A NEW INSTRUMENT FOR COLOUR MEASUREMENT

by

ANDREAS POLYDOROU

Thesis submitted in fulfilment of the requirements of the City University for the award of the Doctor of Philosophy Degree.

City University, 1978

CONTENTS

CHAPTER Page No.

Page No.

List of Tables

List of Illustrations: Page No.

DECLARATION

I hereby declare that for a period of two years from submission of this thesis, it should not be copied in whole or in part.

At the end of the two year period, I grant powers of discretion to the University Librarian to allow the thesis to be copied in whole or in part without further reference to the author.

A. Polydorou

ABSTRACT

This thesis presents an exposition of colour fundamentals and their application to the development of a new instrument for colour measurement.

Sensation and perception of colour vision are considered both from the physiological and the psychological aspects. The physical aspects of colour stimuli is also considered. A number of industrial, commercial and research applications of colour is discussed.

The historical development of colour vision theories, together with the laws of colorimetry and colour matching, are given particular emphasis. The mathematics of colorimetric transformations is described in some detail as a prelude to the definition of the standard observer and the C.I.E. system of colour specification. The transformations to obtain all-positive distribution coefficients is also described in some detail.

Methods of colour measurement and specification discussed, include naming, colour atlases, visual colorimeters, photoelectric colorimeters and spectrophotometers. The operation of specific types of subtractive and additive colorimeters are described. The disadvantages of existing colorimeters and spectrophotometer-computer combinations are outlined. The advantages of the new instrument over the existing systems provide enough justification

for its introduction. A number of applications where the new colorimeter has obvious advantages is described.

The description of the colorimeter is considered by following through each major component as the design evolved. Some problems that arose during the development stages are analysed. Photographs of various sections of the colorimeter are included.

Finally, the performance of the spectrocolorimeter is discussed. Its precision and accuracy are firstly considered from the spectrophotometric aspect and then from the colorimetric aspect. This is done through the evaluation of the errors normally associated with the calibration of an insturment - errors in zero, range sensitivity and non-linearity.

An extensive coverage of the properties of solid state photoelectronic devices is given in an inset in the form of a monograph. In this monograph, the electrical and optical properties of semiconductors are covered extensively. Noise, amplification, speed of response and other limiting parameters are examined and then their ultimate sensitivity compared. A number of practical applications of photoelectronic devices is also included.

INTRODUCTION

A few years ago the author approached Dr Padgham of City University, with the specific request of City University assisting the author in the design of an on-line colorimeter involving no moving parts. During the discussion that followed Dr Padgham suggested that this could be done through a Ph.D. project.

The author had already by that time designed the Refined Oil Colorimeter and was very heavily involved in other industrial colour measurement and control problems. The combination of this background together with the more theoretical approach of the University appeared to offer an ideal situation for the development of a simple, useful and technologically advanced instrument.

The principle of the instrument was unanimously decided within a few minutes of the first meeting.

Nearly all the work for this instrument was done in the author's home during his spare time. During the earlier stages of the project, the photocell problem appeared to be so accute that the project was nearly abandoned. A thorough investigation into photocell principles and performance, however, showed that the new (at that time) silicon photodiode provided the ideal answer.

A second instrument, identical to the first, but using only twelve photocells is now being used industrially for the control of the colour of ice cream. A number of other new applications are now under consideration.

Apart from Dr Padgham and the City University, the author would like to express his acknowledgements and thanks to his colleagues who in many ways have helped to improve the design of this instrument. Acknowledgements are also due to the authors and publishers of the three books from which diagrams have been abstracted: W.D. Wright (1969), Judd and Wyszeski (1963) and Padgham and Saunders (1975).

The author would like to dedicate this thesis to [redacted] who for such a long time had to endure the painstaking task of accepting this amount of work to be done at home.

Also to his father to whom the laws of nature are more obvious than the alphabet, and to his mother whose ability to memorise and remember seem to have no bounds.

CHAPTER 1

PURPOSE OF COLOUR MEASUREMENT

1.1 Equality of Perception (ie matching) not a Measure of Perception

Vision can be defined as the function in which the eye acting jointly with the brain react to radiant energy or light. The action of light (radiant energy in the 380 nm to 760 nm wavelength region) on the retina, the photosensitive part of the eye, forms the first stage of the vision process. Refractive surfaces in front of the retina cause the formation of images on the retina.

Due to the ability of the visual system to adapt to levels of illumination, to differentiate between the existing levels in an instantaneous field of view, to appreciate differences involving small spatial dimensions and temporal features, man can begin to construct a world of sensory objects - visual perception.

Due to the ability of the visual system to differentiate between radiant energies of dissimilar wavelength, man can begin to construct the world of beauty and aesthetic appreciation - colour perception.

Even though colour is enjoyed by everybody, from the scientific point of view it is not an easy subject because colour is fundamentally an optical sensation which combines physiological, psychophysiological, psychological, physical and chemical effects.

Quite often, the study of colour is more of an art rather than a science.

In the English language there is a total of 3500 words which describe colour. The average human eye can, however, distinguish the differences between several million colours and shades. It is thus obvious that a numerical system of colour specification is needed to define adequately the visual quality of colours.

Sensation is the conscious experience that follows immediately upon the stimulation of a sense organ or a sensory nerve. Perception is a sensation that is associated with a meaning. This distinction between sensation and perception is extremely important in the study of how we perceive the world. Fig. 1.1 shows a test which helps to clarify this conceptual discrimination.

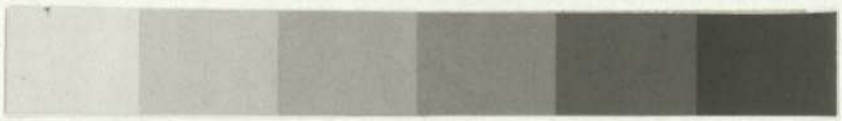


Fig. 1.1 Six shades of grey sensed as uniform but perceived as getting lighter from left to right.

Six shades of grey are shown, each one of which is absolutely uniform. However, each patch of grey appears to get lighter from left to right. This is because of contrast with the adjacent patches. By covering up the adjacent patches with a white paper,

it is possible to see how uniform each patch really is. In this test, the stimulus exciting the eye sensors when looking at any one of the grey patches is exactly the same whether the adjacent patches are of equal brightness (as when covering them with a white sheet of paper) or of unequal brightness. The sensation therefore as detected by the eye is the same in both cases. The perception of the image in relation to its surroundings, however, which is the interpretation by the eye and brain acting in accord, is quite different in the two cases.

Physiological colour sensation results from radiations of wavelengths ranging from approximately 380 to 760 nm entering the eye. In the context of colour measurement, radiation giving rise to colour sensation is termed colour stimulus.

It is not the primary aim of colour measurement to characterise colour perception values. Since colour discrimination depends on physiological and psychological factors (adaptation phenomena, contrast, etc) the perception, which cannot be measured directly, must be substituted by the separate concept of colour sensation for purposes of colour measurement.

The eye evaluates colour stimuli in a certain manner, and the result of this evaluation, viz. the colour sensation, can be uniquely expressed numerically. The object of colour measurement may, therefore, be defined as the determination of colour sensation by numerical coefficients. In practice colour sensation is uniquely

defined by three values, the tristimulus values. The object of colour measurement is accordingly the determination of the three tristimulus values relating to any given colour.

Where a conceptual distinction between colour and colour sensation is immaterial, the term colour may be used without hesitation, even if, in the strict sense of the word, colour sensation is meant. This linguistic simplification facilitates understanding.

Since the three tristimulus values can be shown geometrically in a three-dimensional co-ordinate system as the co-ordinates of one point corresponding to colour sensation, it is understandable that agreement is necessary in regard to the co-ordinate system to be employed. When applied to colour measurement, the choice of a co-ordinate system corresponds to the choice of three primary colours. The co-ordinate system chosen by C.I.E. for specifying colour sensation is known as the C.I.E. system.

Tristimulus values relative to this system are termed C.I.E. tristimulus values.

The natural colour vision of different human beings is not quite identical - even allowing for defective colour vision. Hence, for the determination of unique tristimulus values, an agreement about the average vision of a normal observer is necessary. This factor is incorporated in the C.I.E. system.

