function, the *spectral radiance*. This function specifies, for any vantage point and any direction, the nature of the radiation that can be sampled by a directionally sensitive detector such as the human eye [13,14]. We may conceptualize the spectral radiance as a huge filing cabinet that contains photographs taken with arbitrary filters, from arbitrary points, in arbitrary directions. (Something like this is actually provided nowadays via satelites observing the Earth from orbit: nobody is actually in a position to 'process' all these data.) In practice, the observer samples only part of the available spectral radiance, that is to say, only from a limited number of vantage points, in a limited number of directions, with limited spatial resolution and with a limited spectral resolution and coverage of the electromagnetic spectrum.

In the setting of this chapter we will only be concerned with what we will denote 'beams'. A *beam* is a spectral radiance in such and such a direction, i.e. the (normal) observer will most likely see a 'patch' in that direction 'caused' by the beam. In order for the observer to actually 'see' the patch, the beam must be of sufficient radiance, of the right spectral composition, have the correct geometrical configuration, the observer must have open eyes, must look in the right direction, not be blind, and so forth.

Thus (in the parlance of psychophysics) the 'beams' are the 'stimuli', whereas the 'patches' are the 'percepts'. Of course percepts are only empirical facts for the perceiver (they are personal and 'immediately given'), and in order for colorimetry to constitute a scientific endeavour we need to substitute an objective and operational alternative [7].

The physical parameters that describe a beam are its *direction*, its *extent*,[2] and its *spectral composition*. In colorimetry we typically fix direction and extent, thus we may ignore these parameters. We are then left with spectral composition, or, more precisely, the 'spectral radiant power density'. This is a measure of radiant power, that is to say, radiant power density is measured as radiant energy per unit of time, unit of area and unit of solid angle. Instead of radiant energy, one may exploit the discrete nature of radiation and count

[1] 'Colorimetry' is a discipline that is far removed from 'colour science'. For one thing, it has nothing to say about colour appearances. Whereas colorimetry is a scientific discipline, 'colour science' is—like everything that calls itself a 'science' (e.g. Christian Science)—not science. That doesn't mean it has no intrinsic interest. On the contrary, if you are of the opinion that colorimetry is fairly trivial and not overly interesting, we find ourselves in agreement.

[2] The technical notion is 'throughput' or '*étendue*', basically the capacity of the beam to sustain rays. Informally, the radiance times the *étendue* denotes 'the number of rays' in the beam [13,14].

'photons' per unit of time, unit of area and unit of solid angle. Then the definition of radiance involves only time and space and is indeed very elementary.

In order to specify the *spectral* density we need to analyse the radiation in terms of its wavelength composition. This is usually done by means of 'monochromators' or 'spectroscopes' (Newton's experiment with the prism is the paradigmatic case). The beam is decomposed into other beams with the special property that they only show up radiant power in limited spectral regions.

The wave nature of electromagnetic radiation can be exploited by analysing the beams in terms of 'frequency', this is simply related to the 'wavelength' in vacuum[3] (or, to a good approximation, air). In the 'spectral analysis' we measure radiant power in limited frequency or wavelength intervals. The relevant measure is *spectral density*, that is, the radiant power per frequency or wavelength interval, e.g. per 10 nm. Notice that one should not say that beams are *composed* of 'monochromatic beams', but only that they can be subjected to spectral *decomposition*.[4] (A sausage can be cut into slices, but the (uncut) sausage is not composed of slices!) This is illustrated by the fact that the variance of the spectral density becomes arbitrarily large when we decrease the wavelength interval: thus the 'monochromatic beam' becomes *less* well defined the more we try to isolate it!

Sunlight is essentially 'noise' due to the incoherent superposition of myriads of micro-events in the photosphere of our sun. It has a continuous spectrum. In this case it is evident from the physics that it should not be conceived of (in the ontogenetic sense) as the superposition of many monochromatic beams, but rather as the superposition of many very short-lived *pulses* (much like the acoustical signal due to an applauding audience). Such pulses each have broadly distributed spectral power. Spectral decomposition is, as always, possible, but does in no way reveal 'elementary constituents'.

The discrete nature of the radiation can also be exploited in spectral analysis. Here one counts photons in limited regions of photon energy. Similar considerations apply as discussed above. The photon energy is simply related to the wave frequency via multiplication by Planck's constant.

The 'visual region' involves wavelengths (in vacuum) of about 380–740 nm, frequencies of about $4–8 \times 10^{14}$ Hz, or photon energies of about 1.5–3 eV. This is a biologically highly relevant wavelength range, for many reasons. For instance, sunlight peaks in this region, thermal radiation from the animal's own body can be neglected, photon energies are in the range of biologically important chemical bonds, and so forth.

### What colorimetry is

Historically, colorimetry evolved as a branch of photometry [26]. Thus the methodology derives essentially from early photometric practice. In photometry it was soon realized

[3] The frequency $\nu$ and wavelength $\lambda$ are simply related as $\lambda\nu = c$, where $c$ denotes the speed of light in vacuum. The wavelength in a medium of refractive index $n$ is $\lambda/n$. The energy of a photon is $\varepsilon = h\nu$, where $h$ denotes Planck's constant (*Wirkungsquantum*).

[4] As remarked in the text, whereas a sausage can certainly be cut into slices, the (uncut) sausage can hardly be said to be composed of slices. Radiation can indeed be decomposed into monochromatic beams, but that doesn't indicate that these are more 'elementary' than the beam itself, nor that the beam should be 'composed' of monochromatic beams. Since the theories are *linear*, we can just as well decompose monochromatic beams into other functions as long as we have a sufficiently large collection of them.

that although observers are quite bad at estimating radiant power, they are dependable as 'null indicators', that is to say, they can judge reliably the equality of radiant power in simultaneously perceived patches. A look into the eyepiece of a paradigmatic photometer reveals a circular patch in a dark surround, divided into two hemifields. The observer's task is to distinguish between a uniform patch and a bipartite one, i.e. to judge whether the division between the half-fields is noticeable. Colorimetry further pushes this paradigm to its limits.

In the colorimetric paradigm, two beams are compared by shaping them into patches that appear as contiguous hemidiscs on a black (radiationless) background. The patches appear to the observer as textureless and at no particular distance or attitude (not like the surface of any object). The task is simply to distinguish between a uniform disc and a bipartite field with different colours (patches) on both sides of the division. When the division is not noticeable, one says that a 'colorimetric equation' between the two beams has been obtained. We will write this fact formally as $\mathcal{A} \Leftrightarrow \mathcal{B}$, therewith indicating that the beams $\mathcal{A}$ and $\mathcal{B}$ are not distinguishable. The 'colorimetric equation' $\mathcal{A} \Leftrightarrow \mathcal{B}$ in no way implies equality of the radiant spectral power densities of the beams $\mathcal{A}$ and $\mathcal{B}$. Here we may read that the beam appearing on the left-hand side of the equation fills the left-hand hemidisc, whereas that appearing on the right-hand side of the equation fills the right-hand hemidisc.

Beams can be 'added' by simply superimposing them (for instance, one may take two slide projectors and project their images on the same screen). Such 'incoherent superposition' leads to a simple addition of the radiant spectral power densities [4]. The physics is really easy. It is an empirical fact that if two beams $\mathcal{A}$ and $\mathcal{B}$ cannot be distinguished in the colorimetric paradigm, then the beams $\mathcal{A} + \mathcal{C}$ (here we indicate incoherent superposition of beams by the '+' sign) and $\mathcal{B} + \mathcal{C}$ are also indistinguishable, quite independent of the nature of the superimposed beam $\mathcal{C}$. This suggests a simple complication in our notation. Let $\mathcal{A}$ and $\mathcal{B}$ be indistinguishable beams, then we have $\mathcal{A} \Leftrightarrow \mathcal{B}$. We also write this as $\mathcal{A} - \mathcal{B} \Leftrightarrow \mathcal{O}$, where $\mathcal{O}$ is the 'null beam', i.e. not any beam at all, but the absence of radiation. That this is a reasonable equation is clear when we superimpose $\mathcal{B}$, for then we obtain $\mathcal{A} \Leftrightarrow \mathcal{B}$, since the beam $\mathcal{O} + \mathcal{B}$ is in no way different from beam $\mathcal{B}$ (adding nothing to a beam doesn't change it). This explains plus and minus signs in colorimetric equations: to get rid of the minus sign add beams at both sides of the equation, thus obtaining a realistic situation, which is the one described by the equation (see Fig. 1.2).

Colorimetry is simply the investigation of beams via the colorimetric paradigm. Its eventual goal is nothing more than to be able to predict the truth value of any colorimetric equation involving any conceivable physical beams [6] (but also nothing less!).

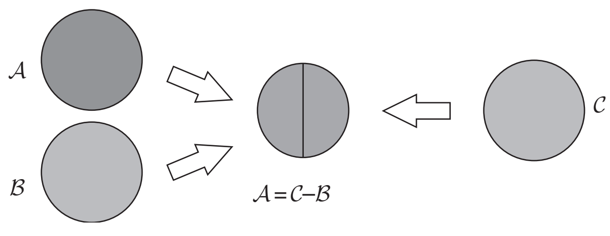


**Figure 1.2** The convention of 'negative' coefficients in colorimetric equations.

Alongside our operational definition of the addition of beams we may also operationalize the multiplication of a beam with a scalar (a 'scalar' is simply a number). This simply changes the total radiant power in the beam but not the relative spectral radiant power density (the 'quality' of the radiation). There are various ways of doing this, one example is to use a 'neutral density filter' ('sunglasses') that *attenuates* the beam. When we start with very intense beams, then virtually all beams used in practice will be attenuated and multiplication with a scalar is operationally well defined.

Since we have a null element (total darkness or no beam at all) and a linear addition and multiplication with a scalar, the space of beams is a 'linear vector space' [2,43]. However, not all elements of this linear space correspond to beams, e.g. when $\mathcal{A}$ is a beam then we can certainly *not* find 'the beam $-\mathcal{A}$', simply because 'negative radiation' doesn't exist. Yet we will admit such entities and call them 'virtual' or 'imaginary' beams, then any beam that *can* be realized will be called a 'real' beam. Thus the space of real beams becomes a subset of the (linear) space of beams. This formal artifice is most convenient because we have so many tools that deal with linear spaces effectively.

**What colorimetry is not**

It is often held that colorimetry involves an extreme stimulus reduction and, for that reason, does not bear on the issue of 'colour appearances', i.e. the descriptions observers venture regarding the colours of objects in their ken. It is indeed the case that colorimetry has nothing to say concerning colour appearances: that is both the reason for its phenomenal success and its fundamental limitation. However, it is not the case that such is due to *stimulus reduction*. Rather, the success (and limitations) of colorimetry are due to extreme *response reduction*.

The reason why people so often stress the stimulus reduction (which certainly applies to colorimetry as conventionally practised) is related to the literature on the 'modes of appearance' of patches [12,17,39]. A patch can appear as a non-localized light (self-luminous), as an irradiated material surface, and so forth. Such modes of appearance depend critically on the context in which a patch appears [25]. Many conventional colour terms (grey, brown) cannot even be applied without such a context present. In the conventional colorimetric setting the patches appear as non-localized, non-material 'film colours' or 'aperture colours'. In such a minimal context, one can have no greys or browns for instance—the world of colours has shrunk to minimal proportions [35].

But exactly because the colorimetric paradigm doesn't require the observer to venture a colour description, the modes of appearance are essentially irrelevant to the paradigm. It is only required that the observer judges indistinguishability of patches: this is a severe form of task or response reduction. It is easily conceivable to perform colorimetry in a natural setting, with full context available. Suppose one presents the observer with a fully textured visual array (a landscape, say) and that one successively alternates a tiny patch (of such a size that it is nearly uniform, say one 'pixel' of a digital image) with a reference, fiducial patch. Then the observer's task is to notice whether that patch appears steady, or alternating in time (toggling between two colours). Such a patch may well appear grey or brown to the observer (there's plenty of context available), but the observer is never asked to comment upon appearances in the first place.

We stress that the fact that colorimetry works is due to a rather extreme form of response reduction and that the severe stimulus reduction typically involved in conventional colorimetry is not essential at all (though it may serve to improve precision and reliability of settings, etc.).

The causal explanation is physiological:[5] colorimetric equations imply equal stimulation at the level of the 'retinal photoreceptor action spectra'. Thus colorimetric equivalence implies identical input to the brain. Consequently, colorimetry involves only the flimsiest interface between the physics and the physiology [7]. In a sense, colorimetry does not involve any neural processing at all and is almost pure physics except for the 'accidental' form of the cone action spectra, which is an empirical datum of physiology.

**The results of elementary colorimetric investigation**

Colorimetric equations can be obtained with a precision and repeatability that is rare in experimental psychology. The ultimate reason is simple: colorimetric equations don't depend on neural processing. If we speak of 'colours' in the context of colorimetry, we don't mean perceptual qualities in any sense, but something very specific (and, many people would say, dull and uninteresting): a *colour* is an equivalence class of beams that can replace each other in colorimetric equations. This is a sensible definition because of the strong empirical evidence that different beams $\mathcal{A}$ and $\mathcal{B}$ (say), such that $\mathcal{A} \Leftrightarrow \mathcal{B}$, may be substituted for each other in any colorimetric equation (for instance, $\mathcal{A} \Leftrightarrow \mathcal{B}$ and $\mathcal{A} + \mathcal{C} \Leftrightarrow \mathcal{D}$ together imply $\mathcal{B} + \mathcal{C} \Leftrightarrow \mathcal{D}$). This definition of 'colours' would really be trivial if colours corresponded in a 1–1 fashion to beams. However, this is very far from being the case: there are infinitely many more beams than colours. The surprising empirical fact is that very many different beams look the same. A colour corresponds to a 'metamer'; that is, an infinite set of indiscriminable beams. The interest, then, is to characterize the correspondence between colours and beams; that is, the partition of the space of beams into metamers, which are mutually disjunct subspaces by construction.

Thus we have 'beams', which are physical entities, which cause 'patches' to be seen, which are psychical entities (percepts). Then we have 'colours' in the sense of colorimetry, which are equivalence classes of beams, where 'equivalence' is to be understood via the colorimetric equations. The colours are *psychophysical* entities. Colorimetric equations, again, are based upon the indiscriminability of patches. Now 'indiscriminability' is operationally well defined and is to be considered fully *objective* because it is established by a second person ('scientist') who investigates the behaviour (confusion of patches) of the first person ('subject').[6] Thus colours, like beams, are well-defined objects for the exact sciences, although beams are physical and colours psychophysical objects.

[5] In this chapter we discuss only colorimetry proper. Of course there are also psychological, physiological, molecular biological, evolutionary, medical, etc. branches of colour study of great intrinsic interest. We will simply ignore them, for instance we won't refer to the 'cone action spectra' and so forth. In colorimetry proper there is simply no need, doing so would complicate the issues unnecessarily.

[6] The scientist may try to use all the tricks in the book to try to ascertain that the subject simply reacts like a reliable physical instrument. Thus 'catch trials' may be introduced, stimuli are presented in random order, and so forth.

The correspondence between colours and beams is characterized fully through a set of empirical laws first formulated by Graßmann [15]. Several alternative formulations have been framed. Perhaps the clearest formulation is the following: first, we notice that the space of beams $\mathbb{S}$ forms a linear vector space. Addition is defined by incoherent superposition, multiplication with a scalar by the introduction of non-selective attenuation (e.g. 'neutral density filters', like sunglasses). The zero element is no radiation at all, or total darkness. The dimension of this space is infinite (intuitively every wavelength is an independent dimension). Then Graßmann's laws boil down to the statement that the space of colours, $\mathbb{C}$ (henceforth called 'colour space') is a three-dimensional linear space, a *projection* of the space of beams. This completely characterizes the relation between beams and colours in a formal way.[7] It is only necessary to find the precise *projection operator*, which is—in principle—a simple enough experimental task. This has been done for a number of normal observers and an average has been canonized by committee (the CIE or Commission International d'Éclairage [9], in 1931). We will use the CIE 1964 10° data for the illustrations in this chapter.

In the above discussion we have schematized things a little, for the sake of initial clarity. First of all, we have omitted 'Graßmann's fourth law', which has a rather separate status. This 'law' states a rule to calculate the 'luminance' of any beam and thus addresses the problem of heterochromatic photometry. The empirical status of this law is in serious doubt. However, because luminance is so important (the customers of the electricity supplies pay for it, it has even a legal status), the law has been instated by committee [8] and is thus true *by definition*. Notice that this makes 'luminance' a purely formal entity. It doesn't have any meaning in terms of the perceptual attributes of patches. In this chapter we will ignore the topic of luminance altogether; in our opinion (and in full agreement with Schrödinger's elegant treatment [40,41]) it doesn't belong to colorimetry proper.[8]

Another fact is that the beams fail to fill the linear vector space of beams. The reason is simply that 'there exists no negative light'. That is to say, given any beam $\mathcal{A}$ (say), the beam $-\mathcal{A}$ that would (theoretically!) yield darkness when added to $\mathcal{A}$, namely $\mathcal{A} - \mathcal{A} \Leftrightarrow \mathcal{O}$, cannot be realized as a physical (existing) beam. Any beam for which the spectrum contains negative spectral radiant power densities likewise cannot be realized. Real beams are characterized by strictly non-negative spectral radiant power densities. That is not to say that non-realizable 'virtual' or 'imaginary beams' don't have their uses—they often occur in colorimetric calculations. However, they have no reality as such, only when they occur in combinations that *are* realizable. Geometrical intuition suggests that all beams fill an '$\infty$tant' (after 'quadrant' in the plane and 'octant' in three-dimensional space) in the linear space of beams characterized by the non-negativeness of all coordinates. Since colourspace, $\mathbb{C}$, is a projection of $\mathbb{S}$, it is geometrically evident that all (really existing)

7 Here we ignore the (both important and interesting) historical development completely. Instead, we try to focus immediately on the modern views on the matter.

8 Here we find ourselves in full agreement with Schrödinger [40,41]. Sadly, preciously few others agree, because luminance is of such enormous utility. However, we think that one should not turn a blind eye to its very low (really rock bottom!) scientific status. It really detracts greatly from the formal elegance (combined with practical utility) of colorimetry proper. Often luminance is so intricately woven into the formal development of colorimetry that it becomes quite unclear what the scientific status of the concepts really is. It is to such formulations of colorimetry (all too frequent) that we object here.

colours are confined to a convex conical volume in the space of colours, the projection of the $\infty$tant. Colours outside this cone may be called 'virtual' or 'imaginary', the others 'real'. These virtual colours are often handy in calculations, but can't occur as such. They only occur in mixtures that *are* realizable.

In the above we have already referred to 'convexity'. This is an important and fundamental issue in colorimetry. Most sets occurring in colorimetry have an essential linear structure, but fail to be linear spaces (by a narrow margin), in the sense that not any point in the linear manifold actually occurs in real life. Only some points are 'real', many are merely 'virtual' or 'imaginary'. The simplest example (and one that may be considered paradigmatic) is the space defined as a 'monochromatic beam'. This space is one-dimensional, for the only freedom is the radiant power of the beam, its nature being fixed (through its wavelength). Thus the space representing a monochromatic beam is a line. However, closer scrutinity reveals that only half of the line represents real beams, the other parts are merely virtual. The reason is simply that the radiant power is strictly non-negative. The set of real beams is convex, it is a half-line including the origin. When two beams, $\mathcal{M}_1$ and $\mathcal{M}_2$, are in the set of real beams, then so is $\alpha\mathcal{M}_1 + (1-\alpha)\mathcal{M}_2$ for $0 \leq \alpha \leq 1$. This is obviously true because $\mathcal{M}_1$ and $\mathcal{M}_2$ must be (positive) multiples of a single beam of unit radiant power, and we can immediately calculate the multiple for the mixture (it is a function of $\alpha$, of course) and ascertain that it is again positive. Thus the monochromatic beams of a given fiducial wavelength form a convex subset of a one-dimensional linear space. This may well be the simplest example.

Other instances of convex sets in colorimetry are the set of real beams (an $\infty$tant in the linear space $\mathbb{S}$), the set of real colours (a convex cone in the linear space $\mathbb{C}$), and so forth. One important geometrical fact is that convexity is invariant under linear transformations, including projections. This is one reason why $\mathbb{C}$ inherits many of the structural properties from $\mathbb{S}$, that is from the physics of electromagnetic beams. Notice that whereas a linear space is necessarily (except for the trivial case of zero dimension) of *infinite* extent, convex sets may also be finite volumes.

Another important geometrical tool is the construction of a 'convex hull'. The convex hull is the set that can be obtained by linear combination of the type $\alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ with $0 \leq \alpha \leq 1$.

For instance, the colour cone is the convex hull of the (non-convex!) set of colours caused by monochromatic beams, the $\infty$tant of real beams is the convex hull of the set of monochromatic beams, and so forth. The convex hull of a finite number of colours is an example of a convex set of finite extent.

The shape of the cone of colours has to be determined empirically.[9] We find that its boundary consists of two connected components: The 'spectral cone' the generators of

[9] The structure of the spectral cone is to be considered 'accidental' from the perspective of colorimetry and optics. We simply have to accept the measurements. Formal investigations reveal that different structures are conceivable that would really upset the nice structure of colorimetry. Such structures would be 'pathological'. For instance, the spectral locus might not be simply connected, or fail to be 'convex' (in the sense that it would not be fully on the boundary of its convex hull). The fortunate 'accident' is no doubt due to evolutionary pressure, like the 'accident' that the retina is close to the focal plane of the eye's optics.

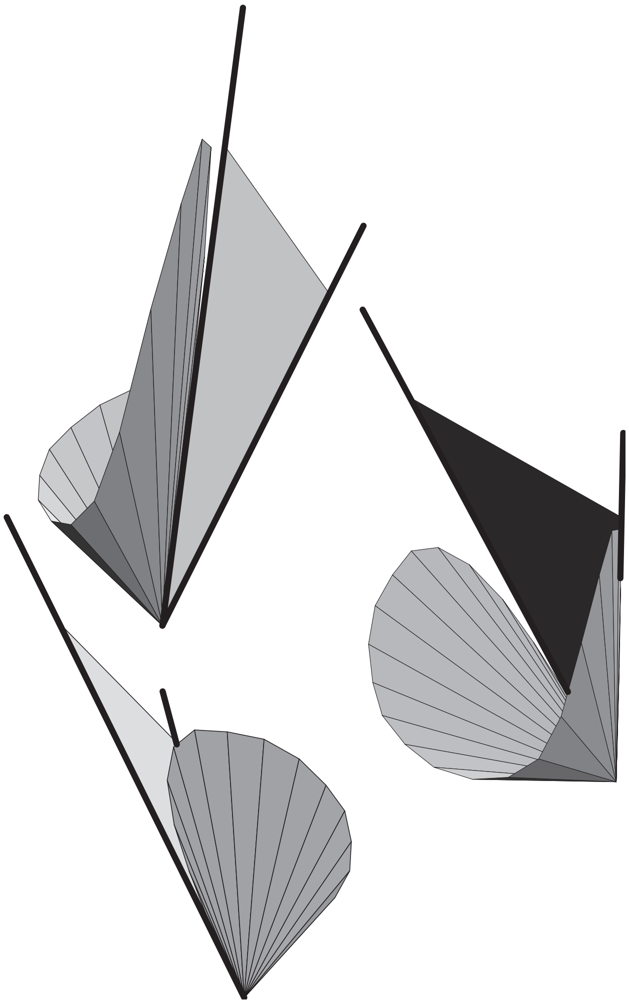

**Figure 1.3** Some views of the spectral cone in the CIE basis. Notice the plane of purples. (See colour Plate 2 in the centre of this book.)

which represent colours due to monochromatic beams, and a planar sector, the so-called 'plane of purples' (to be explained later) (see Fig. 1.3). All real colours find their place in the interior (convex hull) of the spectral cone [42].