The relationship between colour stimulus and colour sensation and the relations between colour sensations among themselves, which form the basis of colour measurements, are the subject of colorimetry. (The term - metry is derived from the analogy between colour sensations and points in a space which may be demonstrated geometrically.)

1.2 Physical Aspects of Colour Stimuli

Before entering into the complex problems of colorimetry, the physical factors characteristic of colour stimuli must be summarised. In terms of energy, the colour stimulus, ie the intensity of the radiation entering the eye, is completely and uniquely defined by the radiance. The radiant flux (radiation energy per unit time), emitted by a given area is observed and divided by this area and by the size of the solid angle operative in the system concerned. The unit of measurement of radiant flux is, therefore, watts/sq cm/steradian. If, as with colour stimuli, the spectral distribution of the radiation is of interest, the radiant flux is differentiated with respect to the wavelength, resulting in spectral radiant flux, whose unit of measurement is watt/sq cm/steradian/nm when the wavelength is measured in nm. In colorimetry, the relative spectral distribution of the radiation is in many cases the only factor of interest; termed Spectral Energy Distribution S_λ. It is immaterial whether radiance, the radiant flux, or any other energy factor is being considered, since the spectral energy distribution is a purely numerical series, which is

only important as a function of the wavelength. If the radiation from a light source strikes the eye directly, the spectral energy distribution S_λ equals the colour stimulus φ_λ (relative spectral distribution of the colour stimulus). In the most important colour measurements, however, a material substance is irradiated by a light source and alters the spectral energy distribution of this source according to its absorptive properties. A colour filter interposed between an incandescent light source and the eye, absorbs light to a degree varying with wavelength, so that the apparent white colour of the incandescent bulb is converted to the colour of the filter. Similar conditions prevail for the diffuse reflection of light at the surface of opaque substances. This is known as diffuse reflectance. Diffusely reflecting substances act as secondary light sources. The absorptivity of the substance in terms of wavelength is defined physically in the known manner by the spectral transmittance τ_λ for clear, transparent substances and the degree of directional spectral reflectance β_λ for diffusely reflecting substances. With the aid of these factors, the spectral energy distribution of the light source S_λ can be linked to the colour stimulus: φ_λ ,

$$\varphi_\lambda = \tau_\lambda \cdot S_\lambda \quad \text{for transparent substances}$$

and

$$\varphi_\lambda = \beta_\lambda \cdot S_\lambda \quad \text{for diffusely reflecting substances.}$$

These interrelations are important for the definition of object colours.

CHAPTER 2

APPLICATIONS IN INDUSTRY, COMMERCE AND RESEARCH

The need of means of colour measurement has always been in limited demand. The last 50 years have, however, seen a rapid increase in the number of applications where colour measurement has become necessary rather than just desirable. The improved performance of available instrumentation, coupled with international agreements on systems of colour specification has triggered a chain reaction of colour measurement applications in Industry, Commerce and Research.

The demand has changed from simple matching methods and colorimeters which were inaccurate, very slow and cumbersome to operate, to very accurate, rapid and fully automated systems. Recent technological advancements in data processing have seen the introduction of computers for almost instantaneous transformation of instrument readings to any other required system of colour specification.

Due to the wide varied assortment of colour measurement applications that exists, a logical classification is extremely difficult. For this thesis, it has been decided to classify these applications into Industrial, Commercial and Scientific (Research).

2.1 Commercial and Industrial

No manufacturer can afford to overlook the fact that the act of purchasing is an emotional experience. Numerous marketing

investigations have conclusively indicated that colour plays a dominant influence on product preference as judged by consumers (J P Favre, 1969 and L Cheskin, 1951). In some cases the tolerance on the acceptable colour range is so critical that a product may be completely discarded even for slight deviations from the preferred colour (J Pilditch, 1961). In other cases products may be graded and priced according to their colours (F Birren, 1962).

A mass produced item (such as a car or a refrigerator) assembled from a number of components, must not show any distinguishable colour differences between the various components that are supposed to be identical in colour when they are brought together on an assembly line. Furthermore, replacing a part produced at a different time, such as replacing a dented car door, should readily be acceptable by the car owner. To achieve such close tolerances, it has been estimated that for the master panels, against which the mass produced panels are compared, the chromaticity co-ordinates must be held constant (W D Wright, 1959) to within ± 0.005 in x or y for near neutral colours and to within 0.005 in luminance factor.

The fascinating aspects of colour can furthermore be used to exploit new industrial and consumer fields in product presentation (packaging systems). A big step towards this aim can be achieved if artistic manipulation (the method used at the moment to improve product presentation), were supplemented with scientific analysis and with the recent technological advancements, especially in the field of data processing.

Colour is very widely used in Commerce and Industry. The following selected examples demonstrate some important applications of colour in commerce and industry.

2.1.1 Colour of Edible and Chemical Oils (Fatty Acids)

The colour of these oils determines their price. A premium price is always paid for light colour oils. Oil colour is normally specified by the use of the Lovibond system. It is usual to leave out the blue and neutral density values and just specify the length of the sample cell (normally one inch or five and a quarter inches) and the yellow and red values obtained on the Lovibond Tintometer.

This is a functional use of colour specification in that colour indicates the degree of purity of the oils. This is also true for the edible oils, where the best quality oils are of very low colour content (nearly colourless). This, until recently, has been unacceptable to a large number of consumers who expect to see a light yellow oil which bears comparison with butter. Thus a dye (normally beta carrotene) is sometimes added to make the colour of oils more acceptable to the consumer.

The recent trend, where calorie consumption is being much scrutinised by an ever increasing number of people, has made the lighter oils much more acceptable.

2.1.2 The Colour of Ice Cream

It is not perhaps obvious that the colour of ice cream can influence its sales. Yet this has been confirmed in a number of investigations. The author has been involved in the following two investigations:

(a) Ten samples of vanilla ice cream with colours ranging from white to a deep yellow (but all with the same flavour) were presented to a panel of 270 people, each of whom tasted three of the ten products. The differences in colour were not drawn to their attention. They were simply asked to rank the three samples of vanilla ice cream in order of preference. Only a fraction of the people concerned classified two of the three samples as having identical flavour. This shows that consumer preferences is biased more by the colour of the ice cream than by its flavour. This investigation eventually led to the more preferred colour of vanilla ice cream. It also gave indications on the range of colour acceptability.

(b) This investigation was specifically designed to find the range of colour acceptability for particular types of ice cream. Two groups of people were shown the same sample of ice cream and asked to select from a colour atlas (the I.C.I. colour Atlas) the colour chips that they thought were either an exact match or a near match to the sample. The first group was told what type of ice cream the sample represented. The other group was not.

The two sets of results were significantly different. The judgment of the people in the first group was biased by their opinion of the ideal colour for that particular type of ice cream.

In another investigation on ice cream, simultaneously conducted in three countries (UK, Germany and Italy), it was convincingly demonstrated that the most acceptable colour was not the same in the three countries. Thus in the UK a "yellow" colour was preferred for vanilla - which shows a connection with butter - whilst in Italy a brilliant white was preferred - an indication of purity. The range of acceptability was also different for the three countries: the UK consumer being more influenced by colour variations than by the Italians or the Germans.

These investigations show how important it is that the colour of ice cream is controlled within specified limits. This led to extensive work by the author in this direction which resulted in a computerised system using the present spectrocolorimeter.

2.1.3 Product Presentation - Colour of Packages

This is a very evident application of colour, which is the most important element of a package. A package should not only be used to cover the product and give information to the customer about the product. By presenting the product in a more striking and attractive way, the package can express something of the quality of

the product and provide an "image" or "personality" for it. It is important that a package attracts immediate attention and recognition on a supermarket shelf.

Our senses are very susceptible to the effect of colour. Colour can stamp itself better than any other factor in our memory and makes the package more easily recognisable.

The following examples (J P Favre, 1969) demonstrate the importance of colour in packages:

(a) For many years Woadbury toilet soap, whose colour was ivory, was sold in a green wrapper. The sales showed a sudden increase as soon as the same soap was presented in a blue wrapper.

(b) The American cigarette Marlboro, which was packed in a white packet, showed a constant decline in the sales figures and it was thought that the white colour of the packet went out of favour with the consumers. A quarter of a million dollars was spent by a team of specialists to redesign the packet. The new packet was still mainly white (to give the impression of quality and purity) with a red angular design. The colour and the angular design were introduced to give the product a dynamic and masculine character. The sales figures showed a sudden improvement soon after the new packet was introduced.

(c) The American firm that markets the electric lamps under the trade name of "Sylvania" decided to change the standard colour of the packets containing the lamps to a range of packets with different colours where each colour identified the lamp wattage. An immediate increase of 66% was effected in the turn-over.

The aspect of package colour design has been of particular interest to the author who suggested two optical systems that can assist the designer in this difficult job.

The first system comprises six diffraction gratings producing six spectra and two fields of view adjacent to each other, each field comprising three superimposed spectra. Thus each field of view can be made to any colour the operator likes. A side by side comparison of the two fields can thus be effected. It would also be possible to alter the relative areas of the two fields. This system can be used for rapid evaluation of optical effects such as simultaneous contrast and harmony of two colours.

The two values corresponding to the positions of the diffraction gratings (determining the wavelength bands used) and of the aperture openings (determining the brightness) can be fed into a computer. The computer can then calculate the amount and type of dyes or inks to be used to match the colours obtained on the screen.

The second system comprises a colour television (Colour Visual Display Unit) together with a computer. It is possible to design

any complicated pattern by the use of the computer and display it on the television screen.

Furthermore, the colours of each section of the pattern can also be varied. Thus the desired design can be eventually obtained on the screen. Figures pertaining to the colours displayed can be fed into a second program which can work out the colour of the inks to be used on the package.

A similar system has been designed by the Digital Equipment Corporation to enable the design of patterns in the textile industry.

2.1.4 Colour and Lighting Conditions in a Production Environment

It is a well known fact, that high lighting levels are not sufficient to ensure the most efficient form of lighting conditions. There should be a balance of spectral emission in artificial light sources.

It has been convincingly shown (Aston et al, 1969) that balanced light sources can improve visual clarity. The warmth of such sources can flatter human complexion, thus aiding employee contentment.

Experiments in a number of Zoos have demonstrated that a balance source can reduce disease, prolong life and encourage mating (J Ott, 1965).

Some plants and some lower animals have also been found not to thrive under the standard cool white fluorescent tubes used almost exclusively in industry and offices. This source is deficient in red radiation (R J Wurtman, 1968).

In spite of the advantages of balanced sources, it is important to avoid monotony of illumination, brightness and colour. Such monotony can cause effects similar to those of the drug LSD.

Different colours can have different psychophysiological effects (F Birren, 1969).

(a) Exposure to red increases the heart action and thus the body temperature rises.

(b) Red-orange light improves the growth of young animals and has been used for the treatment of eczema and menstrual disorders in humans.

(c) Blue light has been used to treat premature infants in which an excess of bilirubin was found in the bloodstream.

On human performance, it has been said that, "A specific colour stimulation is accompanied by a specific response pattern of the entire organism".

Red is inciting to activity. This rise in activity is, however, followed by reactions that fade below normal.

Blue tends to relax the nervous system. This relaxation is followed by a rise in activity above normal.

It would possibly be wise to look into the possibility of illuminating such areas as canteens, rest rooms, toilets etc with blue light. Illuminate working floor areas with balanced light sources with red light introduced a few times a day for short periods (especially before a break) to relieve the monotony and stimulate activity.

2.1.5 General

It can be said that a listing of all applications involving colour in commerce and industry is inexhaustible. The paint industry and the dyeing industry of course lead the field for the necessity of colour measurement and control. The photographic and television industries lead the field in colour reproduction. The fabrics industry has always been associated with a vast number of developments in colour technology. Similarly with the ceramics industry. Enough has been said about the food industry.

To a large extent it is true to say that no industry exists which does not depend on colour science and technology in one way or another.

2.2 Applications of Colour in Research

One of the most common applications is in Chemical tests, where many chemical reactions can be followed by the colour changes that accompany them. The products of the reaction can be quantitatively determined by colour measurement. Such tests can be used for pH measurements, testing the quality of milk, the estimation of sulphanilamide in biological fluids, the estimation of toxic gases in air, and so on.

Indicators that change in colour depending on the hydrogen-ion concentration have been developed (King, 1952).

Colour is also becoming of increasing importance in crystallography.

In structural engineering and in other allied technologies it is normal to make a model of a structure out of bakelite or celluloid and subject it to experimental stresses corresponding to those to be met in practice. The analysis of these stresses is carried out by the photo-elastic effect. Stresses cause the celluloid to become anisotropic and thus a plane polarised beam of white light passing through the sample will suffer optical retardation. Thus the light will in general emerge elliptically polarised and the fraction of

the light transmitted will vary through the spectrum, so that the emergent beam will be coloured. The locus of a particular colour in the model will be a locus of constant stress. This locus provides the stress distribution in the model and comparison of the colours at different points in the model relate the magnitude of the stresses at these points. These colours have been analysed on the chromaticity chart (Band & Wright, 1930).

This analysis provides useful information to the design engineer. It has recently been used for the design of aeroplanes. Model aeroplanes made of celluloid were placed in a wind tunnel and the stresses investigated for different wind velocities (New Scientist, March 75).

In biological studies, optical effects such as scattering and interference, have been used to investigate colour phenomena associated with certain feathers (eg of the Macaw, the humming-bird and the peacock). These studies have resulted in a classification of these feathers in terms of colour.

In meteorology, the study of the visibility and recognition of a distant object is of great importance. This is because the colour of an object as seen from a distance varies, depending not only on the spectral characteristics of the object itself but also on that of the light incident on its surface, the atmospheric absorption and scattering between the object and the observer. It has been shown (Middleton, 1941) that with increasing distance the colours first

shift towards the blue corner of the chromaticity diagram, and then reverse their direction, becoming brighter and shifting towards the point corresponding to illuminant C. Middleton worked out the theoretical considerations as postulated by the Rayleigh formula and an assumed extinction coefficient for slightly hazy air, and converted the obtained values to chromaticity co-ordinates and then compared them with the actual values as obtained by an observer. Good correlation was obtained.

Wright (Wright, 1969) designed a special portable colorimeter to investigate the colours observed during displays of the aurora. As the radiations emitted from an aurora come mainly from three parts of the spectrum (430-470 nm, 577.7 nm, 630-640 nm), trichromatic colorimetry is ideally suited for this application. By locating the chromaticity of an aurora on the C.I.E. diagram, it is thus possible to determine its complete spectral composition.

The same colorimeter was also used to observe the colour of the sky during the total eclipse of the 30th June 1954.

A slightly improved version of the above colorimeter was used for further astronomical observations. Wright (Wright, 1969) used this colorimeter in Sweden, near the centre line of the totality zone, to record the sky colour in an azimuth plane 90° round from the sun and at a height of 40° above the horizon. Due to the cloudy sky, however, the results were of little astronomical value.

Wright (1969) suggested that, "during an eclipse, where changes of luminance and colour occur very rapidly and extend over large areas, a comprehensive investigation might be planned in which visual observations taken with a number of colorimeters pointing to various parts of the sky were combined with continuous photographic records covering large areas. The colorimetric observations could then be used to calibrate the colours of the corresponding points on the photographs, and thus permit interpolation in both time and space to give a much more complete and accurate record than would otherwise be possible by using colorimetry or photography alone".

In the microscopic investigation of minute sections or corpuscles of a specimen, the effect of dichroism has to be taken into account, as the colour of the corpuscles (for example of blood) might bear little or no resemblance to a conglomeration of such corpuscles. Blood corpuscles, for example, are light yellow in colour. Blood in transmission, depending on the size of the sample, can have a dominant wavelength of 575 nm (in a 0.01 mm cell) to 618 nm (in a 10 mm cell).

Dichroism is also very important in the production of dyes, where usually it is an undesirable effect.