If one attenuates a beam ('sunglasses'), one notices a decrease in 'brightness' whereas the 'colour', in some restricted sense, appears to remain invariant (we are talking about 'perceptual attributes' of patches here). This is the reason why one often finds it convenient to regard colours modulo their magnitudes. Thus we form equivalence classes $\mu\mathbf{c}$, where $\mu > 0$ and call them 'the chromaticities of the colours'. The black colour, $\mathbf{o}$, has an

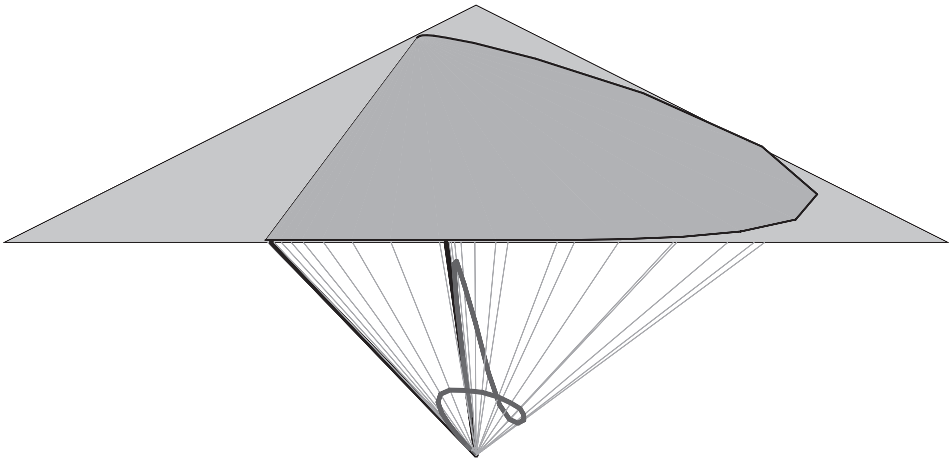

**Figure 1.4** The chromaticity diagram. The cone of colours is projected from the origin on the 'chromaticity plane', **c**. This figure illustrates the CIE basis and choice of chromaticity plane. The red space curve is the locus of the monochromatic beams of unit radiant power: It generates the spectral cone. (See colour Plate 3 in the centre of this book.)

undetermined chromaticity in this scheme. The chromaticities are thus half-rays at the origin of colour space. The set of chromaticities is only two-dimensional, thus we have gained some simplicity by factoring out the magnitudes. A convenient way to represent the chromaticities is to assign a 'chromaticity plane' and to represent each chromaticity by its representative in this plane (this is equivalent to a central projection from the origin of $\mathbb{C}$ on the chromaticity plane, see Fig. 1.4). The choice of chromaticity plane is essentially arbitrary, though it is necessary to take a plane that does not contain the origin, and it is convenient to take one that cuts the cone of colours into two: a finite volume (containing the origin) and an open, infinite part. Then the boundary of the cone of colours appears as a convex curve in the chromaticity plane. It consists of two parts. One part is curved and has a roughly horseshoe shape. This part represents the chromaticities of the monochromatic beams and is the 'spectral locus' in the chromaticity diagram. The other part is a straight line segment and contains the chromaticities of the purples. It is the intersection of the plane of purples with the chromaticity plane. All real colours have chromaticities in the interior (convex hull) of this curve. All points outside represent 'virtual' chromaticities. Notice that—unlike colour space—the chromaticity plane has no natural origin.

The chromaticity diagram has been very important in the history of colorimetry. Most reasoning used to be done in terms of chromaticities instead of colour space proper. Chromaticity diagrams are in common use in practical, everyday colorimetry. However, one should remember that it is only an *incomplete* representation of colour space and may easily lead the unaware astray. One example is the law of additive colour mixture. In $\mathbb{C}$ this is simply vector addition, nothing to it. However, in the chromaticity plane the law becomes rather more involved and beginners in colorimetry are apt to make mistakes that (unfortunately) often go unnoticed.

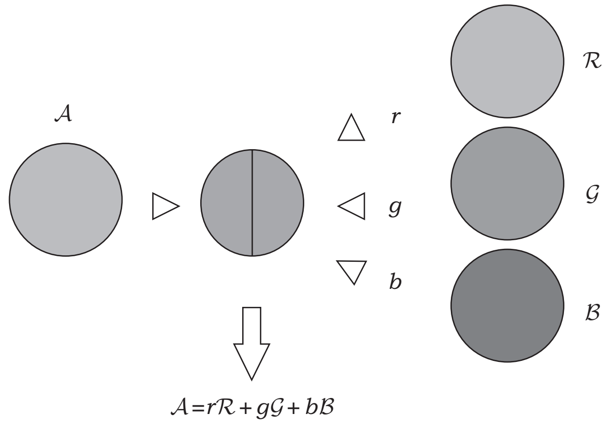


**Figure 1.5** The beam $\mathcal{A}$ is mixed from the primaries $\mathcal{R}, \mathcal{G}$ and $\mathcal{B}$. After adjustment to indistinguishability the fractions $(r, g, b)$ (pure numbers!) are the 'coordinates of beam $\mathcal{A}$ in the basis $\{\mathcal{R}, \mathcal{G}, \mathcal{B}\}$. At the setting of equality the photometer field needn't look like $\mathcal{A}$, in fact this will certainly not be the case if one of the coordinates turns out negative. This is irrelevant.

### Gauging the spectrum

Starting from Graßmann's laws, we may begin empirical colorimetric investigations. Maxwell [28,29] was the first person to seriously practise this. One proceeds as follows. First, one picks a 'basis' for $\mathbb{C}$. This means that one selects a triple of fiducial beams called the 'primaries' $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$. The choice is essentially arbitrary, except for the condition that no equation $c_1\mathcal{P}_1+c_2\mathcal{P}_2+c_3\mathcal{P}_3 \Leftrightarrow \mathcal{O}$ can be established; the primaries should be 'independent'. Next one takes each monochromatic beam of unit radiant power and wavelength $\lambda$, say $\mathcal{M}_{\lambda}$, in turn and establishes a colorimetric equation between it and the primaries. We write this equation $\mathcal{M}_\lambda \Leftrightarrow a_{\lambda 1}\mathcal{P}_1 + a_{\lambda 2}\mathcal{P}_2 + a_{\lambda 3}\mathcal{P}_3$. The coefficients $(a_{\lambda 1}, a_{\lambda 2}, a_{\lambda 3})$ are known as the 'trichromatic coefficients' (see Fig. 1.5). These are pure numbers, dimensionless as the physicist would say, namely simply the fractions of the primaries that enter into the colorimetric equation. We typically sample the wavelength scale at equal intervals at $N$ sample points (say), for clearly we cannot sample all monochromatic beams, nor is this necessary. We sample at wavelengths $\lambda_{\text{min}} + (i-1)\Delta\lambda$ say, where $\lambda_{\text{min}} = 380$ nm, $\Delta\lambda = 10$ nm and we stop at $\lambda_{\text{max}} = 740$ nm, say. In this case we have $N = (740 - 380)/10 + 1 = 37$. In practice, $N$ will be a large number, say of the order of $10^2$, it stands for the $\infty$ dimensionality of $\mathbb{S}$. Now we have a set of $N$ triples of numbers. We collect them in the $N \times 3$ matrix, $\mathbf{A}$ $(a_{ij}, i = 1 \dots N, j = 1, 2, 3)$, the so called 'colour matching matrix'. The columns of the colour-matching matrix are known as the 'colour matching functions', and the rows are the trichromatic coefficients for the corresponding monochromatic beam. Notice that the colour-matching matrix is obtained row by row, the columns only appearing after the whole process has been completed. This whole procedure is known as 'gauging the spectrum'.

Given the spectrum of any beam, sampled at the $N$ fiducial wavelengths, $\mathbf{s} = (s_1, \ldots, s_N)$ (say), we find the corresponding colour via the colour-matching matrix: $\mathbf{c} = \mathbf{A}^T\mathbf{s}$. Here $\mathbf{c}$ is an element of the three-dimensional space $\mathbb{C}$, whereas $\mathbf{s}$ is an element of the $N$-dimensional (notice that $N$ stands for '$\infty$' dimensions!) space $\mathbb{S}$. Thus the three components of $\mathbf{c}$ are simply the total radiant power when the spectrum of the beam is weighted by the corresponding colour-matching function. This is an immediate consequence of the linearity of Graßmann's projection of $\mathbb{S}$ into $\mathbb{C}$. The transpose of the colour matching matrix, $\mathbf{A}^T$ projects the spectrum to the space of colours, dropping $N - 3$ dimensions in the process.

Of course, the initial choice of primaries was quite arbitrary. Any other choice would have done as well. What difference does this make? Well, the colour-matching matrix will turn out different and the trichromatic coefficients for the same beam will turn out to be different too. Even when we change only a single primary, then (in general) all three colour-matching functions will change! Thus these numbers have no invariant meaning at all. However, it is easy to show (because of linearity) that the colour spaces for different choices of primaries are all *affinely equivalent*, that is to say, they are all related through linear maps (deformations, represented by non-singular $3 \times 3$ matrices) that depend only on the change of primaries. Such maps merely 'relabel' the colours and establish isomorphisms between the representations. Seen through 'affine spectacles' there exists only a single colour space and all specific representations only *appear* different, like the different perspectives of a single city.

Of course, this raises many questions. For instance, since all colour-matching matrices (that is colour-matching matrices for different choices of primaries) yield essentially the same colour space except for the inessential representation, what exactly is common between them? And if all representations of $\mathbb{C}$ are equivalent, might not one representation still be considered 'canonical' in some sense? We will return to these—indeed pertinent—questions.

### Warm and cold colours

Consider the colour-matching functions in some basis. We may plot them in $\mathbb{C}$, for clearly each triple of trichromatic coefficients represents a point in $\mathbb{C}$ (See Fig. 1.6). The colour-matching functions thus represent a one-parameter family of points, parameterized by the wavelength of the monochromatic beams of unit radiant power. Such a point set will be a *space curve* in $\mathbb{C}$. Indeed, when we plot the colour-matching functions in some arbitrary basis we obtain a smooth curve. This curve departs from $\mathbf{o}$ as the wavelength progresses beyond about 380 nm (the short wavelength, or 'blue' limit, of the visual spectrum), for shorter wavelengths are invisible (ultraviolet (UV), X-rays and $\gamma$-rays are equivalent to total darkness). Then the curve frolics around in $\mathbb{C}$, only to return towards $\mathbf{o}$ again when the wavelength approaches 740 nm, the long wavelength or 'red' limit of the visual spectrum (for infrared (IR), microwaves and radiowaves are also equivalent to total darkness). The curve leaves $\mathbf{o}$ and returns to $\mathbf{o}$ via different directions. We may construct the tangent directions

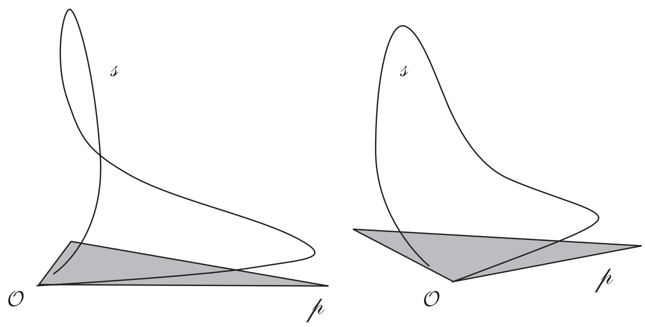


**Figure 1.6** Two views of the spectral locus and the plane of purples. The plane of purples ($p$) is spanned by the two tangents at the origin ($\mathcal{O}$) of the locus of equienergetic monochromatic beams ($s$). Notice how the curve first moves away from the plane of purples, then returns to it. The turning point is the division between the 'warm' and 'cold' spectral colours.

at the two limits of the visual spectrum. These two directions of course span a plane,[10] this is the 'plane of purples' mentioned earlier. Empirically we find that the curve (known as the 'spectral locus') runs on one side of the purple plane only—for the short wavelength end of the spectrum the curve moves away from the plane of purples, then at about 537 nm it runs parallel to this plane, then, for the long wavelength end of the spectrum it returns towards the plane of purples again[11] (Fig. 1.6).

This clearly separates all monochromatic beams into two families: those of wavelengths 380–537 nm and those of wavelengths 537–740 nm (these wavelength indications are only approximate). This dichotomy is quite independent of the arbitrary choice of primaries and is to be considered an invariant property of the structure of $\mathbb{C}$ and the spectrum of the achromatic beam. The latter can only be conventional, of course, say average daylight or a flat spectrum (as here) (the difference is actually slight).

It appears that this dichotomy coincides with the basic dichotomy conventionally recognized by visual artists, that is the division of all colours (here we should really be saying: '(coloured) patches') into 'warm' and 'cold' families. Colorimetry has nothing to say on why the monochromatic beams of 380–537 nm are termed 'cold' and those of 537–740 nm, 'warm'. However, the dichotomy with the transition at 537 nm ('greenish') is unlikely to be coincidental.

### An invariant of the colour-matching matrices

What do all colour-matching matrices (that is, for different choices of primaries) hold in common, that they may generate essentially the same colour space? The answer has to

[10] Thus the points in the plane of purples occur as additive mixtures of the limiting (far red and deep blue) beams. The problem is that *in the limit* we need infinite radiant power to actually *see* something. Thus the plane of purples is really a limiting entity, a tangent plane whose points represent colours that just fail to be 'real'.

[11] One might think that 'distance from the purple plane' were a notion that is alien to the affine nature of colour space. However, we can formulate a fully affine definition easily. The reason is that we can consider planes parallel to the purple plane in affine geometry, and also their order. This suffices to find the farthest point.

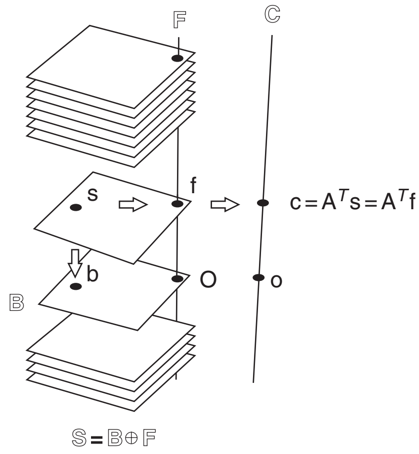


**Figure 1.7** **S** is the space of spectra, here it is represented as a fiber bundle, e.g. we have factored it as $\mathbb{S} = \mathbb{F} \oplus \mathbb{B}$. Notice that though **S** is ${}^{\infty}\mathcal{D}$ (or ${}^{N}\mathcal{D}$), here it is illustrated as ${}^{3}\mathcal{D}$. The ${}^{\infty-3}\mathcal{D}$ space, $\mathbb{B}$, is illustrated as ${}^{2}\mathcal{D}$ and the ${}^{3}\mathcal{D}$ space $\mathbb{F}$, is illustrated as ${}^{1}\mathcal{D}$, simply to obtain an intuitive geometrical picture. Colour space $\mathbb{C}$ is a ${}^{3}\mathcal{D}$ space, quite distinct from $\mathbb{S}$. Here it is illustrated as ${}^{1}\mathcal{D}$ to show the important fact that it has the same dimension as $\mathbb{F}$. In fact $\mathbb{F}$ and $\mathbb{C}$ are isomorphic. Notice that each point of $\mathbb{F}$ carries a translated copy of $\mathbb{B}$. This 'fiber over the fundamental' is the metamer defined by that fundamental. Thus we have represented the space of beams $\mathbb{S}$ as a stack of metamers, each metamer corresponds 1−1 with a colour. When we pick any beam, spectrum **s** (say) we may project it on $\mathbb{F}$ via Cohen's matrix **R**, thus obtaining the fundamental component **f**. We may also project it on the black space $\mathbb{B}$, thus obtaining the metameric black component **b**. Notice that (in $\mathbb{S}$) $\mathbf{s} = \mathbf{f} + \mathbf{b}$. The (transpose of) the colour matching matrix maps the beams on the colours, i.e. **s** on **c**. Notice that this is the same as the map of **f** on **C** since the black component **b** remains causally ineffective. Thus $\mathbb{C}$ is the space of beams $\mathbb{S}$ modulo the metameric black space $\mathbb{B}$.

be found in the parcellation of the space of beams into metamers. Clearly, there is only one such parcellation in terms of the geometry, although the metamers are given different coordinates by the various colour-matching matrices. The set of metamers as such must be identical for all choices of primaries because it doesn't depend on the primaries at all: it is, once and for all, given by the structure of the visual system, that is to say, the cone action spectra. Coordinatization should not matter.

A standard way to put order in this mess is to write the space of beams as a direct sum of two subspaces, 'fundamental space' ($\mathbb{F}$) and 'metameric black space' ($\mathbb{B}$), thus $\mathbb{S} = \mathbb{F} \oplus \mathbb{B}$ (see Fig. 1.7). The symbol '$\oplus$' stands for the 'direct sum' of the 'mutually orthogonal' subspaces $\mathbb{F}$ and $\mathbb{B}$. What does this mean? It means that we will write any beam with spectrum **s** (say) as the sum of two components, an (often virtual) 'fundamental beam' with spectrum $\mathbf{f} \in \mathbb{F}$ and an (always virtual) 'black' beam $\mathbf{b} \in \mathbb{B}$, thus $\mathbf{s} = \mathbf{f} + \mathbf{b}$. We will do this in such a way that the fundamental component yields the colour of the beam, that is to say $\mathbf{c} = \mathbf{A}^T\mathbf{s} = \mathbf{A}^T\mathbf{f}$, whereas the black beam yields the black colour (the origin of $\mathbb{C}$, hence the name), that is to say $\mathbf{A}^T\mathbf{b} = \mathbf{o}$. This can always be done and in a unique manner.

Formally speaking the black space is simply the *null space* of the transpose of the colour-matching matrix, and fundamental space is the orthogonal complement. Since the rank of the colour-matching matrix is 3, the null space (metameric black space) has dimension $N - 3$, whereas fundamental space has dimension 3, that is the same dimension as colour space. Fundamental space and black space together span the space of beams.

Conceptually we have now constructed a very elegant representation: any beam can be decomposed into a fundamental beam and a black beam. The black beam is *causally ineffective*, in the sense that it has no influence whatsoever upon the colour. The physiological explanation is that the black component does not stimulate the retinal cones and thus is unable to modulate nervous activity: the black beam never even reaches the brain! The fundamental component is the unique causally efficient part of the beam. Any nervous modulation is only due to the fundamental component. Moreover, different fundamental components yield different colours, thus the fundamental components stand in a 1–1 relation to the colours. Fundamental space is isomorphic with colour space. A colour corresponds to a *metamer*,[12] which is an infinite equivalence class of beams. In any metamer there exists one canonical representative which is the fundamental component, all other beams in the metamer are equal to the fundamental component plus some black beam. We obtain the full metamer ('metameric suite') when we add all possible black components (that is the metameric black space) to the fundamental component. Since fundamental space and metameric black space are mutually orthogonal, we can write the length of the spectrum $\mathbf{s}$ as $\sqrt{\|\mathbf{f}\|^2 + \|\mathbf{b}\|^2}$ (this is nothing but the Pythagorean theorem). Thus among all metameric beams the fundamental component is the 'shortest' one, in a sense the 'simplest' (thus canonical) representative.

We may then conceive of $\mathbb{S}$ as a fibred space, a three-dimensional linear space of fibres. (Fig. 1.7). Each fibre is a copy of $\mathbb{B}$, attached to a specific fundamental component. The visual system sees only the fibres (a three-dimensional manifold) but cannot resolve the structure *within* the fibres. The fundamentals are sufficient to represent the fibres, but *any* point within the fibre would do as well: the visual system doesn't notice the difference anyway. This is a rather pleasant way to put it because it shows up the fact that $\mathbb{F}$ is not really different from $\mathbb{S}$: it is simply a low-resolution image of it. Moreover, $\mathbb{C}$ is isomorphic with $\mathbb{F}$, that is, nothing but a linearly deformed copy of it. This is an important insight because it is often held that there is a fundamental cleft between the physical world ($\mathbb{S}$) and the world of perceptions ($\mathbb{C}$). This is evidently a nonsensical view since $\mathbb{C}$ turns out to be simply isomorphic to a subset of the space of beams. We will return to this issue later.

Clearly, fundamental space does not depend on the choice of primaries. Thus the null spaces of all colour-matching matrices must be the same: here we have the sought-for invariant of the colour-matching matrices. It is possible to compute an operator in the space of beams (a $N \times N$ matrix) from the colour-matching matrix that projects any beam on its fundamental component: $\mathbf{Rs} = \mathbf{f}$. The matrix $\mathbf{R}$ is 'Cohen's matrix $\mathbf{R}$'. This is a convenient numerical invariant of the colour-matching matrices. Indeed, if we calculate Cohen's matrix $\mathbf{R}$ from an arbitrary colour-matching matrix, we always obtain the same result (See Fig. 1.8). Cohen's matrix $\mathbf{R}$ enables us to calculate metameric beams of any given beam. For instance, consider beams $\mathcal{A}$ and $\mathcal{B}$ with spectra $\mathbf{a}$ and $\mathbf{b}$. Then the beam with spectrum $\mathbf{Ra} + (\mathbf{b} - \mathbf{Rb})$ must have the same colour as beam $\mathcal{A}$ because it is composed of

[12] The term 'metamer', which is in common use today, was coined by Ostwald [34].

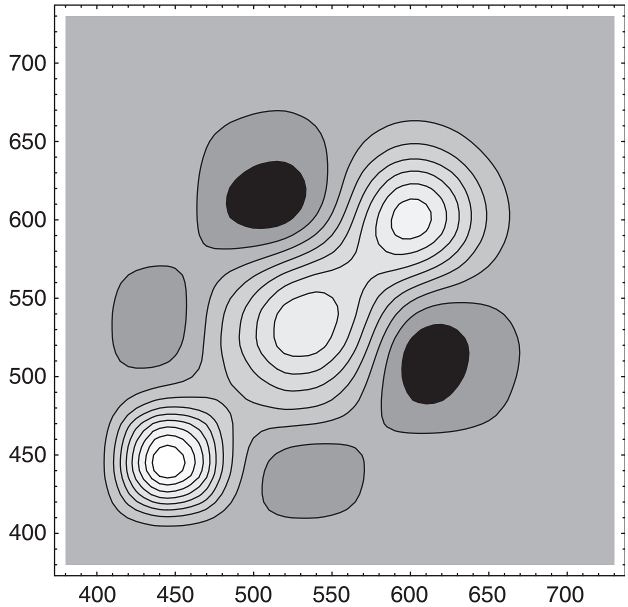


**Figure 1.8** Contourplot of the entries in 'Cohen's Matrix **R**'. Notice the three extrema on the diagonal, at about 440 nm, 530 nm and 590 nm: these are characteristic for the human visual system.

the fundamental of $\mathcal{A}$ and the black component of $\mathcal{B}$. Here the beam $\mathcal{B}$ is quite arbitrary! Here we have a recipe to generate colours in the metamer of **a** at mad abandon.

One thing to take good notice of is that the fundamental and black components of any given beam are unlikely to be physically realizable (the black beam obviously never will!) but are most likely to be *virtual* beams. If this is considered to be a problem, we can try to construct a beam as close to the fundamental one as possible that *is* realizable by adding a suitable black beam.[13] However, in most cases the fact that the fundamental component is only virtual is no big deal.

### Canonical bases

In order to motivate the discussion below we first consider an intuitive example from 'real space' instead of the space of beams.

Consider a linear transformation from three-dimensional space on to a plane. We suppose the three-dimensional space to be *metrical*, i.e. we are able to compare length in three orthogonal directions. Similarly, we assume the image plane to be metrical. Such a transformation is clearly a *projection*, not an isomorphism, for it drops one dimension. Its null space is a *direction* in the three-dimensional space. All points that lie on a line extended in

[13] There exist a variety of technical methods to handle this. The most effective are 'linear programming' and 'iterative projection on convex sets'.

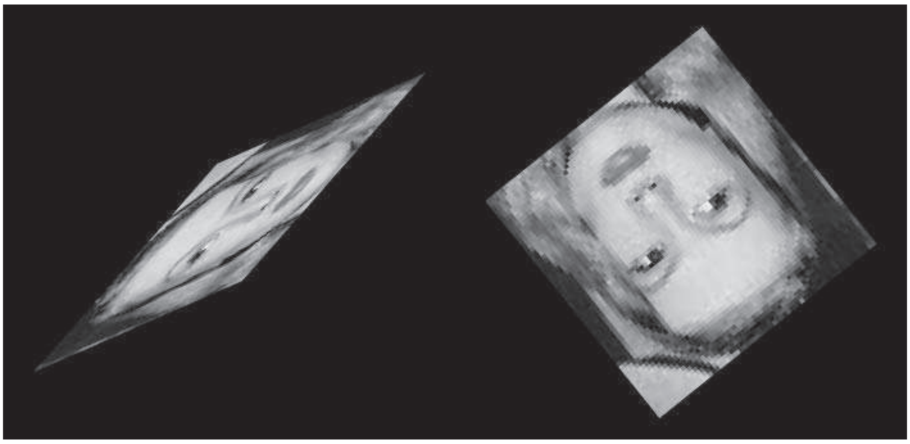

**Figure 1.9** On the left the result of a random projection (same matrix as used in Fig. 1.10) from three-dimensional space on the plane. The subject is a human head. Clearly there are strong deformations in this projection. On the right the same projection in a canonical basis constructed automatically via the 'singular values decomposition'. This projection is clearly free of deformations (except for the loss of the depth dimension) and is clearly to be preferred. The canonical basis is determined up to an isometry in the plane. There is no way to figure out automatically that human observers prefer a specific orientation of 'the vertical'.

this direction will be mapped on a single point of the plane, this is the basic notion of 'projection'. The orthogonal complements of the null space are planes in the three-dimensional space that are perpendicular (in a metric space perpendicularity is well defined) to the null space. We expect of a 'nice' projection that configurations in such a plane are transferred to the plane of projection without suffering any further deformation. For instance, if we take a photograph of a frontoparallel wall (using a long telelens to suit our example—almost 'parallel projection') we expect an 'undistorted image', we should only lose the 'depth'. Now our general linear transformation will fail miserably in this respect: Fig. 1.9 shows an example. It will induce some arbitrary deformation ('shear') that yields a highly distorted image. In order to do better we may attempt to find a basis in the plane (the image) that 'undoes' the deformation. There exist standard methods ('singular values decomposition' SVD) to find such a basis.[14] Such methods are used in many fields where it is sometimes unavoidable that images are distorted because of unfortunate imaging parameters (for instance, in photographing a skyscraper we may have to tilt the camera, which will have the effect of a distorted image, the skyscraper appearing as if it were tumbling down).

[14] Singular values decomposition has quickly become the premier tool of linear algebra. Standard methods are generally available that let one handle basically any matrix of practical interest robustly and in reasonable time. The method exploits the basic fact that any linear map between two 'inner product spaces' (linear spaces with a metric) can be simply factored. When the domain space has higher dimension than the target space, the map can be considered as the combination of a trivial map (dumping everything on a subspace) and a deformation (isomorphism). The deformation again can be considered as a simple rescaling of the axes when one picks the representations in the two spaces shrewdly. The 'SVD' ('singular values decomposition') does exactly this, it presents one with canonical bases for the spaces and a set of 'singular values' which are the required scaling factors.

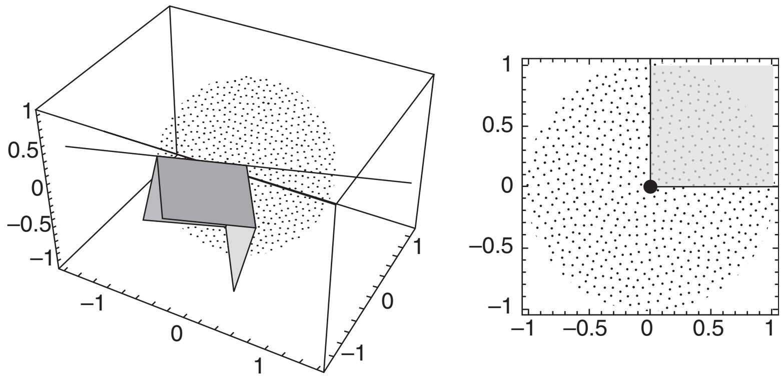


**Figure 1.10** A simple example of the SVD for a map from three- to two-dimensional space. Shown is a cloud of dots, uniformly distributed on the surface of the unit sphere in three-dimensional space. We picked a random map, given by a matrix with rows {0.2364, −0, 1011, 0, 6127} and {0.08872, 0.1462, 0.711}. The straight application of this map produces a cloud in the plane that fills an oblique elliptical area with 1:4.61: the 'image' of the sphere is highly deformed. A straight SVD produces bases in three-space (indicated in the left figure) and the plane (right figure) and scaling factors ('singular values') that let us draw a truthful (not deformed) image: the cloud fills a circular area. The nullspace ('black space') is shaded in the left figure; it is one of the directions of the canonical SVD frame.

We encounter a very similar situation in colorimetry. The colour-matching matrix discards the $N - 3$ dimensional metameric black space and maps the fundamental space in an 1–1 fashion on colour space $\mathbb{C}$. The only part of the space of beams that we will ever 'see' is fundamental space, $\mathbb{F} \subset \mathbb{S}$: *can we at least hope to see an undistorted image of it*? The necessary requirements are given: in the space of beams, $\mathbb{S}$, we can compare lengths in arbitrary dimensions, we have simply spectral radiant power density to compare. Singular values decomposition immediately supplies us with a basis in which $\mathbb{C}$ is an undistorted image of $\mathbb{F}$. We only lose the 'depth' (that is, the black component) but gain a truthful image of the beams otherwise, it is shown in Figs 1.10 and 1.11. The situation is much like the visual projection where we drop one dimension (the depth) out of three (breadth, height and depth); here we drop $(N - 3)$ dimensions (black) out of $N$ (the full spectral radiant power density). This shows that colorimetry is simply 'low-resolution (only 3 degrees of freedom) spectroscopy'. In this respect colour vision is indeed very close to the physics of radiation, the only arbitrary factor is the structure of the null space. What gets lost (misses its way to the brain) depends on the accidental (that is, shaped through evolution of the species) cone action spectra. In Fig. 1.12 we show an 'undistorted' version of the spectral locus and spectral cone.

There is still some freedom left: we may apply arbitrary isometries (rotations, reflections, etc.) to $\mathbb{F}$. Isometries leave all lengths and angles invariant. Similarly, the image in $\mathbb{C}$ will be subjected to an isometry, but all geometrical configurations will be unaffected by this. In summary, we are quite free to pick any orthonormal basis in $\mathbb{F}$ and this will change our $\mathbb{C}$. One unique choice is to pick the achromatic axis and to let another direction lie in the plane that divides the warm–cold colour families. The third orthogonal direction is then

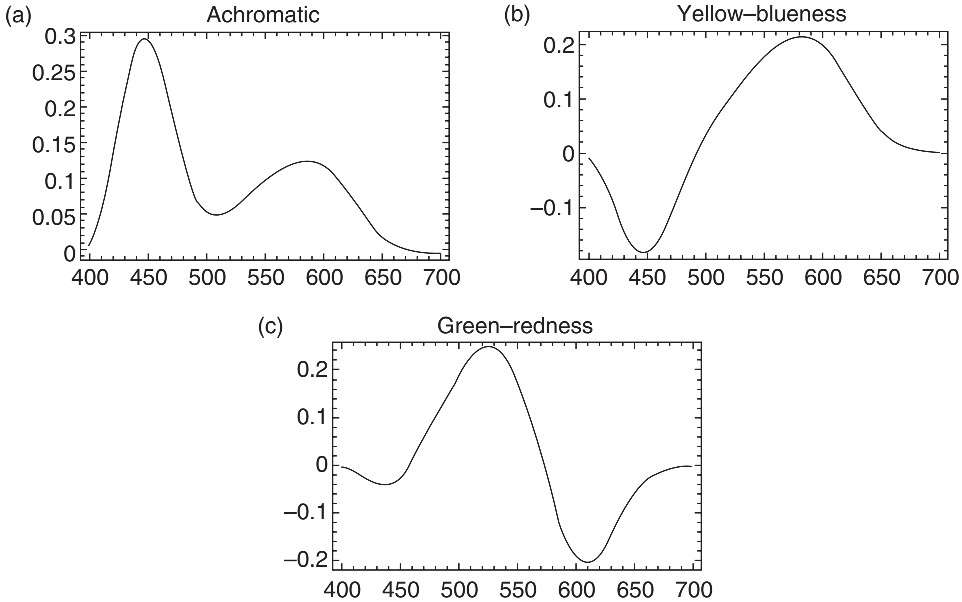


**Figure 1.11** The colour-matching functions of the canonical basis. These function are similar to the classical 'colour moment', 'red–greenness' and 'yellow–blueness'. The latter 'opponent signals' are similar to neural encodings found in the brain. Apparently the human brain uses a representation not unlike the canonical basis, an undistorted representation of the physical structure of beams.

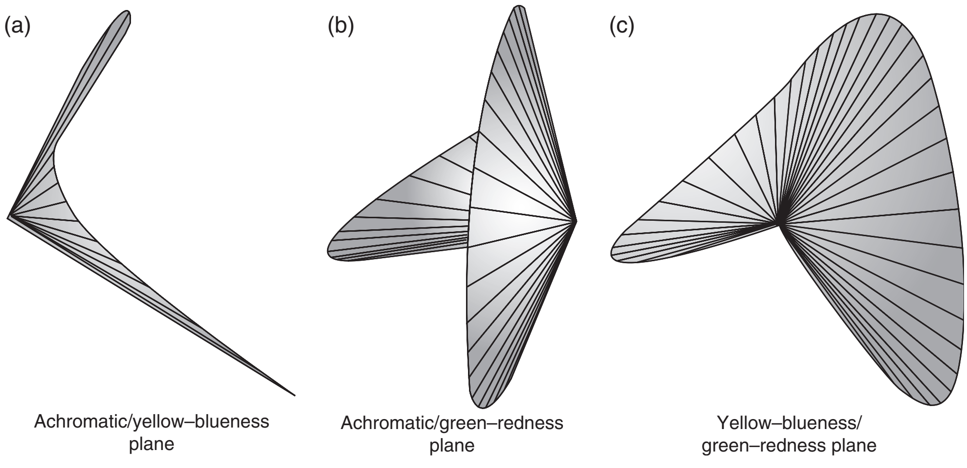


**Figure 1.12** The spectral cone in the canonical basis: an undistorted view.

automatically defined. The only remaining freedom is the orientation, but this can only be settled by fiat (like left and right: which is which?). Notice that this is a representation that is unique and independent of the (fully accidental) choice of primaries. We obtain a truthful image of $\mathbb{F}$ as our $\mathbb{C}$, that is indeed the same in *any* laboratory, because it is fully independent of our accidental choices except for the choice of the achromatic beam (more on that below) and the orientation (discussed above).

In Fig. 1.11 we show the 'primaries' (two of the three are virtual colours) that result from a straightforward singular values decomposition of the CIE 1964 colour-matching matrix and subsequent rotation to the canonical frame. The achromatic beam here is the uniform, flat spectrum. The first primary is simply the fundamental of the equal energy spectrum, the other two are yellow–blue and green–red 'opponent signals'. Notice that the zero crossings (about 490 nm for the yellow–blueness and 570 nm for the green–redness) are indeed close to those of Hering's *Gegenfarben*. In Fig. 1.12 we illustrate the spectral locus in the canonical basis. Here is an undistorted view of this important configuration. We're getting a limited, but true-enough glimpse of the space of beams. The second and third primary are indeed remarkably close to Judd's specification of the Hering system [21], but the first one clearly differs—apparently the Hering frame is far from being orthogonal.

That colorimetry is essentially just low-resolution spectroscopy shows that the discipline isn't really to be considered part of experimental psychology. Rather, it is to be considered simply (applied) physics, or—because of the accidental status of the structure of the black space—perhaps physiology.

The only arbitrary choice left is in the description of the physics of the situation. What is the 'natural representation' for the beams? This is important because it defines the metric of the space of beams which will *induce* a metric in the space of colours. This is not a question of physics but one of 'ecological optics'. As we have seen earlier, the spectral description can be done in terms of frequency (essentially photon energy, see below) or wavelength, whereas the amount of radiation can be measured as radiant power or as photon number density. The choice makes a difference to the eventual representation, but neither physics, nor colorimetry proper, have anything to say on the matter. The conventional choice is radiant power as a function of wavelength, but this is purely for historical reasons, dating all the way back to Newton [31,32]. A, perhaps more rational, representation would be photon number density as a function of photon energy. The reason is that the photoreceptors are essentially photon counters, whereas the photon energy is the major causal factor in the interaction of electromagnetic radiation with matter. Thus there are good biological reasons (that is the 'ecological optics') to prefer the latter representation.

## Additional structure: the achromatic beam

When we circumnavigate the boundary of the cone of colours (avoiding the origin) we experience a continuous change of 'hue'. Since the path is closed, the hues form a *periodic* linear sequence. Since real colours in the interior of the colour cone also have 'hues', one wonders about the loci of constant hue in colour space. They must be surfaces that intersect the boundary of the colour cone transversely. Some topological reasoning soon reveals that somewhere in the interior must exist a singular curve of points of undetermined hue.

One may designate such colours 'hueless' or 'achromatic'. It was already clear to Graßmann [15] that if we want to talk about hue we need an *achromatic locus* inside the cone of colours. However, it is clear that we cannot discuss these things in the context of colorimetry proper, for the simple reason that any talk about 'hue' oversteps the boundaries of the very response reduction that makes colorimetry a viable discipline.

One way out of this dilemma is to point out some fiducial beam as 'achromatic' in a fully arbitrary or deus ex machina fashion. The only requirement is that the beam doesn't map on the boundary of the colour cone. This move has the obvious advantage that it doesn't violate the response reduction (the subject is never asked to agree on the 'achromaticness' of the fiducial beam). It has perhaps the drawback of arbitrariness though. However it may be, such a definition allows us to add considerable additional structure to colour space. It is almost a necessity if we desire to proceed beyond the point that we have now reached. We will follow up the consequences in this chapter.

In practice, one designates a fiducial beam as 'achromatic' for some pertinent reason. For instance, one may pick 'average daylight', or a 'flat (equienergy) spectrum', or the spectrum of the illuminant that is important in a given setting. That is not to say that such a definition is not an arbitrary act from the vantage point of colorimetry. It is particularly useful to pick a spectrum whose spectral radiant power density dominates the spectra occurring in a given setting throughout the spectrum. This is a frequently occurring case in practice: one has the beam of an illuminant (say sunlight) and all other beams are created by taking away (that is attenuating) radiant energy from this spectrum. This is essentially Goethe's notion (dating back to the Greeks) that 'colours are shadow-like entities'. In such a case we assign the illuminating beam as the 'achromatic' one.

The chromaticity of the achromatic beam represents a half-ray that we will call the 'achromatic axis'. This immediately induces additional geometrical structure, namely a sheaf of planes that all contain the achromatic axis. Each such plane meets the boundary of the cone of colours along a generator that either represents the chromaticity of a monochromatic beam, or a purple one. We may thus label the planes with the corresponding monochromatic beam. In the case of a purple we simply extend the plane beyond the achromatic axis and find the monochromatic beam at the opposite side of the cone. We label the planes with the wavelength of the corresponding monochromatic beam, in the case of a purple we prefix a minus sign. One calls this the 'dominant wavelength' of any colour contained in the plane. By extending the planes beyond the achromatic axis we establish a relation between pairs of dominant wavelengths, conventionally these are called mutually 'complementary' wavelengths. In the chromaticity plane the achromatic beam is represented by an achromatic point, and the sheaf of planes by a fan of lines on this point. The dominant and complementary wavelengths are found as the intersections of the lines of the fan with the spectral locus and line of purples.

Using this construction we may add various geometrical configurations to the basic structure of colour space. First we notice that the plane through the monochromatic chromaticity of (about) 537 nm divides the cone of colours into two parts, the warm and the cold colours. Now we have extended the definition of warm and cold to *all* colours, not just the monochromatic ones. There is no way to do this without the achromatic axis and the construction indeed *depends* on the (arbitrary!) choice of this axis. Notice that we also

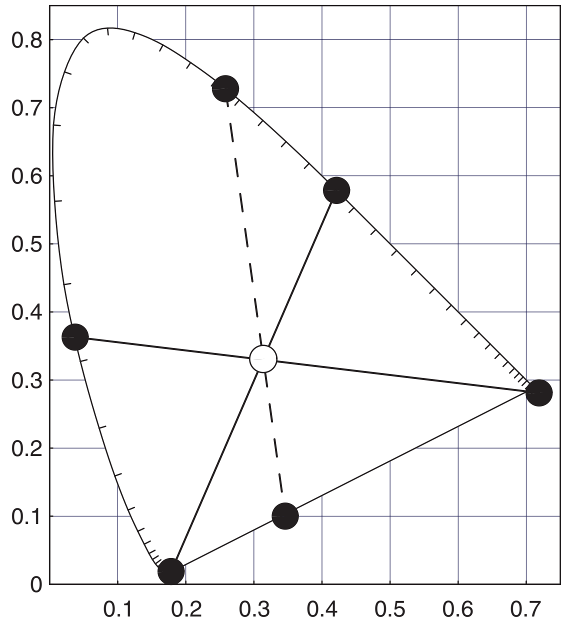


**Figure 1.13** The CIE chromaticity diagram with several remarkable objects: the purple line with the spectral limit points, the warm–cold division on the spectral locus (see colour Plate 4 in the centre of this book). The achromatic point defines the complementarities of the spectral limits and the plane that divides all chromaticities into warm and cold. All these objects depend for their existence on the introduction of an achromatic beam.

obtain a special purple, dividing the plane of purples into a warm and a cold part. Next we notice that the spectral limits ('red' and 'blue') are special generators of the spectral locus (clearly invariant against changes of primaries) and that these also define unique planes with the achromatic axis. Extending these to the opposite side yields two special chromaticities, namely the complementary of the short wavelength end ('yellow') and the complementary of the long wavelength end ('cyan' or 'blue–green'). These geometrical entities turn out to be very significant in the further development of the structure of colour space. They do depend on the (again, arbitrary) choice of the achromatic axis though (see Fig. 1.13).

## Newton's spectrum

From Newton's famous drawing of the spectrum we can estimate that the effective 'slitwidth' must have been somewhere between 50 and 100 nm. Thus the spectrum cannot have been very 'pure'! Yet one easily checks that the spectrum looks *fine* at such a slitwidth (as shown by Newton [31,32] ), despite the fact that the 'spectral beams' are not monochromatic beams ('homogeneous lights' in Newton's terminology) by a long shot, but rather 'confused mixtures' as Newton would have it. Indeed, when one tries to improve the situation by decreasing the slitwidth, the spectrum doesn't really look any better but it becomes dimmer

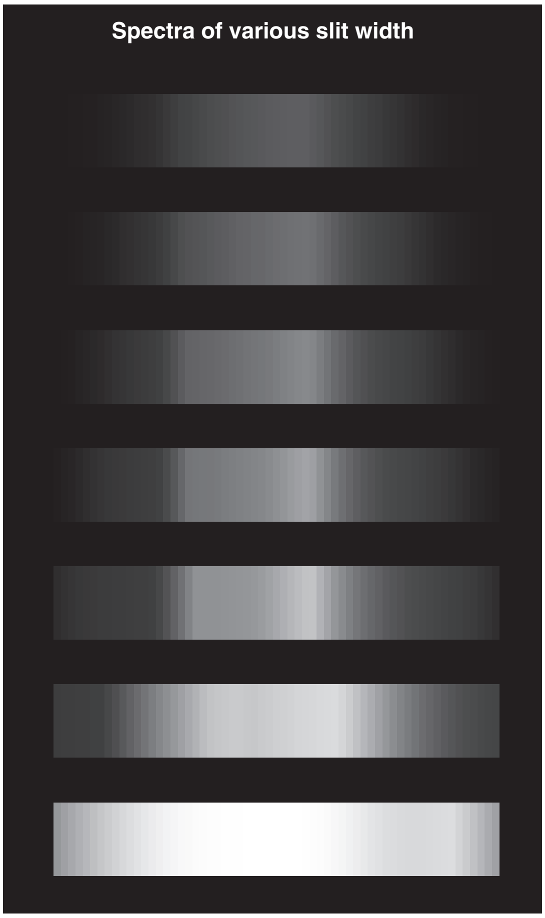


**Figure 1.14** The Newtonian spectrum at various slitwidths (see colour Plate 5 in the centre of this book).

and for very narrow slits (the 'ideal' situation), actually appears black. The reason is simple enough, a very narrow slit hardly lets any radiation pass, thus the beams approach the black beam $\mathcal{O}$. Only in very special laboratory situations does one ever get to see very pure spectra.[15] They really don't look any better than Newton's impure spectrum though, so the effort is really wasted. When one *increases* the slitwidth the colours become brighter (the slit passes more radiation), but from a certain point on they tend to 'desaturate', that is, they approach the colour of the entrance beam (let's call it 'white' for the moment). When the slit is very wide the spectrum is lost and one sees only a white beam. Clearly, there is some optimum slitwidth at which the spectrum appears 'most colourful' (see Fig. 1.14).

[15] One has to increase the radiance of the entrance beam enormously to get some radiation to enter the eye and screen the observer from it in order to avoid dazzling the visual system.

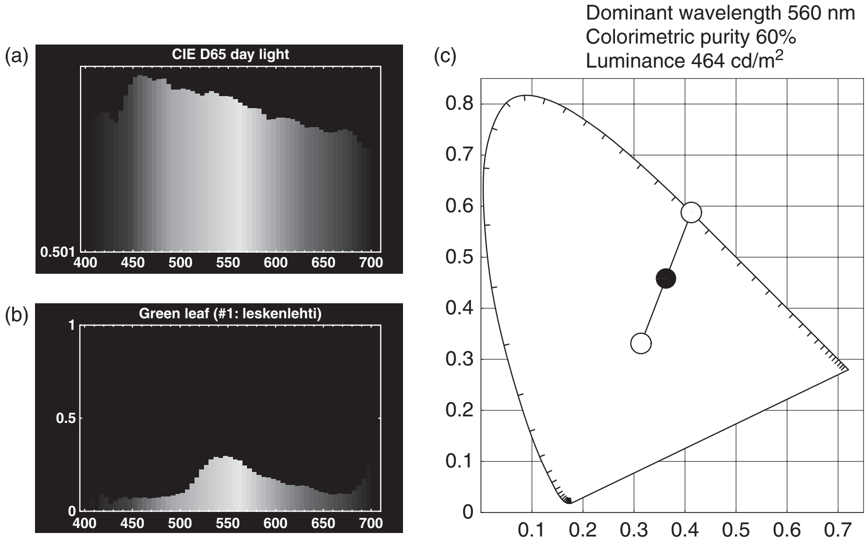


**Figure 1.15** Example of 'Helmholtz coordinates'. The remitted spectrum of a green leaf illuminated with average daylight in the CIE chromaticity diagram. We can obtain the colour as a mixture of the achromatic beam with a monochromatic beam (indicated). The proportions are easily obtained from the chromaticity diagram. (See colour Plate 6 in the centre of this book.)