Vision is a very important area where colour measurement can be of very much interest. A large number of researchers have demonstrated the wide variety of such experiments that led the way towards a more thorough understanding of such a complicated function

as that of vision. A number of these experiments and their findings are reported elsewhere within this thesis - Newton, Young, Maxwell, Helmholtz, Abney, Grassmann, Guild, Wright, Judd etc.

Padgham (Padgham, 1968) used the C.I.E. system to investigate the effects of after-images, and pioneered in the measurements of their intensity and colour. Using a specially designed colorimeter, the colour of an induced after image was measured by matching its appearance against the colorimeter field seen in the right eye. He found that inducing the left eye with an illuminant C stimulus, the locus of the after image in the C.I.E. chromaticity diagram followed a spiral, so that its dominant wavelength changed successively from blue-green to green, orange, yellow, red, magenta, purple, violet and then blue.

D Jameson and L M Hurwich (Jameson and Hurwich, 1961) have reported experiments in which a colour-matching technique was used to compare the chromatic responses to focal stimuli seen first in isolation and then in the presence of surrounding stimuli of specified luminances and chromaticities and of various degrees of complexity. The results were analysed in terms of the chromatic response processes of the opponent colours theory. The concept of physiologically based, opponent chromatic induction is fundamental to the opponent colours theory. A continued quantitative development of this theory to account for colour perceptions, equations and discriminations under various conditions of adaptation and illumination has emphasised the need for systematic

quantification of induced colour effects. The results (using a total of forty observers) have shown that chromatic inductions decrease systematically with decreasing contiguity of focal and surround stimulus areas. For given degrees of contiguity, induced chromatic responses are shown to be opponent to, but proportional in, magnitude to the mean chromatic activities of the inducing field, and the constant of proportionality decreases as a function of decreasing contiguity.

Rowe (Rowe, 1973) also studied the effect of surround luminance on perceived colour. He found that a dark surround reduces the perceived saturation of a self-luminous colour by 65% relative to a light surround. He found no consistent changes in the hue between the two surrounds.

In trying to establish a subjective scaling of hue and saturation in the same thesis, Rowe concluded that "psychophysical estimates of colour appearance are possible by most, if not all, observers under widely varying conditions. The results would indicate that for the hue continuum most observers respond in a similar way to a given stimulus, whilst for saturation a theory of individual modes of response must be put forward. In either case the spread of results for a given observer and stimulus is remarkably small."

CHAPTER 3

COLOUR VISION THEORIES

3.1 Historical Development of Colour Vision Theories

The attempts of the Greek and Roman philosophers to explain the ideas associated with colour were very incomplete. The older writers name only white, black, yellow and red as distinct colours. The Egyptians classified blue and green as one form of black.

The development of new pigments until the middle ages was very slow indeed. Until then, black (powdered charcoal), blue (copper-sodium silicate) yellow and red (from natural ochres) and indigo, formed the basis of all coloured works of art.

Organic pigments were introduced with the rise of chemistry in the 18th century. Even at that time (Ghoete, 1810) the distinction between "colour" and "pigments" was not clearly understood.

Leonardo da Vinci wrote (Leonardo da Vinci, 1877) "And you, painter, who are desirous of great practice, understand that if you do not rest it on the good foundation of Nature, you will labour with little honour and less profit; and if you do it on a good ground, your works will be many and good, to your great honour and advantage. A painter ought to study universal Nature, and reason much within himself on all he sees, making use of the most excellent parts that compose the species of every object before him .....

Then, again, the study of the arrangement of colour of natural objects is almost entirely ignored."

And yet, even the genius of da Vinci, in spite of his observations and advice on colour, gave him no essential ideas in this familiar domain.

The study of colour vision involves physical, physiological and psychological phenomena. The theories of colour vision became more complicated as the knowledge of Physics, Physiology and Psychology developed through the eighteenth and nineteenth centuries. An acceptable theory of colour vision should withstand a critical analysis of all the effects associated with the physiological process of vision, colour vision and the nature of perception.

The first real advance in the understanding of colour was achieved by Newton following his celebrated experiment in which he split up a white beam of light into its colour components by refraction through a prism (Newton, 1704).

Even though purple and magenta were not found in his spectrum (since they are not spectral colours), Newton added these colours which were familiar to painters, to the spectrum and produced his famous Colour Circle. By doing this, Newton showed his lack of understanding of the conceptual distinction between colour and pigments. In the Colour Circle each hue is connected with its neighbours by continuous transitions. Thus the colour circle, as

pointed out by Newton, contained an indefinite number of hues, which naturally fell into a small number of groups. Newton distinguished seven such groups - red, orange, yellow, green, blue, indigo, violet.

The first person to utilise the Newton circle in practice was Le Blond (Frankfurt, 1730) who prepared colour prints for which, in the first place, he employed Newton's seven colours. Soon after this, he discovered that he could achieve this by the use of only three colours: Red, Yellow and Blue. Gautier (Paris 1730) arrived at the same conclusion at about the same time. Thus were laid the principles of the trichromatic origin of colour.

If Newton had, however, studied more deeply the geometrical basis of his Colour Circle and his suggested method for calculating the effects of colour mixtures, he would have realised that his method was inherently based on a trichromatic idea.

Newton conceived the colours of the spectrum (plus magenta and purple) to lie in the circumference of the circle, "and the circle so painted that every radius exhibits a gradation of colour, from some pure colour of the spectrum at the circumference, to neutral tint at the centre. The resultant of any mixture of colours is then found by placing at the points corresponding to these colours, weights proportional to their intensities; then the resultant colour will be found at the centre of gravity, and its intensity will be the sum of the intensities of the components".

It was left up to Maxwell (Maxwell, 1860) to notice that "from the mathematical development of the theory of Newton's diagram, it appears that if the positions of any three colours be assumed on the diagram, and certain intensities of these adopted as units, then the position of every other colour may be laid down from its observed relation to these three. Hence Newton's assumption that the colours of the spectrum are disposed in a certain manner in the circumference of a circle, unless confirmed by experiment, must be regarded as merely a rough conjecture, intended as an illustration of his method, but not asserted as mathematically exact. From the results of the present investigation, it appears that the colours of the spectrum, as laid down according to Newton's method from actual observation lie, not in the circumference of a circle but in the periphery of a triangle, showing that all the colours of the spectrum may be chromatically represented by three which form the angles of this triangle". Thus, was enunciated the famous Maxwell triangle.

Maxwell went on to determine the spectrum locus on a colour triangle, by the use of an instrument consisting of a chamber of about five feet long, nine inches broad and four inches deep, jointed to another chamber two feet long at an angle of about 100°, and investigating the white light reflected from a white sheet of paper illuminated by sunlight in the open air.

Maxwell then deduced mixture curves of the spectrum for himself and his assistant, using as primaries spectral radiations of wavelengths 630.7 nm, 528.6 nm and 457.3 nm.

The results indicate that the spectrum that Maxwell used was very impure, particularly in the blue-green part of the spectrum. This gave an entirely fictitious approximation of the spectrum locus to two sides of the colour triangle and led Maxwell to conclude wrongly that " .... all the colours of the spectrum may be chromatically represented by three which form the angles of this triangle", and that there was "strong reason to believe that these are the three primary colours corresponding to these modes of sensation in the organ of vision".

Even though Le Blond and Gautier used the trichromatic principle of colour in 1730, it was not until 1802 (Young, 1802) when the versatile genius of Thomas Young laid down the foundations of the true Trichromatic Theory of Colour. Young suggested the existence of three primary sensations of colour which are excited in different proportions when different kinds of light enter the organ of vision. According to this theory, the threefold character of colour, as perceived by us, is due, not to a threefold composition of light, but to the constitution of the visual apparatus which renders it capable of being affected in three different ways, the relative amount of each sensation being determined by the nature of the incident light. If it were possible to isolate the three colours corresponding to the three primary sensations as described by Judd

(Judd, 1950) then each of these colours would excite one and only one of these sensations. All other colours whatever must excite more than one primary sensation. The three sensations postulated in Young's Theory are as shown in Fig. 3.1.