For a fixed entrance beam it is not hard to calculate the optimum slitwidth. One uses Helmholtz's and Graßmann's observation [46] that any beam admits of a metameric beam that is made up of an achromatic beam and a monochromatic beam. Thus any beam can be characterized through the intensity of its 'achromatic component', the intensity of the 'monochromatic component' and the wavelength of this monochromatic beam.

This particular description is often known as the 'Helmholtz' coordinates of the beam (Fig. 1.15). In our case we are especially interested in the intensity of this (purely hypothetical) monochromatic component, we're looking for the slitwidth at which it reaches a maximum. (It is a priori clear that there will be an optimum for *some* slitwidth since the intensity of the monochromatic beam clearly decreases for very narrow and very broad slits: in the first case the beam becomes black, in the second it becomes white.) The result depends on the spectrum of the entrance beam in so far that we have to designate it the *achromatic* beam. The solution was guessed intuitively by Ostwald [34,35] and later rigorously proved by Schrödinger [42]. It is very simple, the edges of the slit should be located at complementary wavelengths. (Notice that this result indeed depends on the achromatic beam, in this case the entrance beam.) If this is not possible (there exists no complementary wavelength for the left or right edge of the slit), then the slit should be opened such that only one edge is located in the visual region. In that case either a long wavelength or a short wavelength side of the spectrum is passed by the 'slit'.

That this is a reasonable result is perhaps illustrated by the following. Consider a colour at the optimum slitwidth. Let's take a 'green' for instance, then we have the case of two

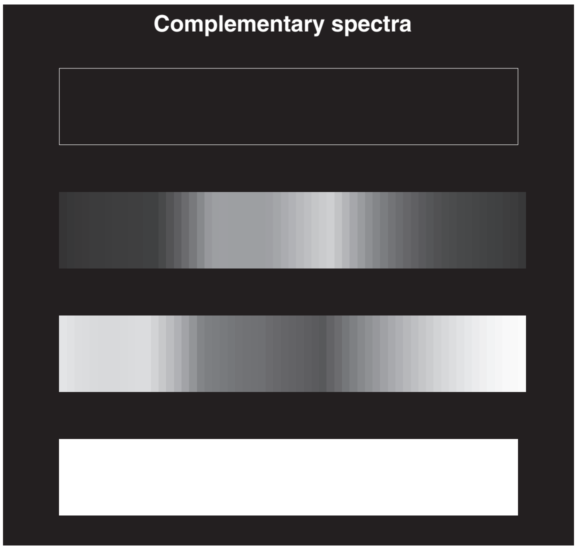


**Figure 1.16** Newton's spectrum and the inverted spectrum at some reasonable slitwidth. Subtractive combination yields black (top), additive combination white (or rather achromatic, bottom). (See colour Plate 7 in the centre of this book.)

complementary slit edges. Suppose we slightly open the slit, taking care to preserve the dominant wavelength of the exit beam. Then we effectively *add* two complementary beams that superimpose to an achromatic beam: we add no colour, we add white! Likewise, consider what happens when we decrease the slitwidth slightly, again preserving the dominant wavelength of the exit beam. In this case we subtract an achromatic beam, or, as Ostwald would have it, 'we add black'. In any case slightly increasing or decreasing the slitwidth doesn't change the amount of colour, thus we must be at an optimum. These colours are known as 'semichromes' or 'full colours' (Ostwald called them *Vollfarben*). They depend on the spectrum, not just the chromaticity of the entrance beam (this is the reason why we introduced achromatic *beams* instead of mere achromatic *chromaticities*). Thus the Newtonian spectrum is indeed most colourful when the slit is very wide, about half of the visual region. This indicates that the monochromatic beams are more like certain *limiting cases* than that they are to be considered the basic building blocks, as Newton would have it.

In the case that one slit edge lies outside the visual region (this happens for the reddish and bluish optimal colours), it is a matter of taste whether one asserts that part of the spectrum is admitted or that part of the spectrum is blocked. We may extrapolate this to the case of the slit proper and consider what happens when we use an obstructing bar instead of a slit (lets call it a 'complementary slit'). The first to try this was Goethe [44,45].[16] He took a prism before the eye and looked at a white stripe on a black card (one sees the Newtonian

[16] We don't want to join the discussion regarding Goethe's and Newton's relative achievements. When Goethe took his first look through a prism at a white wall and didn't see the promised spectral colours he knew then and there that Newton was *wrong*. His acid polemics triggered much unfortunate debate.

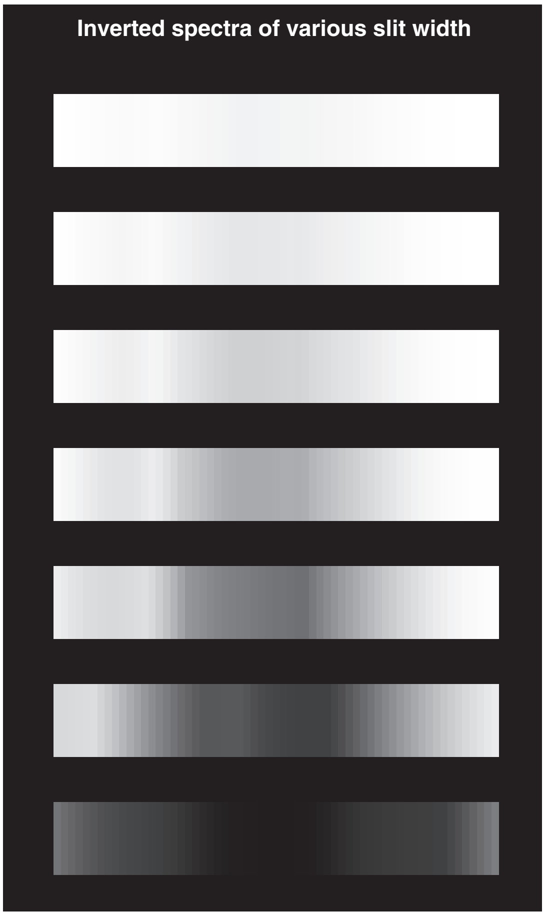


**Figure 1.17** The inverted spectrum at various slit widths. (See colour Plate 8 in the centre of this book.)

spectrum), but also at a dark stripe on a white card. In the latter case one sees the 'inverted spectrum'. In the inverted spectrum all colours are exactly complementary to the colours seen in the Newtonian spectrum (see Fig. 1.16). This is no great surprise since the slits themselves are complementary in the geometrical sense. From geometrical optics it is clear that the exit beams for the Newtonian and the inverted spectrum must always add to the entrance beam (achromatic)—this is technically known as 'Babinet's principle' [4]. Notice that the addition of Goethe's white stripe and dark stripe is simply a uniformly white card.

The inverted spectrum looks as colourful as the Newtonian spectrum (see Fig. 1.17), even more interesting, all experiments that one can do with the Newtonian spectrum one can also do with the inverted spectrum, and with similar results. In particular, Newton's *experimentum crucis*, believed to 'prove' that white light is 'actually' a confused mixture of

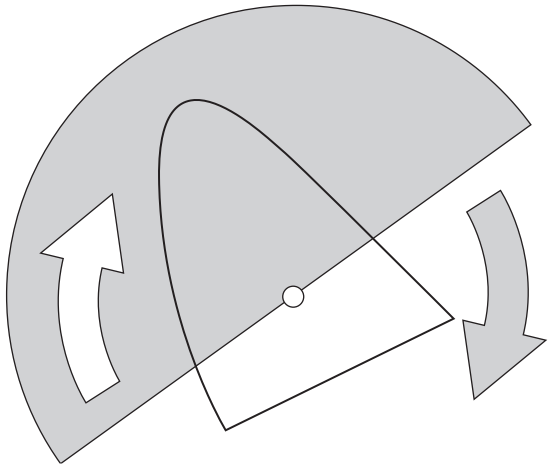

**Figure 1.18** How the closed sequence of full colours is generated from the spectrum (an open linear segment).

homogeneous lights (monochromatic beams), is easily repeated. The conclusion has to be that Newton's experiments were OK, but his conclusions from the experiments overly hasty [18–20]. For colorimetry, the monochromatic beams are nothing special. If it is asserted that colours 'really are' superpositions of monochromatic beams, then—by the same logic—one should be ready to assert that they 'are really' superpositions of beams from the inverted spectrum. This evidently defeats Newton's purpose.[17]

An interesting observation is that whereas the Newtonian spectrum lacks the purples (and thus cannot be closed to obtain the colour circle as Newton erroneously did, freely inventing novel colours for the express purpose), the inverted spectrum contains the purples but lacks the greens. This suggests that *all* colours are somehow represented by a combination of the Newtonian and the inverted spectrum (indeed, we will show later that exactly half of the colours belong to the Newtonian spectrum, whereas the other half stem from the inverted spectrum). The correct way to proceed is to consider the full set of Ostwald's semichromes [34]. The simplest way to obtain an overview of this set is to draw the slit edge positions in a chromaticity diagram. One simply draws a line through the achromatic point and notices the intersections with the spectral locus (see Figs 1.18 and 1.19). There exists either a pair or only a single intersection. By rotating the line over all orientations one obtains a continuous *periodic* series of slits and thus a continuous periodic series of semichromes. This construction proves geometrically that the linear sequence of semichromes is closed, i.e. has the topology of a circle. Here we find the relation between the open-ended linear segment of wavelengths that is the Newtonian spectrum and the colour circle which has always been the intuitive representation of colours by visual artists. One doesn't obtain the

[17] This should not be read as 'Newton bashing'! It would indeed be easy enough to add some negative remarks concerning Goethe's contributions.

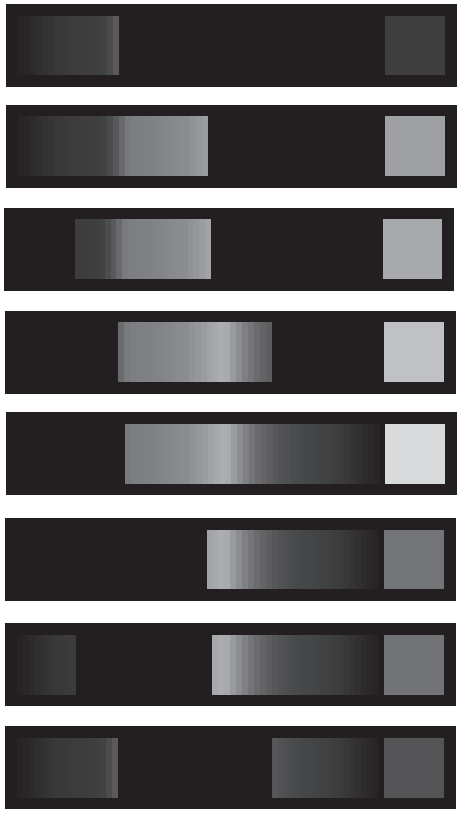


**Figure 1.19** Some representative full colours: spectra (left) and chips (squares on the far right). (See colour Plate 9 in the centre of this book.)

colour circle by simply tying the spectral limits together as Newton did, rather, Ostwald's definition of the *Vollfarben* (semichromes) [34] provides a natural map of all colours on a closed manifold. Notice that this construction depends critically upon the introduction of an achromatic beam. It cannot be done in the austere colorimetry without such an arbitrary fiducial beam. We consider these insights to be a major contribution by Ostwald to colorimetry.

When we circumnavigate the set of semichromes we encounter four different kinds of these, namely (moving in a steady progression in the direction from blue through green to red):

1. The 'short wavelength boundary colours'. Here one edge of the slit is beyond the blue spectral limit, the other moves from the complementary of the red limit to the complementary of the blue limit. These patches tend to appear bluish.
2. The 'band pass colours'. Here both edges of the slit are in the visual region. One edge of the slit moves from the blue end of the spectrum to the complementary of the red

limit, the other edge moves from the complementary of the blue limit to the red limit. These patches tend to appear greenish.

3. The ‘long wavelength boundary colours’. Here one edge of the slit is beyond the red limit of the visual region. The other edge moves from the complementary of the red limit to the complementary of the blue limit. These patches tend to appear reddish.
4. The ‘band stop colours’. Here both edges of a *complementary* slit are in the visual region. One edge moves from the complementary of the blue limit to the red limit, whereas the other edge runs from the blue limit to the complementary of the red limit. Notice that here we meet the short wavelength boundary colours again: the set of optimal colours closes. These patches tend to appear purplish (or ‘magenta’).

Notice that the band pass colours are part of the Newtonian spectrum whereas the band stop colours are part of the inverted spectrum. We have indeed combined these complementary entities!

The ‘boundary colours’ (*Kantenfarben*, Fig. 1.20) were first studied by Goethe [44,45]. One sees them when looking at a light–dark boundary through a prism. By changing the orientation of the prism or the light–dark edge, one switches from the short wavelength boundary colours to the long wavelength boundary colours or vice versa. Only parts of the sets of boundary colours appear as optimal colours (as only parts of the Newtonian spectrum and the inverted spectrum appear as optimal colours). The full progression of short wavelength boundary colours runs from dark blue over cyan to bright white; that of the long wavelength boundary colours runs from dark red over yellow to bright white. As Goethe noticed, the boundary colours contain neither greens nor purples.

If one is inclined to do so, the boundary colours can be mixed to obtain monochromatic beams.[18] Likewise one can mix monochromatic beams to obtain boundary colours.[19] This shows that it is really immaterial whether one bases spectral descriptions on Newton’s ‘homogeneous lights’ [31,32] or on Goethe’s *Kantenfarben* [44,45]. The boundary colours have some advantages, e.g. they can be produced easily and exactly, whereas monochromatic beams can only be produced problematically and approximately (the ideal ones—for zero slit width—obviously cannot be produced at all and in that sense cannot even be said to exist). The whole discussion on which is more fundamental is really of little or no fundamental importance.

We can plot the loci of boundary colours and semichromes in the chromaticity plane or in colour space itself (Figs 1.20 and 1.21). The boundary colours describe spirals from the origin to the colour of the entrance beam [11,27,33]. The semichromes describe a closed, twisted space curve encircling the achromatic axis. We will discuss the fundamental relevance of these curves later. For the moment it is most useful to plot the curves in the chromaticity plane, thus abstracting from the brightness of the entrance beam. The resulting configuration depends on the spectral composition of the entrance beam, i.e. the

[18] One simply looks at a narrow white bar on black paper.
[19] One simply combines parts of the spectrum with Maxwell’s [28,29] or Ostwald’s [34] apparatus.

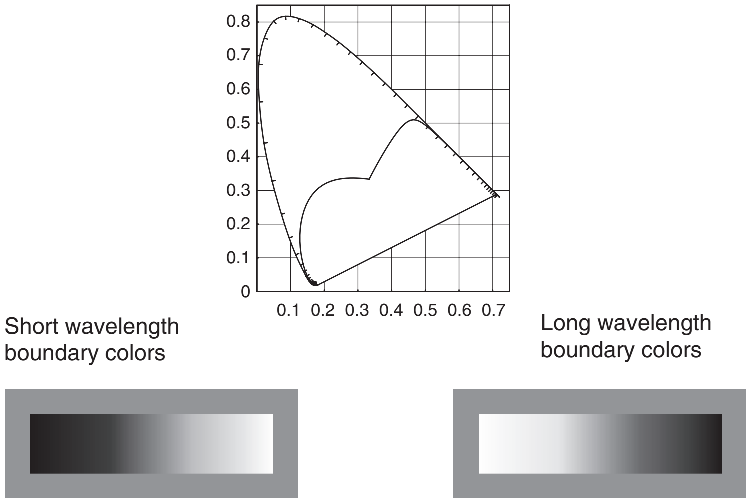


**Figure 1.20** The boundary colours in the CIE chromaticity diagram and impressions of sequence of hues of the short wavelength and long wavelength boundary colours. (See colour Plate 10 in the centre of this book.)

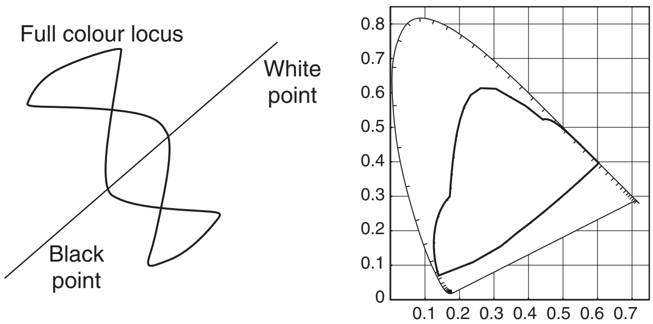


**Figure 1.21** The full colour locus in $\mathbb{C}$ (left) and in the CIE chromaticity diagram.

achromatic beam. We obtain a configuration of three mutually concentric entities:

1. The achromatic point. This is the chromaticity of the spectral colours with wide open slit. They all look white.
2. The closed curve of semichromes. These are the chromaticities of the spectrum and inverted spectrum for optimum slitwidth, i.e. these are the most colourful points in the chromaticity plane.
3. The closed curve made up of the spectral locus and a segment of the line of purples. This is the boundary of the real (as opposed to virtual) chromaticities. These points represent the spectral colours at vanishing slitwidth, i.e. black or total darkness.

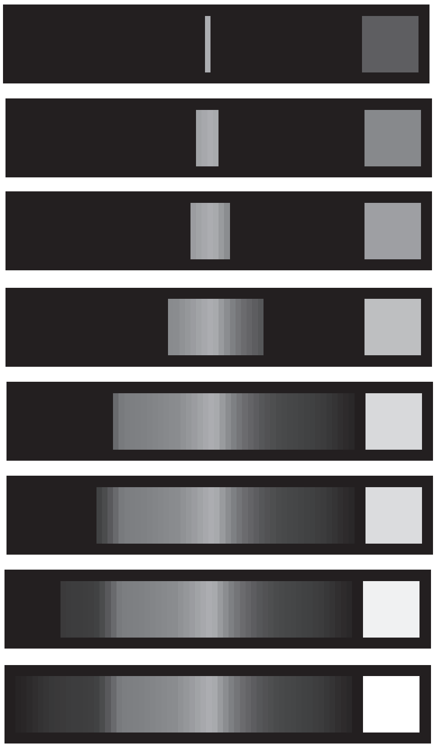


**Figure 1.22** Spectra and samples (chips) of a 'yellow' paint (see colour Plate 11 in the centre of this book). The difference is the width of the spectrum remitted by the paints (they are all optimal colours). When this range is very narrow, the paint appears dark brown. When it is very large (the whole visual region), the paint looks white. The 'best yellow paint' remits all wavelengths above 490 nm. Notice that this paint is a long wavelength boundary colour.

Notice that this picture is quite different from the conventional one where the most colourful points are put at the boundary of the region of real chromaticities. In such a conventional representation it is silently assumed that when you close the slit you simultaneously increase the intensity of the entrance beam. The points at the boundary then represent exit beams at zero slitwidth and infinite radiant spectral power density. Such points are only approximately reached under special laboratory conditions.

In real life, colourful colours are always nearer to the semichromes and the achromatic point, not to the boundary of the region of real colours. This fact was discovered empirically by Ostwald [35], who simply put highly coloured pigments before a spectroscope (Fig. 1.22). He noticed that the best pigments remit about a semichrome. For instance, the best yellow pigments remit all wavelengths over 490 nm and absorb those below it. Since yellow occurs in the spectrum at a wavelength of about 575 nm, it shocked many physicists [35] at the time, who believed that the strongest colours should necessarily be like Newton's 'homogeneous lights', thus a yellow paint was expected to reflect only a narrow wavelength region at about

575 nm. Such a paint would be *black* though, because it would hardly remit any radiation from such a narrow wavelength region. These relations are illustrated in Fig. 1.22.

## The space of object colours

In the case of object colours we meet a situation much like that considered in the previous section. In a rather simplified setting we have an *illuminant* and *surfaces* which remit part of the beam of the illuminant. We assume that only reflection and scattering occur (no fluorescence). Then the remitted beams are essentially equal to the beam of the illuminant with the radiant power at certain wavelength regions selectively attenuated (Goethe's notion of colours as 'shadow-like' entities [44,45]). In order to make this more precise we will define a new entity that should figure in colorimetry next to the concepts of 'beam' and 'patch'. We will call it a 'chip', after the chips to be found in conventional colour atlases. Thus chips unlike beams (not to speak of patches!) are *material* things that you may indeed grasp with your hand. However, we are only concerned with some *optical* properties of these objects.

First we define the concept of a 'black chip'. This is very simple, a black chip is a surface that does not remit *any* of the impinging radiation (at least within the visual region). For if it were otherwise, we could certainly produce an even 'deeper' black. Notice that this is a physical definition that allows one to construct approximations to black chips (e.g. black satin velvet, soot, a hole in a box, etc.) and that the definition is not dependent on colorimetric notions, nor need one store a black chip as an international standard.

Next we define the concept of a white chip. This is rather more intricate. First of all, a white chip should remit *all* impinging radiation. For otherwise I could certainly construct a chip looking brighter than the white chip under the same illuminant, which defeats the very notion of a white chip. But this is not sufficient. For instance, a perfect mirror also remits all impinging radiation, yet no one would be prepared to call a mirror 'white'. The reason is that a perfect mirror, illuminated with a collimated beam will actually look *black* from almost all vantage points, except from the specular direction from which it looks dazzlingly bright (e.g. much brighter than writing paper or white chalk, the whitest materials we can imagine). Some thought reveals that white objects should look the same (namely white) from all vantage points. This actually suffices to define the white chip: it must be a surface that remits all radiation ('unit albedo') and scatters the radiation equally in all directions (a 'Lambertian surface'). Again, this definition does not draw on colorimetric notions, nor will it be necessary to store an international standard. The definition is a recipe to construct white chips. Good approximations are flour, white chalk, matte writing paper or the classical magnesium oxide smoked upon glass from the photometric laboratory. This definition of a white surface is essentially due to Ostwald [34,35] and, again, is to be considered a major contribution to colorimetry.

Once one has defined the white chip, it is easy enough to define general (coloured) chips. These are exactly like the white chip except for a wavelength-dependent attenuation. Thus the chips can be characterized by a spectral remittance factor, which is simply the ratio of the spectral radiant power density remitted by the chip to that remitted by the white chip. The beams remitted to the observers can then be taken as the spectral remittance factor times the white beam, that is the beam remitted from a white chip. Thus given the illuminant (i.e. the beam remitted from a white chip) we can uniquely specify the beams remitted by given chips. Then a *chip* plus an *illuminant* define a *beam* which causes a *patch* to be seen which can be ascribed a *colour* (through the colorimetric paradigm), in this case an 'object colour' (but see below). In this setting the choice of the achromatic beam is obvious: of course, we simply take the white beam. Thus we have completed the canonical setting for the colorimetry of object colours. Notice that whereas an aperture colour depends on the beam only (we take the eye as a constant factor here), the 'surface colours' depend on the illuminant and a chip. One way to handle this formally is to specify an 'object colour' as a *pair* of aperture colours, namely the pair made up of a colour due to the chip and one due to the white chip. Operationally, this would correspond to the chip being presented on a white background, supplying the (really minimal) 'context' necessary for a colour to be a 'surface colour' to start with.

**Change of illuminant**

When one changes the illuminant the remitted beams associated with a given chip change. The achromatic beam changes, too. As a result, the colours of the chips change. The transformation is a complicated one, it is not simply that we obtain a deformation (reshuffling or isomorphism) of colour space. Rather, it may well happen that two chips looking alike under one illuminant look unlike under another illuminant and vice versa. This clashes with the naive notion that chips possess a 'real' or 'intrinsic' colour. It is not quite clear what one could mean by the 'real' colour of a chip. Apparently it should be taken to mean something like the colour of the chip under a standard illuminant (which is a novel entity that enters the picture here). But if this is to be the definition, then the real colour of the chip will *never* be perceived, except in very special laboratory conditions! This can hardly be considered satisfactory.

As a physicist one is perhaps induced to take the *spectral remittance factor* as the 'real' colour of a chip. This definition has at least the advantage that it does not depend on any (arbitrary) 'standard illuminant'. It is a property that is characteristic of the material of the chip, quite independent of its arbitrary irradiation in some photometric setting. A problem is, of course, that the spectral remittance factor as such is never seen. What one may perceive are the colours under various illuminants, like variations on a theme. The invariant theme, then, would be the spectral remittance factor or the 'true' colour. From the vantage point of the physicist this is not unreasonable, for it exactly describes the method of spectroscopy. In spectroscopy one irradiates the sample with monochromatic beams and measures the remittance: this is a direct determination of the spectral remittance factor. A problem is that this requires that one knows the illuminant. This can be solved by measuring the sample against a white background and indeed, this is common spectroscopic practice (it is the preferred method because random fluctuations of the source then tend to cancel out automatically). Our formal device of representing object colours as pairs of aperture colours (due to the chip and the white chip) reflects this.

One way to describe this in colorimetric terms is to factor the multiplicative relation:

$$\text{colour} = \text{colour-matching functions} \otimes \text{spectral remittance factor} \otimes \text{illuminant spectrum}$$

in a different way. One contracts the colour-matching functions with the illuminant spectrum, to obtain 'the colour-matching functions for the given illuminant' and lets them work on the spectral remittance factor. Then the illuminant 'changes the eye', whereas the chips are described in a way that is independent of the illuminant (by their 'true colour'). Thus

$$\text{colour} = \text{eye} \otimes \text{spectral reflectance},$$

where *eye* should be interpreted as

$$\text{eye} = \text{colour-matching functions} \otimes \text{illuminant spectrum}.$$

This is often an advantageous way to frame the problem and perhaps most closely captures our intuitive notions. It is fully equivalent to the other formulation of the same problem: one contracts the spectral remittance function with the illuminant spectrum (the 'remitted beam') and lets the regular colour-matching functions work on it. In this view one looks at the 'apparent colour' of the chip, which is the spectrum under the given illuminant, not a material property of the chip. The two views are, of course, formally equivalent.

### The colour solid

In the history of colour science one encounters many different types of 'colour solid'. One finds colour pyramids (Lambert [24]), spheres (Runge [38]), double cones (Ostwald [34]), shapeless colour trees (Munsell [10,30]), colour cubes (Hicketier [23]), etc. Perhaps the most immediately remarkable aspect is the fact that these colour solids are *finite* volumes (except perhaps for the Munsell tree, which is, at least theoretically, permitted to grow without bounds, although in actual fact we are only given a finite set of samples), whereas the colour cone is an *infinite* volume. The reason is that the colour solids put a simultaneous order on surface colours (chips), whereas the colour cone puts an order on beams. Now the colour of beams may vary from darkness to dazzling bright (infinite distance from the origin), whereas surface colours may only vary between black and white. All chips remit less than the white chip, thus they should map on a finite region of **C**. Let us consider the shape of this region (might it be a tree, pyramid, sphere, etc.?).

First, consider the representation of the beams remitted by the chips in S. In order to simplify the discussion we will start with an illuminant that has a flat spectrum (same spectral radiant power density at all wavelengths). Then the beams remitted by the chips have spectral radiant power densities that vary between zero and the density remitted by the white chip (which doesn't depend on wavelength). Geometrically this means that the region filled by the beams remitted from all conceivable chips is a *hypercube* (same edge length in every dimension). The colour solid then is simply the projection of this hypercube in three-dimensional colour space. One can show that such a projection appears as a fusiform (Zeppelin-shaped) body in three-space (Fig. 1.23). When the illuminant doesn't have a flat spectrum, little changes. Instead of a hypercube we get a *hyperbox* and the fusiform body deforms a bit but doesn't change qualitatively.

Another way to find the colour solid in **C** was developed by Schrödinger [42]. He first solves the problem of 'which paints are the brightest' for a given illuminant and a given

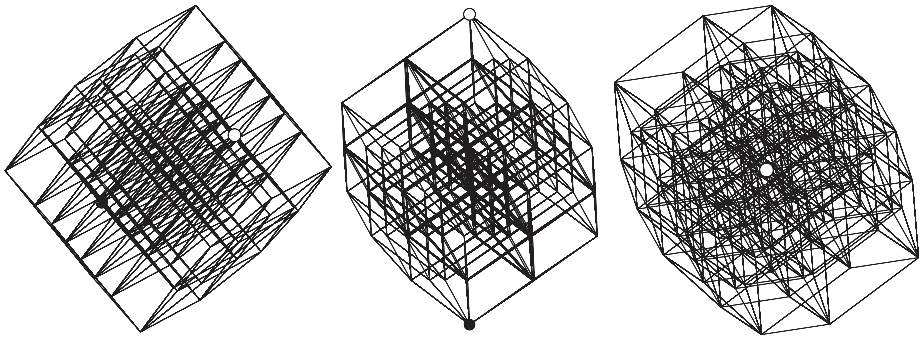

**Figure 1.23** Some projections of the seven-dimensional hypercube. In the middle view the essential fusiform shape of the colour solid is easily visible.

chromaticity. He proves in a very general way that such paints have an 'ideal' spectral remittance factor, namely, their spectral remittance factor is either zero or unity.[20] Moreover there are no more than two transitions in the visual region:[21] such chips are called 'optimal'. (Clearly, the boundary colours and full colours considered earlier are examples of such optimal colours.) Then it is clear that the optimal colours must lie on the boundary of the colour solid, for otherwise one could find a still brighter paint at a given chromaticity. Thus the optimal colours form an explicit parameterization of the boundary of the colour solid. As parameters one may use the transition wavelengths, thus the optimal colours form a two-parameter family, i.e. generically a surface. When we plot such a surface we obtain a fusiform body, indeed the projection of the hyperbox defined by the illuminant in **S** (Figs. 1.24–1.27).

Notice that the boundary colours are only one-parameter families, thus generically *curves*. Indeed, the boundary colour loci are spirals in **C** that lie on the surface of the colour solid [11,27,33]. They divide the boundary of the colour solid into two (congruent) parts, one part containing the band pass optimal colours (Newtonian 'impure' spectral colours), the other part the colours of the impure inverted spectrum. Thus exactly half of the optimal colours are Newtonian spectral colours, the other half are inverted spectral colours, whereas the boundary colours form a set of vanishing measure. Again, this is a striking demonstration that the Newtonian 'homogeneous lights' are nothing special.

The colour solid inherits central symmetry from the hyperbox of which it is the projection. The symmetry centre is the median grey (remittance factor 50% for all wavelengths) chip. We define the 'grey axis' as the line connecting the white and black points. The grey axis is obviously a segment of the achromatic axis. It is a natural question to ask for those optimal

[20] The argument is simple and general. Suppose there existed a region of intermediary reflectance. Then one might perturb the reflectance at three places and thus obtain a colour change, e.g. one might perturb such as to add more of the fiducial colour! This clearly should not work, thus the reflectance should be either zero or one.

[21] The argument is again simple and general. Suppose there existed three or more transitions. Then one could slightly perturb the locations of three of them and thus mix yet more of the fiducial colour! This should be impossible, thus there cannot be more than two transitions.

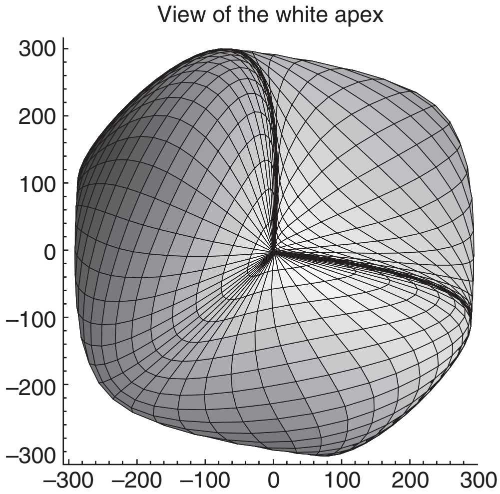


**Figure 1.24** The colour solid in the canonical (SVD) basis. Here is a view of the white pole (see colour Plate 12 in the centre of this book).

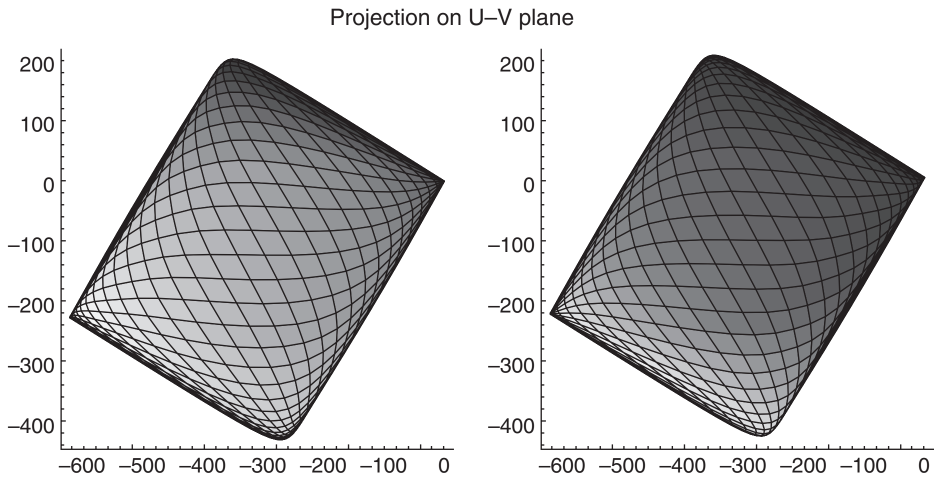


**Figure 1.25** The colour solid in the canonical (SVD) basis. Here is a view in the direction of the third dimension (see colour Plate 13 in the centre of this book).

colours that are as far removed from the grey axis as possible, for these will be the most 'colourful' optimal colours. These colours turn out to be exactly Ostwald's semichromes, or full colours (*Vollfarben*) [34] i.e. optimal colours with complementary transition wavelengths. The locus of semichromes runs as an 'equator' (except that it is not a planar curve) over the boundary of the colour solid, encircling the grey axis. Like the colour solid itself, it has central symmetry about the median grey chip.

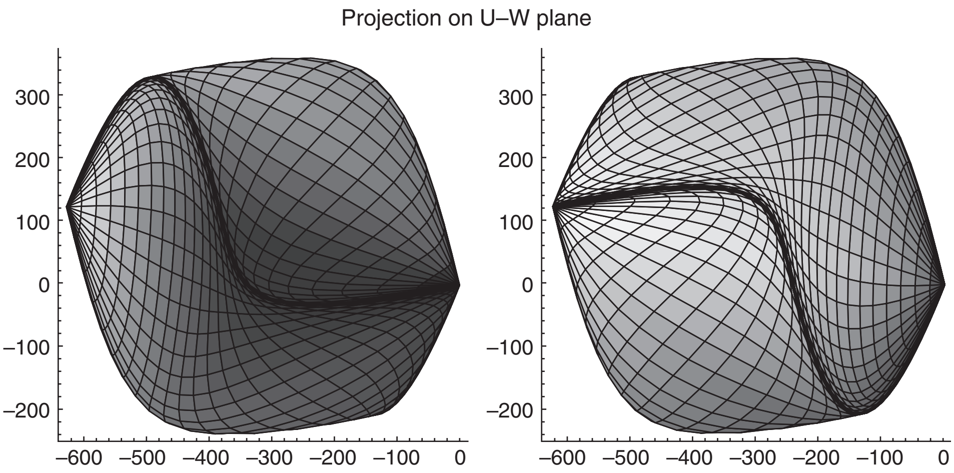


**Figure 1.26** The colour solid in the canonical (SVD) basis. Here is a view in the direction of the second dimension (see colour Plate 14 in the centre of this book).

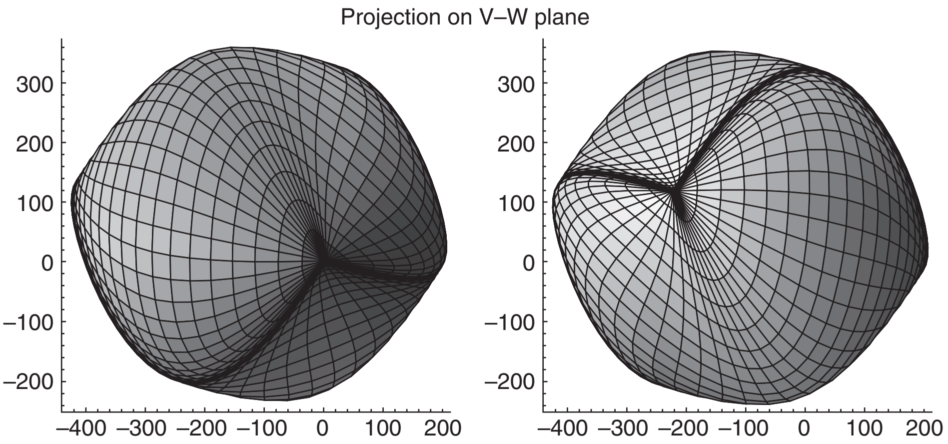


**Figure 1.27** The colour solid in the canonical (SVD) basis. Here is a view in the direction of the first dimension (see colour Plate 15 in the centre of this book).

For very dark optimal colours, the optimal colours approach the monochromatic beams (very 'narrow slits', thus hardly any radiation remitted). This means that the shape of the fusiform body near the black point is approximately that of the spectral cone boundary. In the limit, the spectral cone boundary is *tangent* to the conical singularity of the fusiform body at the black point. If we take a myopic view and see only the neighbourhood of the black point, then the colour solid looks in no way different from the space of aperture colours! By central symmetry we see that at the white point the tangent cone is the *inverted spectral* *cone*. This can be seen as the spectral cone reflected on the median grey, but—equivalently—also as the cone due to the inverted spectrum. Indeed, near the white point the optimal colours are like colours from the inverted spectrum with very narrow (complementary) slitwidth. (The myopic observer at the white point sees an 'upside down' and *complementary* version of the space of aperture colours.)

Thus we see that the colour solid is a centrally symmetric body (like the cube, the sphere, or the double cone) and that it is everywhere smooth, except for conical singularities at the white and black points (unlike the cube which has edges, the sphere which has no conical singularities, or the double cone which has a crease at the equator). It is totally unlike the Munsell tree, which is a (theoretically) unbounded structure. The idea of the Munsell tree is that you can extend it *ad libitum* when novel (ever more colourful) pigments become available. The point (lost to Munsell) is that this will definitely never happen: the colour solid is bounded because the optimal colours are the brightest possible and, indeed, lie on the natural boundary.

This answers some of our initial questions. Although some of the colour solids suggested in the literature have properties not unlike the true colour solid, all deviate in arbitrary ways, and some configurations are clearly very unfortunate or even mistaken. We should regard all properties not present in the true colour solid as irrelevant fancies and distortions of reality.

### Spectral dominance

The surface colours are due to spectrally selective *attenuation* of the white beam. This suggests that it may be of interest to put a partial order on the space of all beams (S) in the following way. Beam $\mathcal{A}$ will be said to 'spectrally dominate' beam $\mathcal{B}$ when it is the case that, for all wavelengths throughout the visual region, the spectral radiant power of beam $\mathcal{A}$ is not less than the spectral radiant power of beam $\mathcal{B}$. This makes the set of beams a poset (partially ordered set). In the case of the surface colours the white beam spectrally dominates all remitted beams. Optimal colour $\mathcal{A}$ spectrally dominates optimal colour $\mathcal{B}$ if the pass band of $\mathcal{B}$ is fully contained within the pass band of $\mathcal{A}$. Thus the subset of optimal colours forms a poset under inclusion of pass bands. The white colour is the unique LUB (lowest upper bound) and the black colour the unique GLB (greatest lower bound) of this subset. Thus the optimal colours satisfy a lattice order, or a true hierarchy.

We may continue in this spirit and define several additional useful relations.

Two optimal colours will be called 'categorically different' when their pass bands are disjunct. When two optimal colours are not categorically different we can always find illuminants under which they appear equal (though not black). When two optimal colours are categorically different one can find no such illuminant, the colours will either look different, or both will appear black.

For any two beams we can define a novel and very strong notion of complementarity. We call two beams 'complementary' if their superposition equals the white beam. Here 'equality' is not colorimetric equality, but actual equality of the spectra. It is evident that such complementary beams will also have complementary dominant wavelengths. However, beams with complementary dominant wavelengths are unlikely to be complementary in this novel, strong sense. The notion of complementarity is important because it is immediately related to the central symmetry of the colour solid. For any beam I can construct the complementary beam, which is again a possible surface colour, and which has a colour that is symmetrically located with respect to the median grey point for the simple reason that the colours should always add to the white point. Notice that when $\mathcal{A}$ spectrally dominates $\mathcal{B}$, then the complementary beam $\bar{\mathcal{A}}$ of $\mathcal{A}$ will be spectrally dominated by $\bar{\mathcal{B}}$.

A better term would perhaps be 'supplementary colours' (German: *Ergänzungsfarben*), but 'complementary' is regrettably the common term.

## Colour atlases

The idea of a colour atlas is both simple and attractive [1]. One produces fiducial chips and 'measures' a given sample by comparison with the fiducial chips under some standard illuminant. Of course there are numerous problems with this concept, both of a practical and of a theoretical nature [22]. Here we only consider some of the conceptual issues.

We start by noticing that there are two quite incompatible approaches to be found in the literature. In one approach one simply forgets about colorimetry and attempts to arrive at a perceptually 'evenly spaced' and 'naturally ordered' system by purely psychological means. (Of course, one may measure the result colorimetrically and attempt to build a bridge to colorimetry in retrospect, but that doesn't interest us at this point.) The Munsell system [10,30] is perhaps the best known example. This is a valid approach in itself though it has nothing to do with colorimetry. Consequently it doesn't concern us here. Another approach is to attempt to construct a colour order system on colorimetric principles. (Of course, one may perform psychological measurements and attempt to build a bridge to perceptually even spacings in retrospect, but that doesn't interest us at this point.) The best example is perhaps Ostwald's atlas [34]. There exist many attempts of a mixed nature (e.g. the CIE's Lab-system, the DIN system) but such mongrel attempts are not very relevant from a principled point of view. Since the Ostwald system is by far the most rational attempt, we consider it here in some detail. Historically it is a unique attempt at a rational systematization of object colours. Unfortunately Ostwald committed some (relatively minor) errors in the process. We don't consider that these glitches detract from the fundamental importance of the attempt though. In any case, no one has done better since. The fundamental advances proposed by Ostwald have largely been lost on the colorimetric community though. Although one finds it mentioned (a rare enough occasion!) we know of no author past the 1950s (say) who recognizes the *essential* conceptual differences with, for example, the Munsell system.

Ostwald bases his atlas on the semichromes. He considers any object colour as made up of fractions of optimal colour, white and black, the fractions (colour content, white content and black content) adding up to 100% (Figs 1.28 and 1.29). This immediately leads to his double cone representation. The oversight is that there exist colours (for instance, the optimal colours other than white, black or the full colours) that cannot be incorporated in this scheme and will fall outside the double cone. The double cone only represents *part* of the colour solid, namely the convex hull of the full colours, the white and black point. However, this oversight is relatively minor and the problem can easily be mended in Ostwald's own spirit (see below).

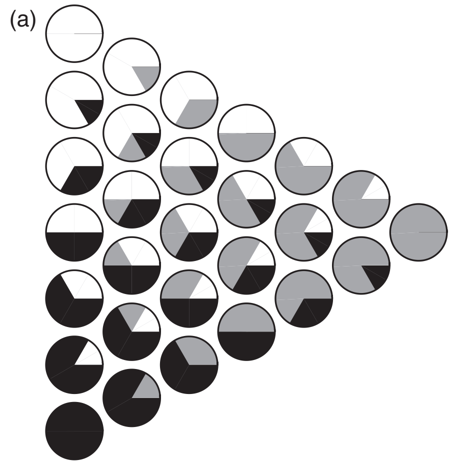


**Figure 1.28** (a) Principle of an Ostwald page. The page consists of partitive ternary mixtures of white, black and an optimal colour. (b) The same Ostwald page as in (a), but with the white, black and colour sectors mixed (one may think of a set of Maxwell tops being spun). (See colour Plate 16 in the centre of this book.)

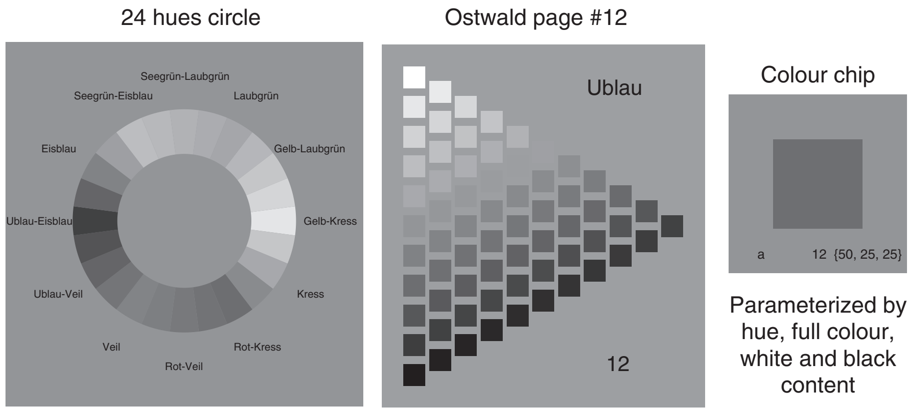


**Figure 1.29** The basic structure of the Ostwald atlas: colour circle (mensurated set of *Vollfarben*), single page, single chip. (See colour Plate 17 in the centre of this book.)

The double cone is a rather arbitrary topological deformation of the colour solid as we have discussed it. Although complementarity and linear structure in planar sections through the grey axis are conserved, the general linear structure of $\mathbb{C}$ is destroyed. This is a severe (and unnecessary) disadvantage.

Notice that with the introduction of the colour, white and black content, Ostwald has solved the simultaneous order in the planes of constant dominant wavelength. In order to complete the ordering, he has to put a rational order on the semichromes themselves. As Ostwald remarks, the semichromes are like 'beads on a string', in the sense that one

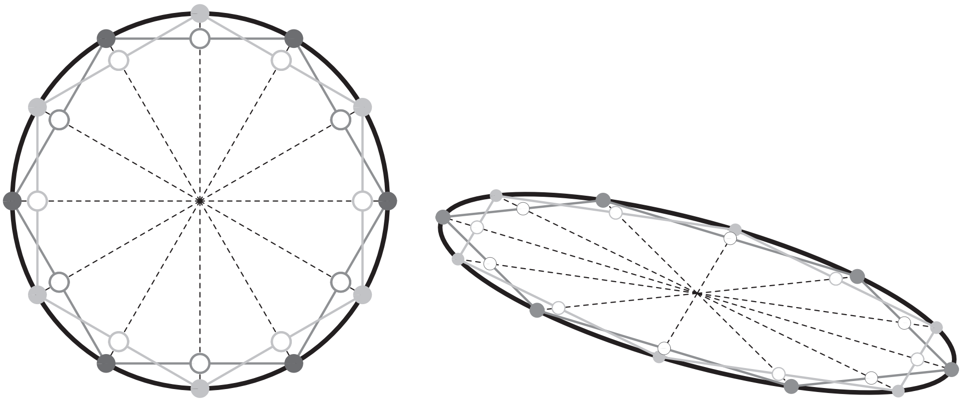

**Figure 1.30** Ostwald's principle of internal symmetry in action. Here we have mensurated a circle and deformed it into an ellipse. The ellipse is automatically mensurated because Ostwald's principle is affinely invariant. (See colour Plate 18 in the centre of this book.)

may push them about as one pleases. Some principled method has to be found to fix their positions.