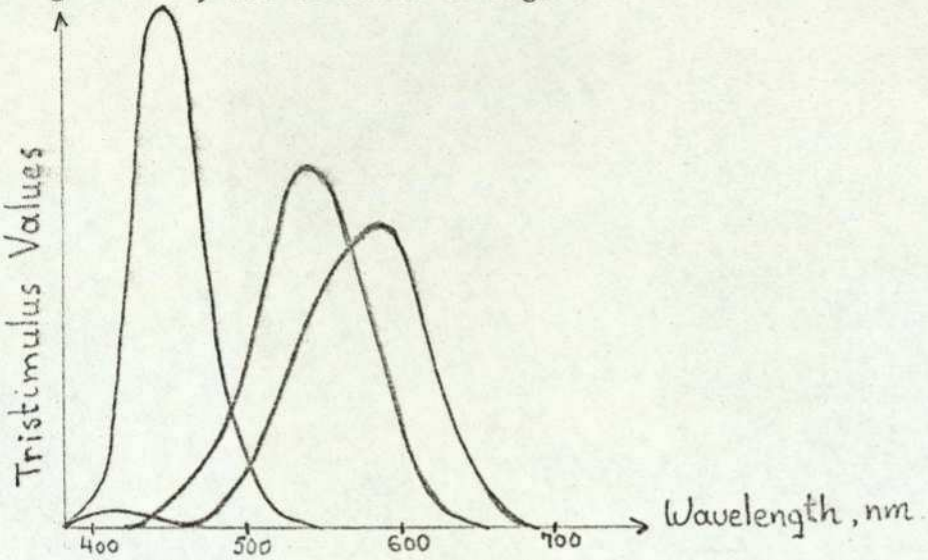


Fig. 3.1 Relative spectral-response functions postulated in Young's Trichromatic Theory.

Even though dichromatic vision could in principle be explained by Young's Trichromatic Theory, the response functions (sensations) postulated cannot in practice account for dichromatic vision. The primaries, in Young's theory, correspond to an extreme spectrum red, an extreme spectrum violet and an imaginary green, that is, a green located outside the gamut of real colours. Fig. 3.1 shows the response functions derived from such primaries.

Helmholtz fifty years later modified Young's three component theory to account for dichromatic vision. Helmholtz proposed that dichromatic vision could be explained if one of the three sensations (sensors) was suppressed. Thus in the case of protanopia, the

"red" process is missing, for deuteranopia the "green" process is missing and for tritanopia the "violet" process is missing.

To account for dichromatic vision, Helmholtz carefully chose the primaries to correspond to those missing in dichromatic vision. The response functions resulting from such primaries are shown in Fig. 3.2.

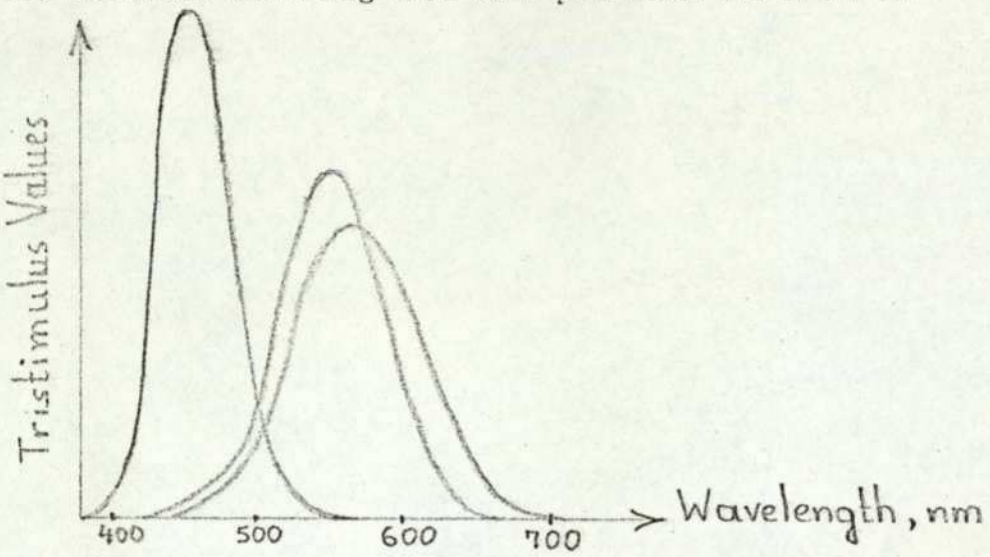


Fig. 3.2 Relative response functions postulated in Helmholtz's Theory.

The Young-Helmholtz theory explains the main facts of colour vision: after images can be explained by assuming fatigue of one or more of the processes in varying degrees. For example, after prolonged exposure of the eye to green light (thus fatiguing the eye to green light) a white surface appears an unsaturated purple-pink. Helmholtz explained the problem of simultaneous contrast by assuming that this was the result of "false judgment". While it may be purely psychological, it appears probable to some that it is actually physiological in nature, one part of the retina being influenced by stimulation of another region.

The Young-Helmholtz Theory, however, is completely unable to account for total colour-blindness, as no mechanism is postulated that can sense all colours equally.

Detailed analysis of Helmholtz's theory predicts that protanopes should see all colours in shades of green and violet; deuteranopes should see all colours in shades of red and violet. However, experimental evidence provided by observers with one normal and one dichromatic eye shows that protanopes and deuteranopes see all colours in shades of blue and yellow.

Thus the Young-Helmholtz theory, even though able to account for most colour phenomena, cannot be accepted unconditionally, as it fails to explain colour perceptions of protanopes and deuteranopes and also complete colour blindness.

A number of variations have been suggested (Fick 1879, Konig 1884 and 1892 and others) to the Young-Helmholtz theory to correct its failures, but each of them has produced new weaknesses. Other attempts to reconcile this theory with all the observed facts have resulted in highly complex models that had eventually to be abandoned.

The Duplicity Theory, which attempts to differentiate colourless and colour vision, was first suggested by Schultze in 1866, and later on advanced by Von Kries and Muller. It is based upon anotomical evidence of the existence of "rods" and "cones" in the retina. The

former were assumed to be responsible for achromatic sensations and the latter for both achromatic and chromatic sensations. The rod action is chiefly associated with twilight illumination, whilst the cones are involved at higher levels of illumination. Examination of the retina shows that the cones are situated in the centre of the retina (fovea centralis); the rods predominate in the outer zones of the retina.

The Duplicity theory explains the following facts (perhaps because it was chiefly built up from these facts):

(a) Colourless vision over the whole retina in dim light.
(b) The decreased sensitivity of the fovea in twilight.
(c) The Purkinje effect.
(d) The absence of a "Purkinje shift" for foveal vision.
(e) The absence of achromatic threshold for any light for foveal vision.
(f) The absence of achromatic threshold for red light for any region of the retina.
(g) Colourless vision over the whole retina in the case of the totally colour blind.

The supporting evidence in general is represented by more dependable and convincing data than in the case of any theory of colour vision. The Duplicity Theory does not attempt to explain colour vision, but is of interest here because of the attempt to separate vision into chromatic and achromatic processes.

A number of experimenters have tried to confirm the existence of three types of cones, as postulated by Young.

Measurements of the spectral absorption of individual cones in a human retina kept active for a few hours after death, using the microspectrophotometer (Mark et al 1964, Brown et al 1964), tend to confirm completely Young's postulate of the three types of colour receptor.

Hering (Hering, 1878) abandoned physical analysis in favour of a psycho-physical analysis to explain colour vision. Instead of arranging colours according to their wavelengths, the physiologist Hering arranged them according to those similarities and contrasts which are directly felt. He assumed the existence of six fundamental sensations coupled in pairs namely, white and black, red and green, yellow and blue.

Hering, while emphasising its hypothetical character, introduced a physiological assumption, that each pair of sensations is associated with a substance somewhere in the retinocerebral apparatus. Each one of these substances is capable of building up (anabolism) or of breaking down (katabolism) under the influence of radiant energy or its effects. In the katabolic state the "warm" sensations white, yellow and red are produced. In the anabolic state the "cold" senations black, blue and green are produced.