The solution offered by Ostwald is the principle of internal symmetry (Fig. 1.30). What he means is the following. Suppose we want to subdivide the locus of optimal colours in $M$ 'cardinal colours' (our term). Let's take $M = 24$ for concreteness, as Ostwald did in his second attempt (the exact number is not important). Then the central symmetry means that cardinal colours $i$ and $i+12$ must be complementary. Thus one need only order the first 12 cardinal colours. The principle of internal symmetry states that cardinal colour $i$ should have the same dominant wavelength as the equal mixture of cardinal colours $i-1$ and $i+1$. It is not clear how Ostwald conceived of the idea, but it is certainly an algorithm that leads to a unique mensuration of the semichromes through purely colorimetric calculations. The idea is contained in a letter by Graßmann [16] and is used, as a matter of fact, by some painters [37]. It must be said that Ostwald was somewhat confused on the issue and added irrelevant and even inconsistent axioms to the principle of inner symmetry. This has been cleared up admirably by Bouma [5].

In retrospect, the principle of inner symmetry is quite sophisticated. In the limit for large $M$ it defines a kind of parameterization of the full colour locus by arc length. Were the semichrome locus a flat curve, then it would indeed be a parameterization by *affine arc length* and would lead to an affinely invariant mensuration of the semichromes, quite independent of the choice of primaries. As the case is, the semichrome locus is a general (twisted) space curve and the situation is much more complicated. Again, though defective in practice, this idea of Ostwald's contains the nucleus of a great idea, namely the invariant mensuration of the semichromes by purely colorimetric means. Ostwald has been severely criticized for his errors, but the fact is that all later attempts are feeble or arbitrary by comparison. Instead of critique, it might be more constructive to attempt a rationalization

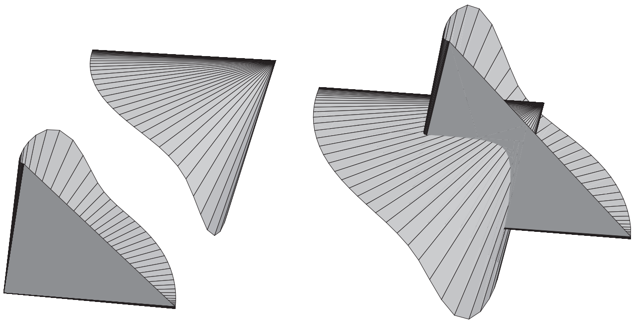

**Figure 1.31** How the spectral cone and its complementary image (inverted spectral cone at the white point) define a double conical volume in **C**. (See colour Plate 19 in the centre of this book.)

of the simultaneous order of object colours by colorimetric means in the spirit of Ostwald, but avoiding his less fortunate decisions. We offer exactly such an attempt here.

Perhaps the first unfortunate decision of Ostwald was to base his atlas upon the semichromes. For one reason, the colours of the semichromes depend on the spectrum, not just the chromaticity of the illuminant. Moreover, many colours (i.e. the optimal colours) appear as 'supersaturated'. It appears to be more natural to let the semichromes be less than maximally saturated since one may certainly spot some whiteness in them. For instance, especially in the red, people have often remarked that some of the darker chips of Ostwald's atlas look more strongly coloured than the *Vollfarbe*. (Bouma [6] has a good discussion.) There exists a simple way to meet these problems (Figs 1.31 and 1.32). Instead of basing the atlas on the semichromes, we base it on a certain family of virtual colours that we will designate 'characteristic colours'. We define them in the following way. The colour solid is a smooth body except for the conical points at the white and black points. If we construct these tangent cones (spectral cone and inverted spectral cone), we find that they intersect in a curve that, like the semichrome locus, encircles the grey axis. We define this curve as the locus of the 'characteristic colours'. The characteristic colours depend on the chromaticity of the illuminating beam (the position of the white point in $\mathbb{C}$), but not on its spectrum (thus unlike the semichromes). If one varies the spectrum of the illuminating beam, leaving the white point invariant, the colour solid (and thus the semichromes) changes, but the double cone (not to be confused with Ostwald's 'double cone'!) with the characteristic colours remains invariant. Moreover, the double cone is the *envelope* of all colour solids obtained in this way. Thus we may unambiguously describe any colour in Ostwald's tradition as characteristic colour content plus white content plus black content. In this scheme the semichromes have less than 100% colour content and finite white and black content. No supersaturated colours occur. This essentially solves the major problem found in Ostwald's scheme.

The other problem involves the mensuration of the characteristic colour locus. We cannot simply use Ostwald's principle of inner symmetry, since the locus is not a planar curve. Instead, we use a similar principle that is manifestly affinely invariant. We use the fact that

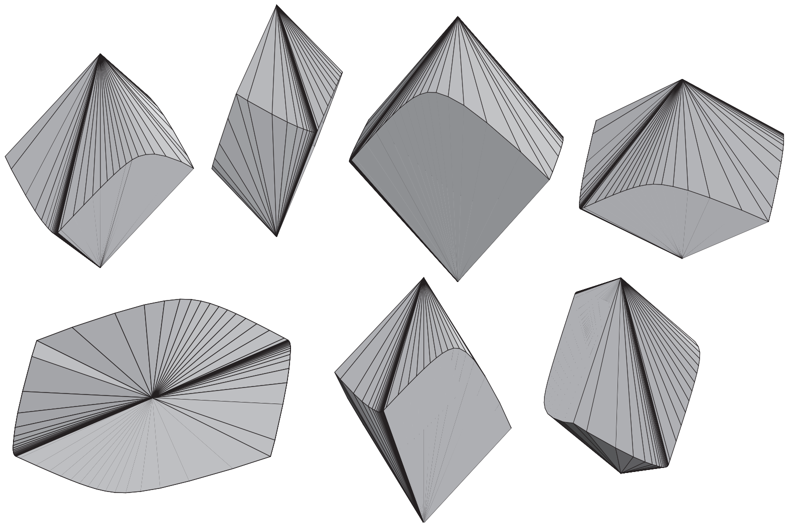

Figure 1.32 Various views of the intersection of the spectral cone at the black point and the inverted spectral cone at the white point. The sharp edge (equator) is the locus of characteristic colours. Notice that the overall shape is strongly determined by the (flat) purple sector and its inverted copy. (See colour Plate 20 in the centre of this book.)

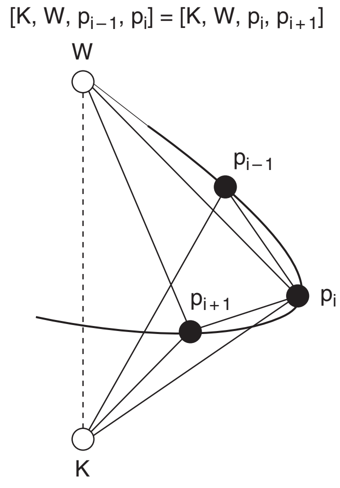


Figure 1.33 The geometry of the mensuration by volume ratios.

we have a special segment on the origin, namely the grey axis. If we consider three nearby characteristic colours in sequence, $\mathbf{p}_1, \mathbf{p}_2$ and $\mathbf{p}_3$, say, we may find the volumes (Fig. 1.33) of the tetrahedra $\mathbf{kp}_1\mathbf{p}_2\mathbf{w}$ and $\mathbf{kp}_2\mathbf{p}_3\mathbf{w}$ (here $\mathbf{k}$ denotes the black and $\mathbf{w}$ the white point). We notice that the ratio of these volumes is an affine invariant, in particular it will be invariant against changes of the primaries. This ratio can serve to define an affine invariant arc length of the locus of characteristic colours. When we simply cumulate incremental volumes and divide by the total volume we have constructed an affinely invariant parameter that increases by unity when one circumnavigates the locus of characteristic colours once. This completes the colorimetric mensuration of the colour solid in the Ostwaldian spirit. It yields a fully rational simultaneous order of object colours.

When we analyze the volumetric mensuration analytically we find that, perhaps surprisingly, for small increments (formally 'infinitesimal', in practice a 24-hue division will do) the equal mixture of hues $\mathbf{p_i}$ and $\mathbf{p_{i+2}}$ lies in the plane $\mathbf{kwp_{i+1}}$. Thus $\mathbf{p_{i+1}}$ has the same dominant wavelength as the equal mixture of $\mathbf{p_i}$ and $\mathbf{p_{i+2}}$, which, again, is exactly a reformulation of Ostwald's principle of internal symmetry: we have found the correct generalization of Ostwald's principle. When formulated in this way, the principle is a flawless affine invariant.

Such a principled scheme is interesting because it yields a *global* order on the surface colours. A major problem with the psychological schemes based on colour difference judgments is that their order is only *locally* well defined, but globally not well determined. This is the case because it is very hard to judge equal differences between very different chips. Thus such systems are locally even but globally uneven. Of course, experiment has to decide whether the principled order is anything close to a perceptually even one, for colorimetry proper has nothing to say on this issue.

In Fig. 1.34 we present a comparison of the results of a mensuration of locus of characteristic colours via the (new) principle of inner symmetry with the (perceptually uniform) Munsell scale. Deviations are of the same order as deviations between the various perceptually uniform scales (one to several steps on a 24-hue scale). We conclude that 'perceptually uniform' essentially coincides with metrical uniformity. From a pragmatic point of view, colour atlases are probably best constructed from first principles, rather than via laborous and noisy perceptual judgements.

## The colour cube: RGB-display colour space

The colours on CRT screens are produced by way of additive mixture of three beams that can be attenuated in programmable proportions. A typical program statement controlling the colour of a 'pixel' is 'RGBColor[$r, g, b$]'. Here numbers should be substituted for the actual parameters $(r, g, b)$. Typically the beams are off (black pixel) for (0, 0, 0), and white (also the brightest colour) is produced for (1, 1, 1). The phosphors are such that RGBColor[1,0,0] looks red, RGBColor[0,1,0] green and RGBColor[0,0,1] blue. Then RGBColor[1,1,0] will produce yellow, RGBColor[0,1,1] cyan (blue–green) and RGBColor[1,0,1] magenta (purple).

Notice that in the case of these CRT screen colours the space of beams, $\mathbb{S}$, is very limited, it is a mere three-dimensional space. This means that there exists a 1–1 map of $\mathbb{S}$ into $\mathbb{C}$, although (depending on the phosphors used) only part of the colour cone can be reached. This is an instructive example because of its low dimensionality. We will use it to illustrate several of the configurations of $\mathbb{C}$ in an intuitively very obvious way.

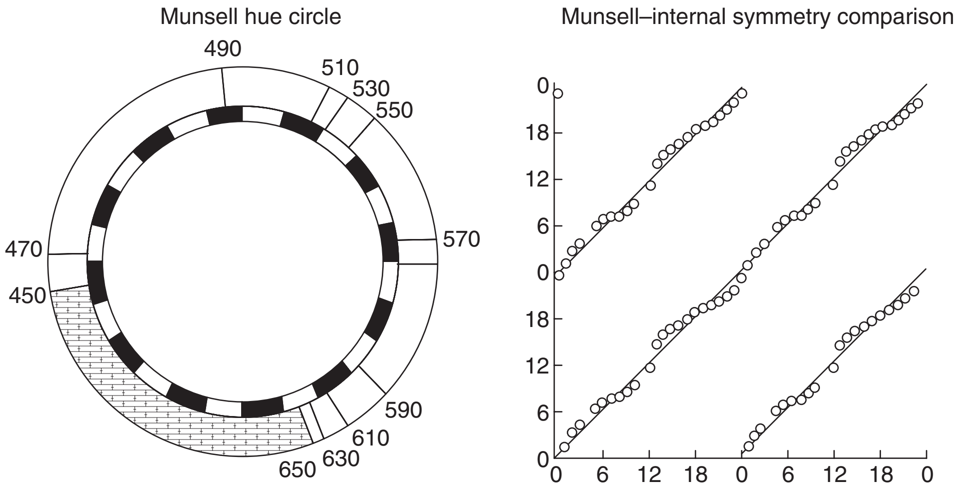


**Figure 1.34** A comparison of the mensuration of the characteristic colours with the Munsell scale (right). Both scales have been resampled to 24-hue scales. Maximum deviations are about one and a half unit on this scale, the root mean square deviation is about half a unit. Such deviations are of the same order as the deviations between perceptually uniform scales among each other, e.g. the Munsell scale and the DIN scale. Notice that the scales themselves are very non-linear, as is illustrated with the Munsell color circle (left).

First notice that all beams can be understood as selective attenuations of the white beam (1, 1, 1). The region of $\mathbb{S}$ filled with real beams is easy to construct: it is simply the *unit cube* $0 \leq r \leq 1, 0 \leq g \leq 1, 0 \leq b \leq 1$ in (*rgb*)-space. The projection in $\mathbb{C}$ will be a parallellopiped, but when we pick a nice basis for $\mathbb{C}$ we may actually use the cube in $(r, g, b)$-space as a (in this case fully congruent) model of the colour solid. We will refer to this as the RGB-cube. We can think of it as residing in $\mathbb{S}$ or in $\mathbb{C}$, in this case it makes no difference (Figs 1.35 and 1.36.)

We can use this RGB-cube to illustrate many of the geometrical properties of the colour solid in a very simple way.

First notice that spectral dominance is simply defined as $\mathcal{A}$ dominates $\mathcal{B}$ if $r_A \geq r_B \wedge g_A \geq g_B \wedge b_A \geq b_B$. Thus white (W) dominates all colours, in particular the binary colours cyan (C), magenta (M) and yellow (Y). Cyan dominates green (G) and blue (B), magenta dominates red (R) and blue (B), and yellow dominates red (R) and green (G). All colours dominate black (K), in particular black is dominated by the unary colours red, green and blue. When we plot the lattice structure as a Hasse diagram, we notice that this has the structure of the projection of a cube, in a way, the RGB-cube can double as the Hasse diagram of its dominance hierarchy. The topological structure of the RGB-cube is also apparent in the familiar tricolor diagram of additive ternary mixing of beams (it has the topological structure of the 'Schlegel diagram' of the cube, Fig. 1.37).

It is easy enough to determine the complementary beams of the ternary, binary and unary mixtures, the complementary pairs are W–K, C–R, M–G and Y–B. Thus we can

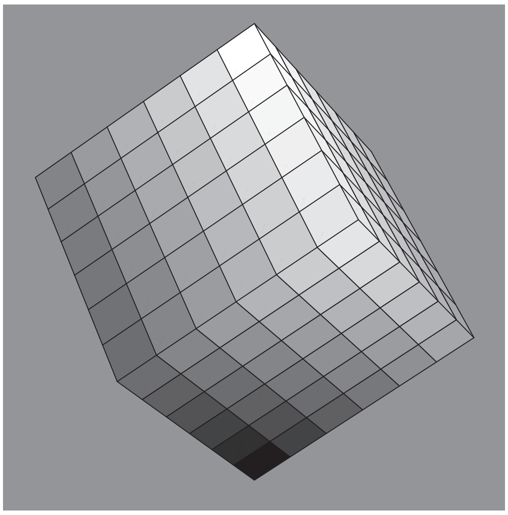

**Figure 1.35** A generic view of the RGB-cube. (See colour Plate 21 in the centre of this book.)

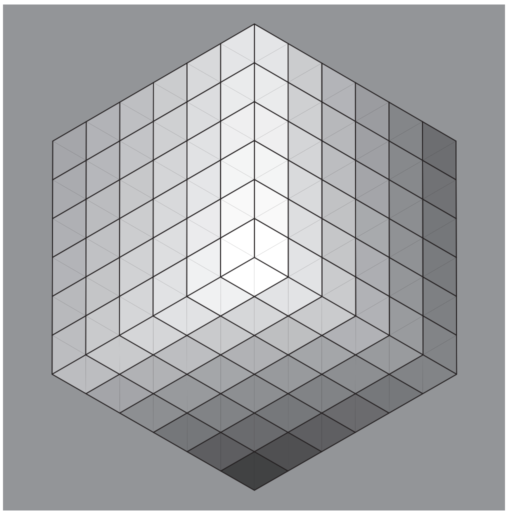

**Figure 1.36** A view at the white pole of the RGB-cube. (See colour Plate 22 in the centre of this book.)

immediately construct the inverse dominance hierarchy. In the RGB-cube the grey axis is the body diagonal W–K. Its midpoint is the median grey and it is indeed a symmetry centre of the cube. By inversion in this centre the vertices $(K, R, G, B, C, M, Y, W)$ go over into $(W, C, M, Y, R, G, B, K)$ that are the complementaries.

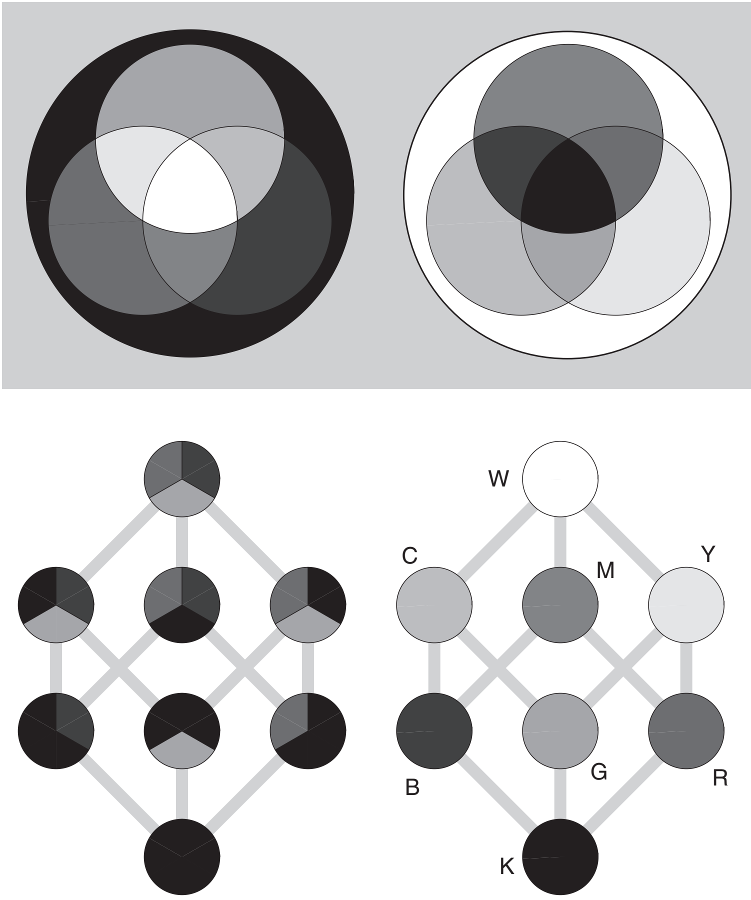


**Figure 1.37** Tricolour spot diagrams for subtractive (top left) and additive (top right) colour mixture. At the bottom, the Hasse diagram of spectral dominance, on the left with the RGB contributions explicitly drawn, at the right with the hues indicated. (See colour Plate 23 in the centre of this book.)

Notice that the 'spectral cone' is made up of the three faces of the cube that meet at the black point, whereas the inverted spectral cone is made up of the three faces that meet at the white point (they go over into each other through the central symmetry). The spectral cone and the inverted spectral cone together exhaust the surface of the RGB-cube. They intersect in the closed polygonal arc *RYGCBM*. This is the locus of the characteristic colours. In this case the colour solid coincides with the envelope (the intersection of the spectral cone and the inverted spectral cone), thus the closed polygonal arc *RYGCBM* is also the locus of the full colours or semichromes (Fig. 1.38).

The boundary colours (Fig. 1.39) are spirals between the white and black point. The long wavelength series of boundary colours lie on the polygonal arc *KRYW*, whereas the short

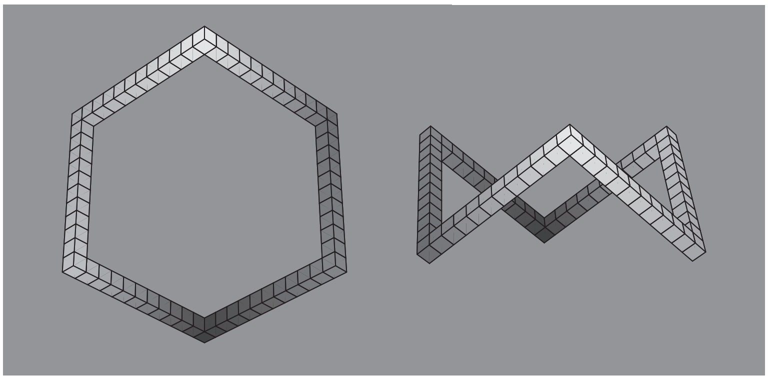

**Figure 1.38** Two views of the locus of full colours on the RGB-cube. (See colour Plate 24 in the centre of this book.)

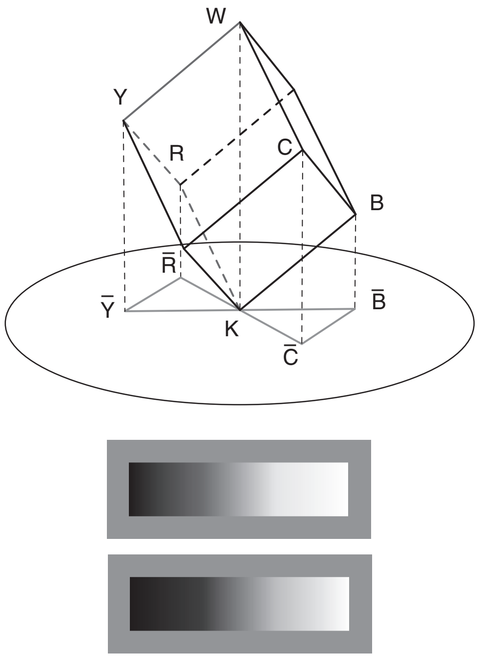


**Figure 1.39** The loci of boundary colours on the RGB-cube with sequences of boundary colour hues. (See colour Plate 25 in the centre of this book.)

wavelength series of boundary colours lie on the polygonal arc *KBCW*. When you trace these arcs on the RGB-cube you see that they are congruent twisted space curves (spirals) of opposite chirality. They divide the surface of the RGB-cube (the optimal colours) into two equal areas, the band pass optimal colours (colours from the—impure—Newtonian spectrum) and the band stop optimal colours (colours from the—impure—inverted spectrum).

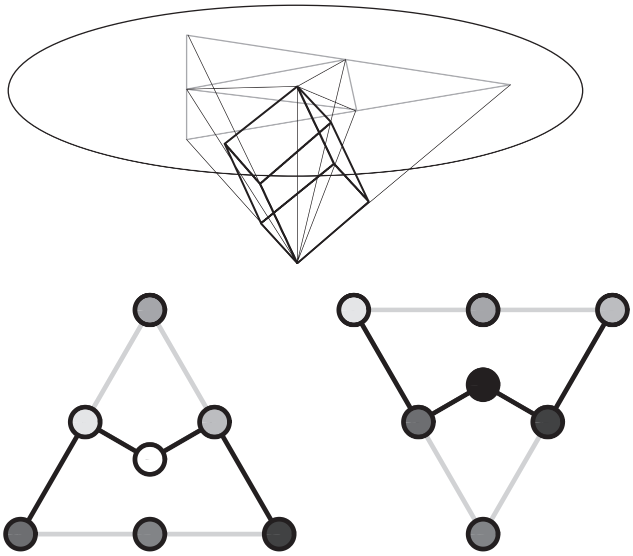

**Figure 1.40** The RGB chromaticity diagram. The boundary colour loci are plotted in the RGB colour triangle and in the complementary (inverted) triangle. (See colour Plate 26 in the centre of this book.)