It is believed that Hering's theory has the advantage over the Young-Helmholtz theory in that it deals more directly with the sensations of colour.

This theory explains most of the colour phenomena such as after-images. For instance, blue light will cause anabolism in the yellow-blue substance resulting in an accumulation of this substance. If now yellow light is allowed to start the katabolic process, it will originally do so at a greater rate and the yellow sensation is thus greatly enhanced. The continuation of the katabolic process results in the complete breakdown of the relevant substance, resulting first in reduction of the yellow sensation and eventually ambiguity of the sensation. Hering's theory fails, however, to give an account of protanopia and tritanopia, without ad hoc assumptions.

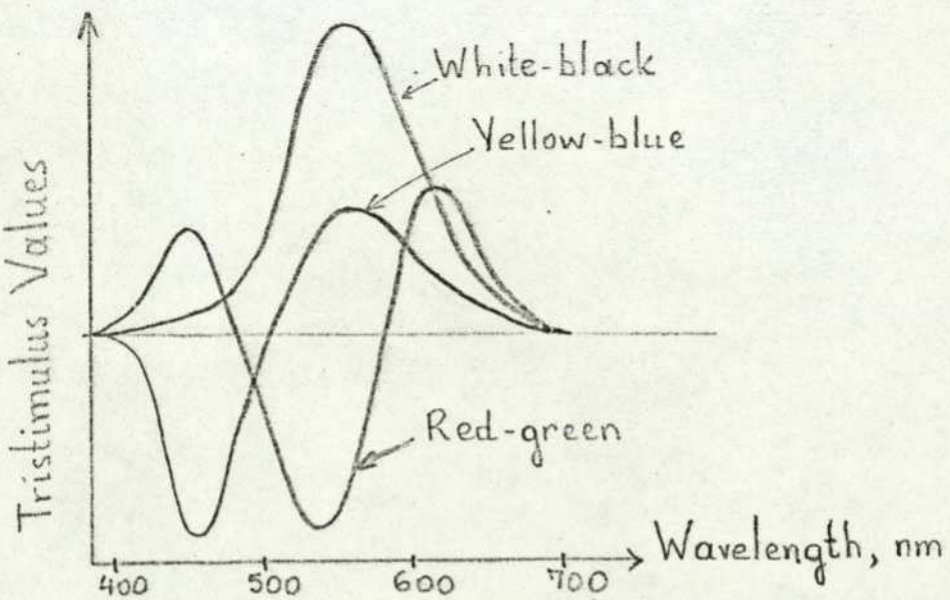


Fig. 3.3 Relative response functions postulated in Hering's Theory.

For a few decades following Hering's theory, a large number of experiments were carried out to learn more about the precise nature of colour perception. These studies confirmed beyond doubt both the three dimensional character of colour vision and the plausibility of Hering's opponent-colours theory. Mü ller (1930) suggested that the Young-Helmholtz theory and the Hering theory were not competing but rather supplementing each other.

Mü ller then proposed in his "three-stage zone theory" that the Young-Helmholtz concept of three types of colour receptors in the retina of the eye was correct, but that responses from these three receptors were converted in an elaborate nerve-signal switching area within the eye and optic nerve to opponent-colour signals such as Hering postulated.

Judd (Judd, 1949), in a detailed analysis of the Mueller theory, successfully explained the observed effects associated with the different forms of colour blindness.

In 1953 (Motokawa, 1953 and Svaetichin, 1953) work on fish retinae revealed that some cells respond by slow potential changes. The magnitude of these potential changes were proportional to the strength of the stimulus. Some other cells produced a negative potential with blue light but positive with yellow light. Also, other cells were found that produced negative potentials with green light but positive potential with red light. These findings were later confirmed by De Valois (De Valois, 1958) who worked on the

macaque monkey whose colour vision is almost identical to that of man. In these experiments, the generated electrical potentials detected by the use of micro-electrodes (whose tip was 0.1 mm) placed in the lateral geniculate body in the base of the brain, were recorded. Thus these recordings may be either due to the responses from the retinal ganglion cells, or may be the result of these modified by further activity in the lateral geniculate.

The first type of cells which produces just positive potentials are called non-opponent cells. These cells seem to be responsible for the signal which carries the luminosity information. The other two types of cells which produce both positive as well as negative potentials are called opponent cells, as they assess the relative strengths of opposing pairs of colours.

The results of these investigations, together with the microphotometric studies of single cones by Marks (Marks,1964) and Brown and Wald (Brown and Wald, 1964) which proved the existence of three different types of cone receptors absorbing light in three different regions of the spectrum, have given convincing evidence which resulted in the development of the modern theory of colour vision. This theory can be followed on the chart of Fig. 3.4.

According to these modern ideas, the Young-Helmholtz theory is true in the first stage of the process, whilst Hering's opponent theory can be applied to the following stages. As shown in Fig. 3.4 the outputs from the three types of receptors are firstly converted to

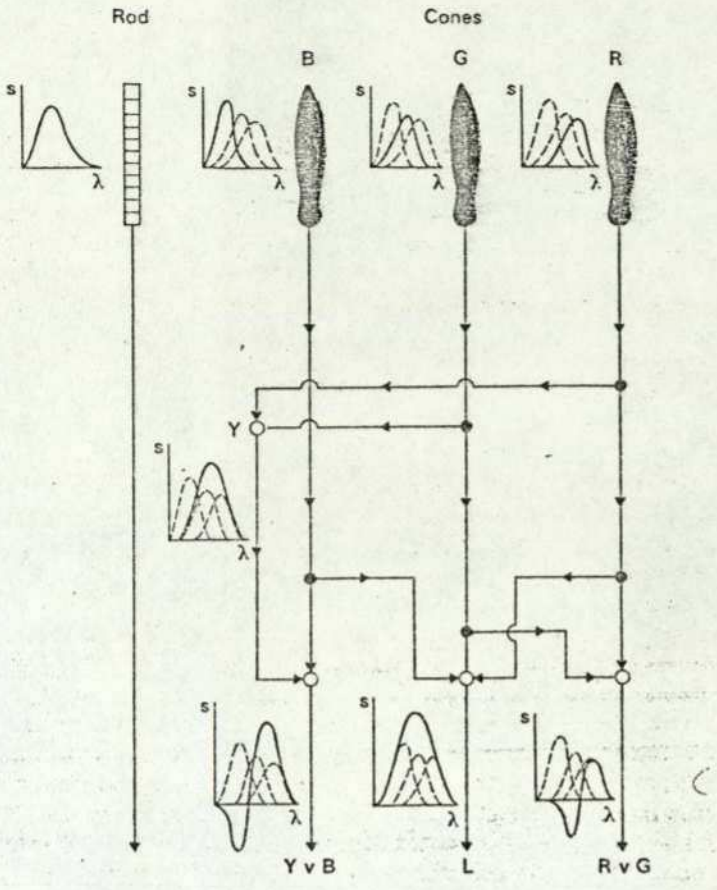


Fig. 3.4 Principle of the modern theory of Colour Vision.

electrical potentials which are then frequency coded for transmission to the brain. The summation of these outputs serves as the luminosity signal. A luminosity signal is also derived from the independent rod receptors.

The yellow signal is derived from a summation of the Green and Red signals. The derived Yellow signal together with the Blue signal form the first colour difference signal which is then frequency modulated for transmision to the brain. The second colour difference signal is obtained from the Red against the Green signals.

3.2 Laws of Colorimetry and Colour Specification

Newton was probably the first person to consider colour matching and colour mixing from a scientific point of view. As pointed out earlier on, in devising his Colour Circle he also enunciated his laws on colour mixture.

Plateau was the originator of the revolving disc for colour evaluation. He was, however, blinded before he could finish his experiments on complementary colours and colour mixtures.

Maxwell used revolving discs to check the validity of the laws of mixtures enunciated by Newton.

The founder of colorimetry was, however, the distinguished mathematician Grassman who in an amazing theoretical study (Grassman, 1853) laid down the principles of colorimetry:

(a) The eye can distinguish only three kinds of difference or variations (expressible as variations in dominant wavelength, luminance and purity).