We can also construct the representation of the various geometrical loci in the chromaticity diagram. One obtains a neat, symmetrical representation when one selects a chromaticity plane that is orthogonal to the grey axis. We may attach it to the white point, for instance, and project from the black point. In this case the spectral cone at the black point projects into an equilateral triangle (it degenerates into a curve). Two sides of this triangle are the 'spectral locus' (namely *RGB*), the remaining side (*BR*) is the line of purples (Figs 1.40 and 1.41). The inverted cone at the white point maps into the full interior of this boundary. We can easily find the loci of semichromes *RYGCBM*; it coincides with the boundary. The loci of boundary colours *KRYW* and *KBCW* appear much as in the CIE chromaticity diagram for the general case.

Finally, we can put an Ostwald-style atlas structure on the RGB-cube. The locus of characteristic colours is the closed polygonal arc *RYGCBM*. We may consider it mensurated (six divisions) since the vertices are equally distributed according to the equal volumes scheme. The characteristic colours can be used to label the planes of a sheaf of planes on the grey axis. Each plane on a certain characteristic colour defines a triangle *WCK* (*C* the characteristic colour). Such a triangle is a 'single page' from the atlas. We may indicate colours on the page by their colour, white and black content. Thus each point inside the RGB-cube is uniquely specified via the four numbers: index of characteristic colour (0–5), white content ($0 \leq w \leq 1$), black content ($0 \leq b \leq 1$) and colour content ($0 \leq c \leq 1$) (there are only three degrees of freedom because of the constraint $w + b + c = 1$). This is a much more intuitive and practical way to specify a colour than by red-fraction, green-fraction and blue-fraction (the 'RGBColor$[r, g, b]$' method). It is used in some 'colour pickers' for

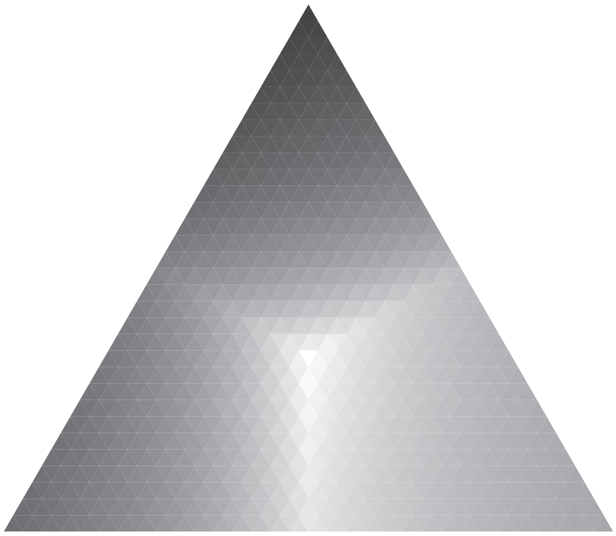

**Figure 1.41** The RGB colour triangle. At each chromaticity we have plotted the brightest RGB colour. (See colour Plate 27 in the centre of this book.)

graphics applications, notably the 'Painter' application of Fractal Design Corporation. Of course one can easily convert between the RGB and Ostwald denotations.

Thus we see that we find exactly the same geometrical structures in this simple and intuitive situation as we find in the general case (though here the formal structure is much more complicated because of the colour matching functions). Indeed, the structure of the RGB-cube can often be used to gain a more intimate understanding of the essential structure of the general case.

This throws a new light on one of the questions posed at the beginning: Is there any reason to prefer one of the many 'colour bodies' (spheres, double cones, cubes, trees, etc.)? It appears that the cube has definitely some advantages over the others since it preserves most of the structure that characterizes the general case. The reason is simply the three-dimensionality of $\mathbb{C}$.

## Discussion

Did we find any answers to the problems posed in the introduction? Here we take them in order:

- That there exist so many colour solids is largely the result of human fancy. The one feature that is common to (almost) all colour solids that have been proposed is that they are convex, finite bodies with pronounced singularities at the white and black poles. This feature, at least, has firm roots in colorimetry. The only aberrations here are the 'colour trees', which do not explicitly endorse convexity (nor finiteness for that matter), and systems such as Lambert's and Helmholtz's, which are really based on colours in the aperture mode (no black). The latter are strikingly at odds with the (very fundamental) central symmetry of the colour solid.

- Since biological systems (by all odds) developed outside the laboratory, they are set to deal with the structure of continuous spectra, not monochromatic beams. The surface colours are most naturally parameterized via their common boundary, the optimal colours. The 'equator' of the optimal colours (full colour locus) is a closed curve. Thus the hues quite naturally fall into a *periodic* order. The open, linear segment of the Newtonian spectrum is to be considered a laboratory artefact with little relevance for natural vision (monochromatic paints are black!).
- The warm and cold colour families have not been formalized as colorimetric concepts up till now (at least as far as we are aware). We have shown that the dichotomy is implicit in the most primitive, affine colorimetric structure, one does not even need the notion of an achromatic beam for its definition. It is to be considered one of the most basic invariants of colour space.
- Newton's 'homogeneous lights' are nothing special. Although they often occur in the formalism they rarely occur in real life and are to be considered more of a laboratory artefact than a basic building block. Ostwald was on the right track when he replaced the monochromatic beams with his *Vollfarben*. Indeed, one of the more rational bases for object colours are the attenuated optimal colours. At least they are complete. The Newtonian (impure) spectral colours represent only half of the object colours. Goethe's *Kantenfarben* are in many formal respects similar to monochromatic beams. They are more robust than monochromatic beams and can actually be produced easily in the laboratory [technically, they are the 'primitives' (cumulated integrals) of the monochromatic spectra]. The relation between the full colours and the boundary colours is an intimate one. 'White' is simply one of the ideal colours, the highest (LUB) in the spectral dominance hierarchy. To consider white a 'confused mixture of homogeneous lights' buys one nothing, to consider it the 'mother of all colours' is closer to the actual state of affairs.
- There does exist a natural metric, it is simply the metric provided by the physics. Since colour space is essentially a low-dimensional image (linear projection) of the space of beams, we have an induced metric for the colours. Thus not all affinely equivalent copies of colour space are equal—the large majority of them shows us only a deformed view of fundamental space. Only a small set (related via isometries) shows us an undeformed, 'true' view of the structure of fundamental space. Even among those 'nice' colour spaces we can make a rational choice and pick a truly 'canonical' one: we might pick a preferred spatial orientation on the basis of the fiducial directions specified via the achromatic beam and the plane separating the warm and cold colour families. Such a choice has the obvious advantage that it can be re-established from first principles (and colorimetric experiments, namely 'gauging the spectrum') without arbitrary international agreements such as the CIE 1931 basis.
- There are indeed principled ways to mensurate the colour circle. This may be understood in different ways—one may either mensurate the full colour locus (Ostwald's choice) or the locus of characteristic colours. As we have shown, the latter choice has the advantage

that it depends only on the colour, but not on the spectrum of 'white'. Principled methods are *invariant* against variation of the primaries. Ostwald's principle of internal symmetry is an attractive idea, and indeed contains the nucleus of a solution, but fails because the full colour locus is non-planar. We have identified two invariant methods: a division by arc length in the canonical basis and a division by volumes (incremental volume between two characteristic colours is the volume of the tetrahedron defined by the two full colours, white and black). The methods seem to offer slightly different advantages and we won't venture a final choice here though the latter can be shown to be the natural (and correct!) generalization of Ostwald's somewhat mystical principle of internal symmetry. An exciting empirical finding is that the results of 'eye measure' (purely psychological 'uniform spacing') are quite similar to each of the principled methods. This indicates that vision uses a metric close to the one induced from the physics, and suggests that one should prefer a rational method in practice. Results are probably not significantly different from a pragmatic point of view. The principled methods are well reproduceable and guarantee globally uniform results, whereas 'eye measure' scores badly on both counts.

- The colours of colorimetry are indeed a truthful (though limited) reflection of the physical structure of electromagnetic radiation. Colour vision—in the approximation of colorimetry—is largely a form of *low-resolution spectroscopy* and the 'observer's share' appears to be minimal. This is both an exciting and prima facie surprising fact (in retrospect, however, one should probably expect the results of evolution to converge to such a state of affairs). It is exciting because it offers us a powerful handle on the interaction of humans with their physical environment. It is surprising because only the minutest move from pristine colorimetry gets us into situations where we are at a loss to predict even the simplest empirical facts relating to 'colour vision'; this is where (puritan) psychophysics ends and psychology starts.

Apart from the key questions, we have discussed several more technical points. In the process we have introduced a few novel concepts and developments that readers with some background in colorimetry may have noticed.

## References


1. **Anonymous**. ISCC–NBS color-name charts illustrated with centroid colors. Standard sample No. 2106, Supplement to NBS Circular 553.
2. **S. Axler**. *Linear algebra done right*. Springer-Verlag, New York, 1996.
3. **F. W. J. Billmeyer**. Survey of color order systems. *Col. Res. Appl.*, **12**:173–186, 1987.
4. **M. Born and E. Wolf**. *Principles of optics, electromagnetic theory of propagation, interference and diffraction of light* (2nd edn). Pergamon Press, Oxford, 1964.
5. **P. J. Bouma**. Zur Einteilung des Ostwaldschen Farbtonkreises. *Experientia*, **2**:99–103, 1946.
6. **P. J. Bouma**. *Physical aspects of colour*. N.V. Philips Gloeilampenfabrieken, Eindhoven, 1947.
7. **G. Brindley**. *Physiology of the retina and the visual pathway*, Edward Arnold, London, 1970.
8. CIE. *CIE proceedings*. Cambridge University Press, Cambridge, 1924.
9. CIE. *CIE proceedings*. Cambridge University Press, Cambridge, 1931.

10. **M. C. Company**. *Munsell book of color*. Munsell Color Co., 10 East Franklin Street, Baltimore, MD, 1929.
11. **C. Dolland**. Über den Luther-Nybergschen Farbkörper. *Die Farbe*, **5**:113–136, 1956.
12. **R. M. Evans**. *An introduction to color*. John Wiley & Sons, New York, 1948.
13. **A. Gershun. Sur une théorie du champ lumineux**. *Rev. Gen. de l'Elec.*, **44**:307, 1938.
14. **A. Gershun**. The light field. English translation by Moon and Timishenko. *J. Math. Phys.*, **18**(51), 1939.
15. **H. Grassmann**. Zur Theorie der Farbenmischung. *Ann. Phys.*, **89**:69–84, 1853.
16. **H. Grassmann**. Bemerkungen zur Theorie der Farbenempfindungen. *Va. Selbstanzeige von V. Königsb. Rep. II*, pp. 213–221, 1879.
17. **E. Hering**. *Grundzüge der Lehre vom Lichtsinn*. Julius Springer, Berlin, 1920.
18. **T. Holtsmark**. Newton's experimentum crucis reconsidered. *Am. J. Phys.*, **38**:1229–1235, 1970.
19. **T. Holtsmark**. Das Experimentum Crucis und die Theorie der Dispersion. *Opt. Acta*, **18**:867–873, 1971.
20. **T. Holtsmark** and **A. Valberg**. On complementary color transitions due to dispersion. *Am. J. Phys.*, **39**:201–204, 1971.
21. **D. Judd**. *Color in business, science and industry* (1st edn). John Wiley & Sons, New York, 1952.
22. **J. J. Koenderink**. Color atlas theory. *J. Opt. Soc. Am.*, **A4**:1314–1321, 1987.
23. **H. Küppers**. *Das Grundgesetz der Farbenlehre*. DuMont Buchverlag, Köln, 1978.
24. **J. Lambert**. *Photometria sine de Mensura et Gradibus Luminis Colorum et Umbrae*. Eberhard Klett, Augsburg, 1760.
25. **E. H. Land**. Experiments in color vision. In R. Held and W. Richards, editors, *Readings from scientific american* (originally *Scientific American* May 1959), number 28, pp. 286–298. W. H. Freeman and Company, San Francisco, 1972.
26. **R. Longhurst**. *Geometrical and physical optics* (1st edn). Longmans, Green and Co., London, 1957.
27. **R. Luther**. Aus dem Gebiet der Farbreizmetrik. *Z. f. techn. Physik*, **12**:540–558, 1927.
28. **J. C. Maxwell**. XVIII. Experiments on colour, as perceived by the eye, with remarks on colourblindness. *Trans. of the Roy. Soc. Edinburgh*, pp. 275–298, 1855.
29. **J. C. Maxwell**. On the theory of compound colours, and the relations of the colours of the spectrum. *Phil. Trans.*, pp. 57–84, 1860.
30. **A. Munsell**. *A color notation* (10th edn). Munsell Color Co., 10 East Franklin Street, Baltimore, MD, 1947.
31. **I. Newton**. A new theory about light and colors. *Am. J. Phys.*, **61**:108–112, 1993.
32. **I. Newton**. *Opticks*. Dover Publications, New York, 1952.
33. **N. Nyberg**. Zum Aufbau des Farbenkörpers im Raume aller Lichtempfindungen. *Z. f. Physik*, **52**: 406–419, 1928.
34. **W. Ostwald**. Das absolute System der Farben. *Z. f. physik. Chem.*, **92**:222–226, 1917.
35. **W. Ostwald**. *Er und Ich*. Theodor Martins Textilverlag, Leipzig, 1936.
36. **C. Parkhurst** and **R. L. Feller**. Who invented the color wheel? *Col. Res. Appl.*, **7**:217–230, 1982.
37. **S. Quiller**. *Color Choices, Making Color Sense out of Color Theory*. Watson-Guptill Publications, New York, 1989.
38. **P. O. Runge**. *Farben-Kugel oder Construction des Verhältnisses aller Mischungen der Farben zu einander, und ihrer vollständigen Affinität, mit angehängtem Versuch einer Ableitung der Harmonie in den Zusammenstellungen der Farben*. Friedrich Perthes, Hamburg, 1810.
39. **J. Ruskin**. *The Elements of Drawing*. Dover Publications, New York, 1971.

40. **E. Schrödinger**. Grundlinien einer Theorie der Farbenmetrik im Tagessehen. Parts 1 and 2. *Ann. Phys.*, **63**:397–456, 1920.
41. **E. Schrödinger**. Grundlinien einer Theorie der Farbenmetrik im Tagessehen. Part 3. *Ann. Phys.*, **63**:481–520, 1920.
42. **E. Schrödinger**. Theorie der Pigmente von grösster Leuchtkraft. *Ann. Phys.*, **62**:603–622, 1920.
43. **G. Strang**. *Linear Algebra and its Applications*. Academic Press, New York, 1976.
44. **J. W. von Goethe**. *Theory of Colours* (6th edn). The MIT Press, Cambridge, MA, 1982.
45. **J. W. von Goethe**. *Farbenlehre, mit Einleitungen und Kommentaren von Rudolf Steiner*, volumes 1–3 (3rd edn). Verlag Freies Geistesleben, Stuttgart, 1984.
46. **H. von Helmholtz**. *Handbuch der Physiologischen Optik* (2nd edn). Voss, Hamburg, 1896.
47. **G. Wyszecki** and **W. S. Stiles**. *Color science. concepts and methods, quantitative data and formulas.* John Wiley & Sons, New York, 1967.

## Commentaries on Koenderink and van Doorn

## From physics to perception through colorimetry: a bridge too far?

*Donald I. A. MacLeod*

Koenderink and van Doorn's 'Perspectives on colour space' is a landmark contribution, a uniquely scholarly and insightful synthesis of old and new ideas. Unlike the authors I regard it as a contribution to the literature of 'colour science' as well as of colorimetry. It can, at least, be read with profit by all colour scientists (who, whether they like it or not, have to deal with colorimetry). I admire it for its range and clarity, its bold originality and its sometimes amazing ingenuity. But my admiration is tempered by alarm.

One of the things I like about the essay is its liberating effect. It loosens the grip of the Newtonian paradigm on current thinking about colour and colorimetry. By focusing on surface colours instead of on the monochromatic lights into which Newton decomposed the spectrum, it draws attention to interesting but neglected avenues for exploration in the physics of colour, and revives the unduly neglected tradition of Goethe, Schopenhauer, and Ostwald. The chapter has much to teach us, both by precept and example. For instance, it is the best extant demonstration of the advantages of adopting a full three-dimensional geometrical description of the colour stimulus, rather than the two-dimensional projections (chromaticity diagrams) that most discussions have preferred to concentrate on.

What is alarming is that, in their attempt to say something interesting about colour based on physics only, the authors have enhanced the attractiveness, to unwary readers, of what Mausfeld (2002) has called the 'physicalist trap'. The physicalist trap is the attempt to give purely physical accounts of perceptual phenomena, and it is a conceptual pitfall into which we are all prone to fall (with the frequent exception of physicists themselves, from Newton with his famous insistence that 'the rays are not coloured' to the authors of the present chapter). We fall into the trap because evolution has made us victims of what might be called the *illusion of objectivity*: the pre-scientific starting point for any scientific consideration of perception is the intellectually naive but biologically necessary conviction that the phenomenal world of perception is nothing different from the world of external physical reality. This illusion of objectivity remains in the cognitive background as a conceptual pitfall when we try to develop scientific accounts of perception.

The authors are not naive realists, and they are in no real danger of falling into the physicalist trap themselves while they so successfully embellish it. They simply invite the reader to note the intriguing qualitative and quantitative correspondences they have discovered between the world of physical colour stimuli and the world of colour appearance. Their project is to promote a view in which the naive notion of identity of the phenomenal and the physical is left behind, but the notion of a simple (albeit limited and qualified) isomorphism between the two is retained.

Although the tendency of their work is in this sense merely quasi-reductionist rather than genuinely reductionist, I believe it is dangerous. Even if the physicalist trap fails to capture the authors or their sophisticated readers outright, it can have insidious effects by its mere presence in the cognitive background, because it lends a spurious significance to parallels between the physical and phenomenal worlds—parallels sometimes so indirect that they may be almost accidental. This chapter identifies many such parallels that are novel and intriguing, but whose significance becomes questionable on close scrutiny. And while it might be interesting to endow these distant parallels with significance within either a mechanistic framework (with a basis in physiology) or an evolutionary framework (with a basis in ecology, including natural scene statistics), it should be clear that any such project would have to invoke an elaborate set of new and disagreeably uncertain assumptions. The authors have understandably chosen not to emphasize such difficulties and uncertainties in their enterprise of tracing colour to its physical roots. In what follows, I try to do this for them.

## From physics to colorimetry

The authors want to limit themselves to colorimetry, excluding other aspects of colour perception (although, as we will see, they occasionally stray beyond that self-imposed boundary). Colorimetry is the part of colour vision most closely tied to physics. Indeed, Koenderink and van Doorn characterize colorimetry, with the colour-matching functions on which it is based, as 'almost pure physics'. They recognize that all colorimetry depends on the form of the cone spectral sensitivities, but they refer to those as 'accidental', and dismiss them (in their note 5) as an unnecessary complication, as if the colour-matching functions belong to physics, but the cone excitations to physiology.

Yet in reality the cone excitations determine the colour-matching functions. The judgements of a subject about a colour match—on which colorimetry depends—are therefore at least as 'accidental', and at least as divorced from 'pure physics', as the cone excitations that the compared stimuli elicit in the subject's retina. How, then, could physics include the colour-matching functions but exclude the cone excitations? That conception of the scope of physics may draw its plausibility from the illusion of objectivity, which encourages us to identify the phenomenal world of colour with the physical world of colour stimuli, and to neglect, when making this identification, the intervening physiological processes on which phenomenal experience undoubtedly depends.

Be that as it may, the cone photoreceptor excitations that the authors are careful to exclude from their discussion have provided natural and convenient coordinate systems for characterizing colour stimuli, both in the early days of colorimetry (Luther, 1927) and in more recent discussions (MacLeod and Boynton, 1979). By abandoning this obvious, but overtly physiological, choice of coordinates, Koenderink and van Doorn find themselves faced with an embarrassment of choice among various more or less arbitrary coordinate systems for representing the colour stimulus. I consider next the choices that they make and that determine their perspective on colour space.

## The quest for physical invariants

Koenderink and van Doorn opt to use as a starting point the CIE XYZ coordinate system, whose claim to reality is less physical than sociological–historical. Recognizing the conventional nature of that framework, they cleverly transform it, following Cohen and Kappauf, to a physical stimulus representation that is invariant with the initial choice of primaries in terms of which the colour-matching functions of a given observer are expressed. As they explain (p. 17), in this representation, any given spectral energy distribution is decomposed into two component spectral energy distributions: a 'fundamental component' and a black component. The fundamental component is an energy distribution (virtual, because sometimes negative) constructed as a linear combination of three primaries, each with a spectral energy distribution proportional to one of the colour-matching functions for an equal energy spectrum—indeed, it is the one such combination that matches the given beam. Each row and each column of matrix **R** is the fundamental component, in that sense, of some monochromatic beam from the equal energy spectrum. If now the colour matches of the observer are described in terms of new primaries—say, red, green, and blue instead of X, Y, and Z—the colour-matching functions themselves will change. But their weighting as primaries in the construction of the fundamental component of any given beam must change correspondingly, so as to keep the fundamental component itself the same (since it must match the given beam).

But what is gained by this detour back to the colour-matching functions in the choice of the primaries from which a beam's 'fundamental component' is built up? After all, the beam could equally well be regarded as the combination of *any* of its metamers with some suitably chosen black: the fundamental component preferred by the authors is just one choice among the infinity of possibilities. It could be misleading, therefore, to think of this 'fundamental' spectrum as 'the unique causally efficient part of the beam' (p. 17). It has that status only if the particular, somewhat arbitrary decomposition favoured by the authors is adopted.

Not only the fundamental status of the fundamental components and of the matrix **R**, but their invariance as well, must be qualified: the fundamental and matrix **R** remain completely tied to convention in their dependence on the equal energy spectrum. There is no obvious physical warrant for giving that spectrum a special status (as opposed to, for instance, one with equal quantum flux per unit frequency), and no ecological one either. In view of these considerations it is not clear when, or why, we might prefer to calculate the 'fundamental' spectral component of a stimulus, instead of (for instance) the corresponding triplet of cone excitations. The decomposition into the invariant fundamental spectrum plus black is formally elegant but lacks practical utility. The only use the authors suggest for it is to generate metamers for a given beam. Surely the cone excitations have a less hollow claim to fundamental status: even in the assessment of the degree of metamerism between two given beams, the cone excitation triplet provides a much simpler alternative representation of the visually effective stimulus, in the form of three quantities that are literally fundamental for colour vision, including colour discrimination.

Much as Koenderink and van Doorn would like to banish them, the spectral colours not only appear in the matrix **R** but return to haunt them persistently in their later enlightening discussion of full colours and of the colour solid for reflecting surfaces. The difficulty here again arises from a lack of invariance. The colour solid, including the locus of full colours or semichromes (p. 27) in particular, is illumination dependent. The envelope of that solid for all possible illuminants is very different from any particular realization of the solid, and indeed is none other than the convex hull of spectral colours: with no restriction on illuminant power, the colour solid can contact that spectral cone at any point, as is clear in the limiting case of intense monochromatic illumination. Koenderink and van Doorn show that a particular closed curve of 'characteristic colours' will form, when linked to white and black, the envelope of the colour solid for all illuminants that share a particular colour (metameric illuminants). But these characteristic colours are again simply the colours of monochromatic lights. (Each has the maximal intensity that could be remitted by a white surface under any illumination metameric with the prevailing one. That intensity is achieved when the illuminant is just a mixture of the monochromatic light and some other spectral light or purple.) The search for invariance here forces a return to the decomposition of the spectrum into monochromatic lights.

This lack of invariance is an unwelcome complication for the chapter's enterprise of providing a general treatment of the colorimetry of reflecting surfaces. But it may be helpful to the visual system in the context of colour constancy. The gamut—the solid shape in colour space that can be filled by surface colours under a given illuminant—is an unduly neglected physical constraint in natural colour vision, a constraint that this chapter explains with unprecedented clarity and elegance. And just because it is illuminant-dependent, the gamut affords cues that could be very useful for estimating the illuminant and achieving a relatively illumination-invariant estimate of surface colour in natural scenes. This possibility has been explored by Forsyth, and is a focus of Chapter 7 (this volume) by MacLeod and Golz. The vague notion of those authors that 'when the light gets red the reds get lighter' has a definite realization in the behaviour of Koenderink and van Doorn's illuminant-dependent locus of characteristic colours. Under a red illuminant, the inverted spectral cone that tapers to white is translated redward, and the locus of characteristic colours, where the origin-centred spectral cone meets the inverted cone, gets tilted away from the origin on the red side. This serves to illustrate how the shape of the full colour solid, hence the illuminant colour, can provide a cue to illuminant colour (if there are enough visible samples from the colour solid to give information about its shape). But the theoretical utility of the spectral characteristic colours in this and other contexts may be limited because they are a very extreme case, encircling the colour solids for typical natural illuminants at a far greater distance from the achromatic axis. Here invariance comes at the price of the artificiality that the authors have been keen to avoid.

## Colorimetry and perception

The discussion of canonical bases (p. 18) suggests that a particular linear transform of colour space offers a natural or undistorted view of it, unique except for a choice of viewpoint through rigid rotation. The very idea that some particular view of colour space can be regarded as undistorted, with a faithful rendition of distances and angles, is a provocative one. In proposing it, the authors stray beyond the domain of colorimetry into that of colour perception, since (as they note) colorimetry proper does not constrain distances or angles. This makes it all the more interesting if an undistorted view can be determined on the basis of physics (as represented in the colour-matching functions) alone. And it is, indeed, intriguing that the distances in the undistorted view of colorimetric colour space tally so well with the Munsell distances that represent phenomenal colour differences.

Unfortunately, though, the 'undistorted' basis functions obtained by singular value decomposition (SVD) have hardly a better claim to physical, biological or psychological reality than the colour-matching functions from which they were derived. Like the untransformed space of 'fundamental components' discussed above, the structure derived using SVD is satisfyingly independent of the initial choice of primaries, but is strongly affected by the conventional choice of the equal energy spectral colours as a stimulus set. So the SVD 'canonical' basis can be faulted on the same grounds that its authors fault so much of contemporary colorimetry: it gives an unjustified primacy to the spectral lights (and to the equal energy spectrum in particular). To be consistent with the authors' position elsewhere in this chapter, the suggestion that the brain achieves 'an undistorted representation of the physical structure of beams' (Fig. 1.11) should be evaluated with reference to natural colours, not the Newtonian spectrum. Recent investigations of the coding of natural colours, for instance the chapter by MacLeod and von der Twer (Chapter 5 this volume), illustrate this approach. But any set of colours is a somewhat arbitrary choice, so the derived view of colour space will always lack the objectivity and generality that we expect for a datum of physics.

The canonical orientation suggested for colour space adopts as one axis the '*achromatic axis*' (p. 23) that is said to represent points of indeterminate hue. But the existence and location of such an axis are facts of psychology, not of physics. An observer might, for instance see all realizable colours, including any candidate 'whites' suggested by physics, as reddish. The authors acknowledge such difficulties by saying that the choice of which axis to identify as achromatic is completely arbitrary; yet if the choice is truly arbitrary, surely nothing but confusion can result from labelling it the achromatic axis.

This problem is compounded when the authors designate Ostwald's semichromes as the most 'colourful' colours: the semichromes (or full colours) must indeed be optimal in *colorimetric* purity in the sense illustrated in Fig. 1.15, but this carries no implications whatever about their appearance. More generally: *the cone sensitivities, and the colorimetric data that they determine, place no constraint whatever on the way that colour appearance varies across physical colour space.* Failure to appreciate this point is a serious, though ubiquitous, theoretical error. It is an intellectual blind spot fostered by the illusion of objectivity: we assume an objective physical basis for qualities of our experience that are in reality accidents of our physiology (or if you prefer, of our psychology). This point deserves elaboration. For concreteness, consider observers who have the normal three cone photoreceptors together with three or more postreceptoral neural signals that determine colour appearance. The postreceptoral signals each depend, let's say continuously but non-linearly, on all the cone excitations. Even this minimally complex and familiar colour vision system allows each of the three colour signals to take any value at any point in colour space (subject only to the continuity constraint). There are many possible structures for the phenomenal colour space of observers of this general type (a class which doubtless includes humans). Such an observer might, for instance, perceive the entire locus of semichromes as perfectly achromatic rather than maximally coloured, with colours inside that locus in the chromaticity diagram appearing greenish (for example) and colours outside it appearing reddish. The 'achromatic axis' could be perceptually the greenest colour of all! And this is among the *least* outlandish of the possibilities . . . .

In fact, of course, the semichromes are indeed relatively colourful in appearance (more so than the necessarily very dark, narrow-band reflectances), but this has nothing directly to do with their special physical nature. This chapter provides no physical, physiological, or functional rationale for expecting any correspondence between colorimetric purity and colourful appearance (nor does that correspondence hold strictly even as a matter of observation: exceptions are noted later in the chapter).

Although physics does not tell us why stimuli on a particular colorimetric axis appear perceptually achromatic, the chapter by MacLeod and von der Twer (p. 155) attempts an answer based on ecology, and on efficiency of neural coding. There it is suggested that for efficient use of noisy and compressively non-linear neural signals, extremes of sensation (and maximal signals) should be associated with stimuli close to the boundary of the probability density function for natural stimuli, while the most frequent stimuli are perceptually encoded as close to neutral. This simple postdiction is approximately fulfilled: in any reasonable colour space, the most frequent natural stimuli are ones that appear nearly white (though often slightly yellowish and/or greenish). Koenderink and van Doorn may be excused, or even congratulated, for having avoided such messy lines of enquiry. But, is it realistic to hope that less messy answers could ever suffice for such inherently messy questions?

Along with the achromatic axis, the authors identify a new axis in colour space, that links the origin with the spectral boundary between *warm and cold* colours. This boundary they locate in the spectrum at 537 nm, which they show is, in any linear colour space, the farthest point on the equal energy spectrum locus from the plane of purples. This colorimetric definition of warm and cool colours is a novel and intriguing proposal. But readers should ask themselves: (1) how the proposed correspondence should be explained or understood, if not as a coincidence; and (2) why it should hold for the equal energy spectrum in particular (other spectra, such as that of sunlight, place different wavelengths farthest from purple).

A further technical point is pertinent. *Strictly, the unique plane of purples is a fiction.* The notion of a limited spectrum has no support from physics, and behavioural data also fail to support it in the required sense of suggesting tangent directions at the two spectral extremes. Very short wavelengths, in particular, exhibit no well-defined limiting chromaticity, so no 'plane of purples' is objectively given. Instead, a conventional choice of wavelengths for the ends of the spectrum defines the set of colours identified as the plane of purples, and this in turn determines the wavelength most distant from that plane.

Less arbitrary, though more overtly biological, axes can be suggested for colorimetry. An axis representing *luminance* is a standard choice, roughly a substitute for the 'achromatic' axis. Koenderink and van Doorn cheerfully characterize the status of luminance as 'rock bottom'. Yet in the context of human vision, the status of energy is lower still (unless one's interest is in cooking the retina). To represent a local light stimulus by a single number, luminance is the only reasonable choice. It not only corresponds roughly to perceived brightness, but also determines, among other things, visual performance in acuity, reaction time, and temporal resolution. Luminance-normalized cone excitations provide serviceable chromaticity coordinates (Luther, 1927; MacLeod and Boynton, 1979), and the addition of luminance itself to those (Derrington, Krauskopf and Lennie, 1984) provides a useful, albeit not physically derived, three-dimensional representation of the colour stimulus. In these physiologically based representations of colour, the $\mathbb{S}$ cone excitation axis provides a second axis, naturally orthogonal to the luminance axis and likewise capturing directly information relevant for (among other things) spatial and temporal resolution and distinctness of borders.

As these competing examples suggest, the considerations that have been adduced for favouring certain candidate axes for colour space do not come from the domain of physics. They derive instead from psychophysics or physiology. But merely by preferring certain axes to others, we are leaving the domain of colorimetry proper. As this chapter makes clear, colorimetrically relevant properties of colour space are unaffected by origin-preserving linear (affine) transformations. Thus if our concern is purely with colorimetry, the very enterprise of searching for canonical axes is at best unnecessary and at worst misguided.

It is natural to want to use the freedom of axis choice to capture something meaningful outside of the domain of colorimetry, and that is what all the candidate choices attempt to do. This is harmless in itself. But in making such a choice, it is important to remember how restricted are the explanatory limits of colorimetry *per se.* For colorimetric purposes, a colour space merely has to indicate whether colour stimuli match (have the same coordinates). If we take the liberty of adopting the cone spectral sensitivities (or if you prefer, the colour-matching functions) as a physical datum, physics determines that much of colour vision—but no more: the physically determined coordinates need not indicate anything at all about how two distinct colours relate to one another in their appearance. Two avenues, neither of them through physics, allow us to advance our understanding of how colours look. Mechanistic hypotheses must involve physiological postulates or data, as in the tradition of Hering, together with implicit or explicit psychophysical linking hypotheses (Brindley). And teleological or evolutionary hypotheses must involve ecological postulates or data that pertain to the organism's visual interactions with its environment.

### Conclusion

Although my purpose in this postscript has been to register objections and suggest qualifications to a few of the chapter's claims, I hope those remarks will only encourage readers to investigate Koenderink and van Doorn's magnificent analysis of the structure of colour space more comprehensively. But to those readers I say again: beware—this is treacherous, though fascinating, territory, to be explored with caution! The physicalist trap, that ever present pitfall in the study of perception, has here been most dangerously camouflaged by the artful physicist authors. Even if their intentions are innocent, avoiding the trap is up to you alone.

## Reference

**Brindley, G. S.** (1970). *Physiology of the retina and visual pathway.* Williams and Wilkins, Baltimore.

**Derrington, A. M., Krauskopf, J., and Lennie, P.** (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology (London)*, **357**: 241–65.

**Luther, R.** (1927). Aus dem Gebiet der Farbreizmetrik. *Z. Tech. Phys.*, **8**, 540–58.

**MacLeod, D. I. A. and Boynton, R. M.** (1979). Chromaticity diagram showing cone excitation by stimuli of equal huminance. *Journal of the Optical Society of America*, **69**, 1183–6.

**Mausfeld, R.** (2002). The physicalistic trap in perception. In *Perception and the physical world.* (ed. D. Heyer and R. Mausfeld), pp. 75–112. Wiley, Chichester.

## Commentaries on Koenderink and van Doorn

## Colorimetry fortified

*Paul Whittle*

Colour has long elicited geometrical representation. In this chapter, by two physicists steeped in the intuitive geometry of Hilbert and Cohn-Vossen (1932), we have the best summary exposition and tutorial of colorimetry and its associated colour spaces that I know of. Each reading raised something new for me, not least gaps in my own understanding. The authors are at pains to stress the austerity of their subject matter, how it deals only with arid and abstract things of little interest to the general reader, overlaps hardly at all with 'colour science' and eschews completely the 'mess' of 'the world of colour'. At least the first disclaimer can be taken with a pinch of salt. For these two physicists are no more able than the rest of us to avoid the seductions of colour, and so cannot resist introducing the fertile notion of an achromatic point, and later on several 'intuitions' that take them further. It is *fortified* colorimetry ('supplied with added nutrients', *Oxford English Dictionary*). For me one of the challenges of the chapter was to spot what was being slipped in and where, to work out whether its appearance of going a long way on remarkably little fuel was a conjuring trick or not. I try to do it here for the colour circle.

It should be noted that the authors give a *principled* account. They are not at all concerned with empirical details. They do not mention, for example, that the results of metameric matching can vary considerably with the stimulus parameters. I find this refreshing in a field that is often so empiricist that *understanding* is continually postponed. I wish we had more of it.

### Colorimetry is independent of photometry

The beginning student of colour naturally assumes that colorimetry must be more complex than photometry. Is it not three-channels versus (effectively) one? The authors show that the opposite is the case. Colorimetry has a formal elegance that quite escapes photometry. At first sight this is a little puzzling, because colorimetry, of course, includes the intensity dimension. But its intensity is purely radiant intensity: a scalar multiplying the absorptions of all three pigments by the same factor. The photometry question is quite unrelated: the relative effectiveness of lights of different chromaticity. There are many criteria for effectiveness and the results do not agree exactly ('very low (really rock bottom!) scientific status'). Consensus has been imposed only by committee. Colorimetry, however, for all its formal elegance, has not escaped the same fate. The idealizations of science are one thing, their application another—see Johnston (2001), for a good account of the committee wrangles in both fields. But the contrast is instructive: colorimetry is biophysics, it only scrapes the surface of the organism, the photopigments, and ignores the nervous system proper. Whereas the effectiveness that photometry is concerned with involves the whole organism.

### Colorimetry does reduce the stimuli

The authors point out that the conventional use of isolated lights in dark surrounds for metameric matching is unnecessary. It can just as well be done with a small patch in the middle of a natural scene. It is response reduction, not stimulus reduction, that is crucial. The subject has only to respond 'same' or not and this can be reduced to the detection of an edge in a bipartite field or a flicker in a temporally alternated one. No reports of colour are needed; no colour concepts. Animals can easily be trained to do it.

But there is stimulus reduction too. If for you the essence of colour—your most vivid experiences of it—is in the play of colour, light and shade in a complex changing scene, such as a wood on a sunny spring day with a slight breeze moving the foliage, or a glistening salmon run in the hills, then working with one small textureless element at a time is as severe a stimulus reduction as could be. It excludes the greater part of what makes the world colourful for you. And that will never be restored. Colorimetry is resolutely atomistic.

But the authors' fortified colorimetry is not quite so atomistic. Their exposition deals with the full (though static) manifold of all visible 'beams', and the colour concepts that fortify it depend, as we shall see, on having the manifold there.

## Spectrum to colour circle

A 'major topic of this paper' is 'Why are most colour order systems based on the "colour circle", that is a *periodic* linear sequence whereas the *spectral* colors are naturally ordered as a linear open segment?' After reading it I felt I understood this much better than before. In this section I try to say how, and to pick up what presuppositions are being introduced along the way.

First, where does the cone of colours in colour space come from? We start with a 'space of beams' in which each axis corresponds to the radiation at one wavelength. 'Geometrical intuition suggests that all beams fill an "∞tant" . . . in the linear space.' It is this manifold that projects into the cone of colours, via the projection operator that 'gauging the spectrum' empirically determines. It gives ready-made a single bounded cone with all colours nicely inside the monochromatic ones which, together with the purples, constitute its surface. Note that the cone can expand or contract a little depending on the choice of wavelength interval, which can only be pragmatic: neither too wide nor too narrow. Note also that the nice projected cone is not pure geometry. Biology also comes in (note 9). Where do the non-monochromatic purples come from that close off the cone? They arrive because we are projecting a *filled* ∞tant (the manifold of all visible beams) so these mixtures of far red and far blue are already included. Their particular plane in the space of beams just finds itself on the outside of the cone of colours because of the form of the projection operator (our photopigments plus the particular experimental procedure).

I'm jumping the gun in referring to 'colours' and 'purples'. So far these are purely formal entities: 'metamers', projections of subspaces of indistinguishable beams. To start talking about colours as we know them, we need *appearances*, which, as the authors note, takes us outside pure colorimetry. And to get from the bounded cone of metamers to the colour circle proper, we also need more structure: in order, chromaticities, chromaticity planes, an achromatic point, and semichromes.

Chromaticities first. 'If one attenuates a beam ("sunglasses") one notices a decrease in "brightness" whereas the "colour" in some restricted sense appears to remain invariant . . . This is the reason why one often finds it convenient to regard colours modulo their magnitudes . . . and call them "the chromaticities".' A lot enters at 'in some restricted sense'. These are *our* concepts, which took centuries to develop (Gage 1993). In many cultures, and in many contexts in ours, a light and a dark red that differed only via 'sunglasses' would not be seen as qualitatively similar. The authors' description is particularly persuasive in a laboratory where one can continuously attenuate beams, or change wavelength or purity. The concept depends for its persuasiveness (this is rhetoric, not logic) on having the manifold of colours present in practice or in imagination and navigating it in a particular way. Factoring out an intensity dimension is fundamental to radiation-based systems of colour, but also marks their restricted context of applicability. For instance, black is an anomaly in such systems; it has 'undetermined chromaticity'. But in a set of paint samples, black is a colour like any other. Of course all this is why the authors say ' "colour" in some restricted sense'. The restricted sense is precisely that which makes their scheme work. For their purposes it fits the world. But not for all.

They then introduce a 'chromaticity plane' that transects the cone. In the section 'Additional structure . . .' much more comes in. 'When we circumnavigate the boundary of the cone of colours . . . we experience a continuous change of "hue" . . . Since real colours in the interior of the colour cone also have "hues", one wonders about the loci of constant hue in colour space. They must be surfaces that intersect the boundary of the colour cone transversally. Some topological reasoning soon reveals that somewhere in the interior must exist a singular curve of points of undetermined hue.' Here we have formal definitions ('labelling', 'calling' . . .), some natural history of colours ('experience', 'noticing'), implicit assumptions (e.g. of continuity of hue 'in some restricted sense' as one moves inwards from the spectral locus; a very big assumption), and gestures to mathematics ('some topological reasoning'). Again the persuasiveness depends on navigating the manifold in particular ways. It is hard to be clear about the relative roles of the components in this section, and to separate the formal structure from one's understanding of it in terms of colour experience.

In the same section, an 'arbitrary achromatic locus' was introduced. 'Arbitrary' to maintain colorimetry's independence of appearances, although the illuminant is a 'useful' choice since its spectrum then dominates all the other (reflected) lights in a scene. 'All other beams are created by taking away . . . radiant energy from this spectrum.' Colours can then be thought of as spectral 'shadows', a notion to be found in ancient Greek thought and in Goethe. It is surely one of the most appealing and suggestive answers to the question 'What is colour?' The authors forbear from mentioning it, but the arbitrariness of the achromatic point is of course matched by the eye's ability to adapt, to take, to a considerable degree, any prevailing illuminant as achromatic ('discounting the illuminant'). The same phenomenon is alluded to later on, where the option is raised of formally treating the illuminant spectrum as part of 'the eye'. At these points the principled account could get slightly further into the nervous system. As also in the derivation of optic-nerve-like opponent colour functions from considerations of a canonical basis for the projection of beams to colours. I note that both these scarcely get beyond the retina.

We now have a closed curve of colours, a proto-colour-circle. The authors go on to show some interesting things about this circle, and to do that they need to get a better look at it, that is, to get away from the spectral locus, whose colours in the limit are invisible, and find conditions where the gamut of hues is best displayed. For this, they turn to *material* considerations, viz., how colours can be produced with a spectrometer. How does the colour vary with slit width? A very narrow slit doesn't pass enough light to see: it generates black. A very wide slit passes the whole spectrum which recombines to white (if that is the colour of the entrance beam; the arbitrary achromatic point has to be really achromatic for you to follow the exposition in this section). There must be an optimum somewhere in between: a slit that generates the most colourful colours. This occurs for a surprisingly wide slit, with its ends at complementaries (Ostwald, Schrödinger). These colourful colours are therefore known as 'semichromes': made from half of the spectral-locus + purples (Fig. 1.18) (also known as 'full colours' or *Vollfarben*; confusingly, 'optimal colours', defined later, are different).

There is a paradox here. The argument holds for any given spectrometer. But if we have a more powerful light source and use narrower slits, the 'high purity' colours produced are different from less pure ones. We can discriminate purity, at constant luminance, up to the highest value we can make (Wyszecki and Stiles 1967, p. 509), and increasing purity certainly doesn't make colours look less 'colourful'. My memory is that they acquire an intensely coloured, jewel-like quality. However, the authors assert that a spectrum made up of such colours 'really [doesn't] look any better than Newton's impure spectrum though, so the effort is really wasted'. The tone is pragmatic, as also in 'In real life, colourful colours are always nearer to [sic] the semichromes'. That invocation of 'real life' is significant. This is a point where material considerations are important: the spectroscope or Ostwald's most colourful paints. How relevant is it that the colourfulness of a spectral display is a different criterion from that of individual colours? We perhaps shouldn't worry exactly how well the semichromes fit 'the mess' of "the world of colour". As I've already noted, it is good to have a principled exposition.

As the wide slit that produces semichromes is moved across the spectrum (rotated around the achromatic point in Fig. 1.18, a very helpful diagram), one end of it will go beyond the visible region. One end of the spectrum is then passed by the slit, but we could equally well say that the other end is blocked. What happens if one uses from the start, instead of a slit, a stop (a 'complementary slit'), which blocks off part of the spectrum? This, in effect, was Goethe's procedure, and the colours produced are the 'inverted spectrum' which shares yellows, reds, and blues with the Newtonian (slit) spectrum, but now includes the purples but *not* the greens. The complementary geometry of the slit/stop generates complementary colours. As far as generating colours is concerned, Newton's monochromatic lights were nothing special; Goethe's 'boundary colours' will do just as well. (The authors do a fine job of historical rehabilitation on both Goethe and Ostwald.)

The story of the colour circle continues into what is perhaps the most original contribution of the chapter: the development of the surface-colour solid and the mensuration of the colour circle, but I stop here for reasons of space. So, what have we learnt so far about the relation of spectrum to colour circle? We learnt how we get a cone of colours by projecting a manifold of beams; we learnt where in the chromaticity diagram to look for the best colours, and how to produce them; and any residual impression that Newton's spectrum was primary was removed by the slit/stop complementarity. But note that the crucial substrate of appearances was brought in from outside; just taken as given. It was then woven into the development. The situation is not that of footnotes in a mathematics text mentioning applications, where the body of the text is an independent formal argument. This, of course, is just repeating that the chapter is fortified colorimetry, which the authors would not deny. What I've tried to do is to pinpoint some of the added nutrients.

### What's left out?

When I first read this chapter, I was struck by how many of the topics of colour science were touched on, even though some, like adaptation, are implicit rather than explicit, and others enter surreptitiously. I found it an effort to write a list of topics that were completely excluded.

But I can also see it in quite a different perspective, in which I agree with the author's disclaimers about its relevance to colour at large. Look at it this way. The chapter is about trichromacy. But is trichromacy any more relevant to most of what we do with colour than is the writing system to the content of a book? Anyone who has transcribed a spoken tape knows that much of what went on—prosody, hesitations, changes of tone and so on—is lost in transcription. Similarly, our photopigments filter out much of the spectral variety of the world. But from most points of view they are both just interfaces. They explain very little indeed of what we do with colour or what is written in books.

## References

**Gage, J.** (1993). *Colour and culture.* Thames & Hudson, London.

**Hilbert, D. and Cohn-Vossen, S.** (1932). *Geometry and the imagination.* English translation 1952. Chelsea, New York.

**Johnston, S. F.** (2001). *A history of light and colour measurement.* Institute of Physics Publishing, Bristol.

**Wyszecki, G. and Stiles, W. S.** (1967). *Color science* (1st edn). Wiley, New York.