(b) If one of a two-component mixture of lights is continuously altered (whilst the other remains unchanged) the impression of the mixed light is also continuously changed.

(c) Two colours with identical dominant wavelength, luminance and purity give constant mixed colours, no matter of what homogeneous colours they may be composed.

(d) the luminance produced by the additive mixture of a number of lights is the sum of the luminances produced separately by each of the lights.

Grassmann's laws are so vital to colorimetry that they have been scrutinised by the most critical experimental procedures. They have been found to hold within the range of luminances usually met in normal daylight.

From the third principle it follows that if light A matches light B, and light C matches light D, then the mixed light produced by A added to C colour matches the mixture of B and D. It also follows that the remainder produced by removing A from C colour matches the remainder produced by removing B from D. Furthermore, it follows that increasing or decreasing by a factor the radiance of two lights of the same colour but keeping their spectral compositions unchanged will not destroy the colour match, regardless of the composition of the two matching lights.

3.2.1 Geometric and Algebraic Representation of the Laws of Colour Matching by Mixtures of Lights

Following Newton and Maxwell, Abney (Abney, 1900) proceeded to elaborate on the algebraic representation of the laws of colour mixing, whilst Ives (Ives, 1915 and 1923) made very useful contributions in the geometrical illustration of these laws in a tridimensional space.

In the geometrical interpretation of colour space, three vectors $\vec{R}$, $\vec{G}$ and $\vec{B}$ corresponding to the three primaries Red, Green and Blue, will always describe the tridemensional colour space, provided that they meet at a common point (the origin O) and that they do not collapse into a single plane.

Fig. 3.5 shows a co-ordinate system of the three primaries $\vec{R},\vec{G},\vec{B}$ and a colour $\vec{S}$ with tristimulus values R,G,B defined by these three primaries.

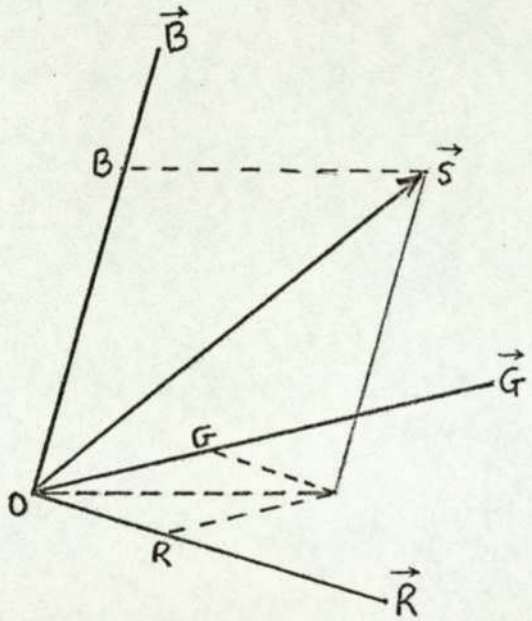


Fig. 3.5 Geometrical interpretation of colour space.

The unit lengths of the three axes corresponding to the unit amounts of the three primaries may be chosen in an arbitrary way - normally chosen so that equal amounts of $\vec{R},\vec{G},\vec{B}$ result in a neutral colour. Thus in Fig. 3.6 vector $\vec{N}$, corresponding to a neutral colour, intersects the unit plane R+G+B = 1 at the point N the central point of the triangle RGB.

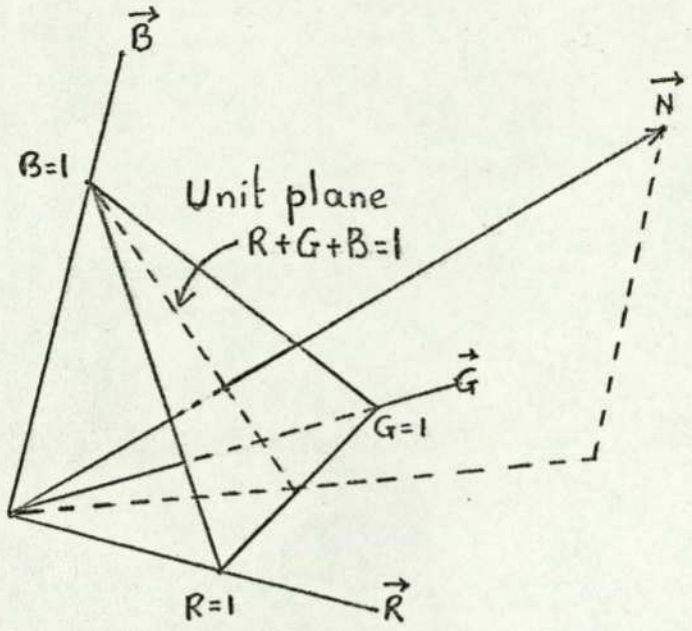


Fig. 3.6 Neutral colour $\vec{N}$ crossing the unit plane.

A colour match between $\vec{S}$ and $\vec{R},\vec{G},\vec{B}$, can be expressed algebraically by,

$$\vec{S}\ (R,G,B) = R\vec{R} + G\vec{G} + B\vec{B}$$

From Grassmann's laws of the additivity of luminances, the luminance L of $\vec{S}$ (R,G,B) is, $L = l_R R + l_G G + l_B B$

where l_R, l_G and l_B are the luminances of the primaries at R = G = B = 1. It also follows from Grassmann's laws that the length of the vector S is proportional to the luminance of the colour S.

When a colour $\vec{S}_1$ (R_1,G_1,B_1) is mixed with a colour $\vec{S}_2$ (R_2,G_2,B_2) to produce a colour $\vec{S}_o$ $(R_o,G_o,B_o) = \vec{S}_1\ (R_1,G_1,B_1) + \vec{S}_2\ (R_2,G_2,B_2)$

$$= (R_1\vec{R} + G_1\vec{G} + B_1\vec{B}) + (R_2\vec{R} + G_2\vec{G} + B_2\vec{B})$$

$$= (R_1 + R_2)\,\vec{R} + (G_1 + G_2)\,\vec{G} + (B_1 + B_2)\,\vec{B}$$

Thus, $R_0 = R_1 + R_2$

$G_0 = G_1 + G_2$

$B_0 = B_1 + B_2$

ie, the tristimulus values of the mixture are the sum of the tristimulus values of the component colours. This relationship can be extended to apply to any number of colours.

In the geometric representation the colour of the mixture can be obtained as shown in Fig. 3.7. S_0, S_1 and S_2 lie in a common plane and thus the three points lie in a straight line.

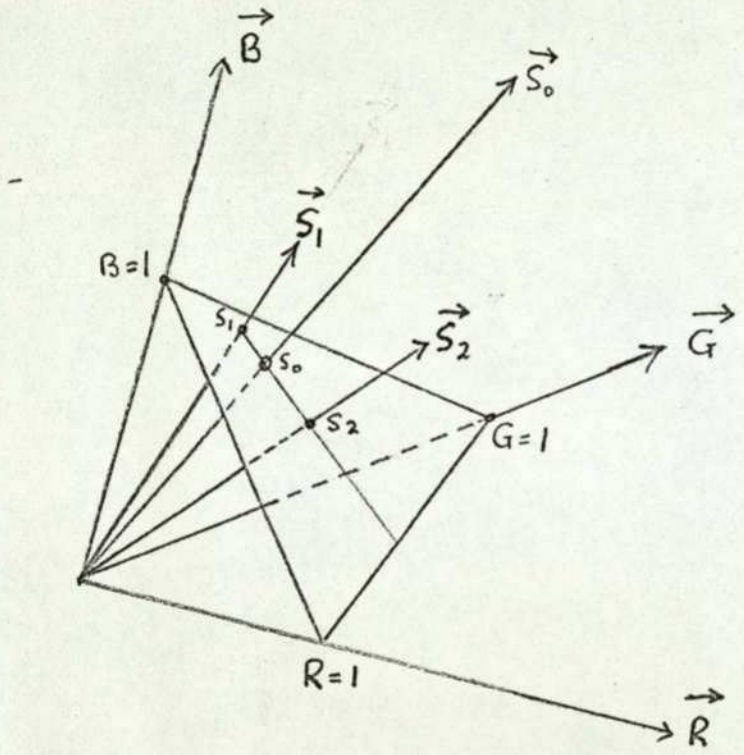


Fig. 3.7 The addition of two colours in vector space.

Through the use of the unit plane, a useful simplification of the representation of colours in the geometrical model can be obtained. The intersection point S, of the colour vector S and the unit plane defines the chromaticity of S. The unit plane is called the chromaticity diagram. This is shown in Fig. 3.8.

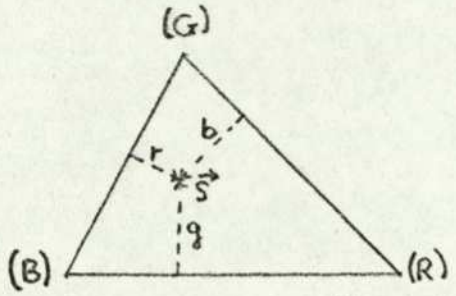


Fig. 3.8 The chromaticity diagram (unit plane).

The chromaticity co-ordinates r,g,b are introduced to define the position of $\vec{S}$. The corners of the triangle R,G,B, can be assumed to be located at the points (1,0,0), (0,1,0), (0,0,1). From the geometry of the diagram,

$$r = \frac{R}{R+G+B}$$

$$g = \frac{G}{R+G+B}$$

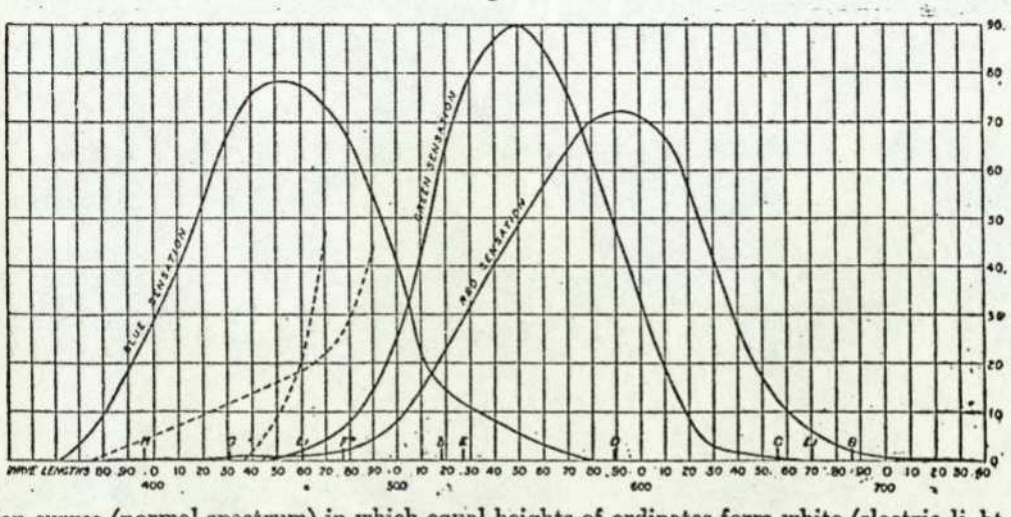
$$b = \frac{B}{R+G+B}$$

from which r+g+b=1.

The original form of chromaticity diagram, the equilateral triangle suggested by Maxwell, which had the advantage of showing the three primary lights symmetrically, has been replaced by the right angle form to enable the use of ordinary graph paper. This was suggested by Guild (Guild, 1925).

3.3 The Colour Sensations

Following the theoretical work of Young and Helmholtz and also of Konig who postulated the various response functions (colour sensations) discussed earlier in this chapter, Abney (Abney, 1899) decided to redetermine these functions by a luminosity method. To achieve this, Abney used the apparatus described on p.57 (photocopy from his original paper). The light used was from the crater of the positive pole of an electric arc. The sensation curves obtained by Abney are shown in Fig. 3.9. These curves correspond to those of a normal spectrum (equal heights of ordinates form white). Fig. 3.10 shows the equivalent snesations reported by Konig. All the observations relating to Fig. 3.9 were made by Abney himself; no other observers were used.

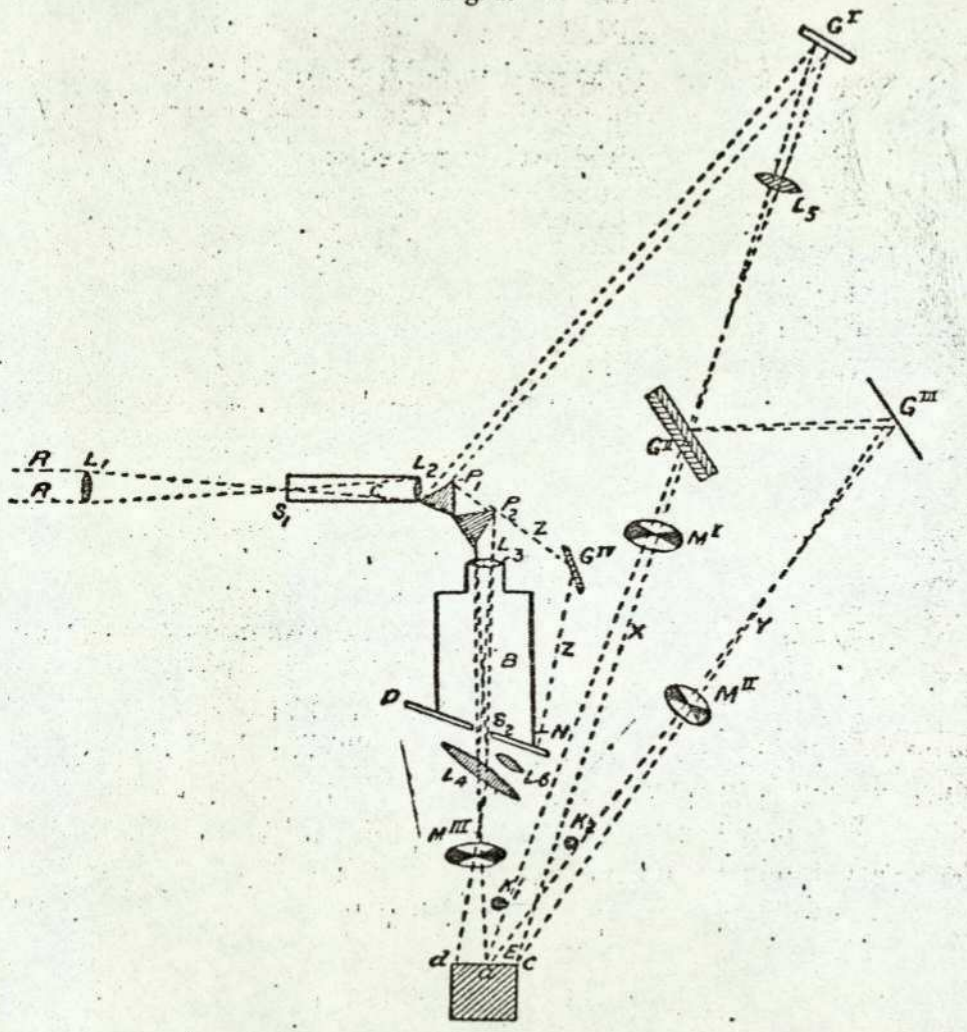


Sensation curves (normal spectrum) in which equal heights of ordinates form white (electric li ht crater).

Fig. 3.9 Sensation curves obtained by Abney.

After passing through the collimator C, the rays emerged as parallel rays; part passed through the prisms P_1 and P_2, were collected by a lens, L_3, and a spectrum was formed on a slide, D (which will be more fully described), in which slits could be placed, and an image of the surface of the first prism was formed on the white-red surface of a cube, E, by means of the lens L_4, so arranged that the image of one edge of the prism fell at *a*, the other edge falling outside *d*. The other beam which passed through the collimator was reflected from the surface of the first prism to a mirror, G^I, and passed

Fig. 2.



through a lens, L_5, then through a bundle of glass, G^{II}, placed at an angle to the beam, and on to the surface *ac* of the cube, a rod, K_1, being placed in its path, to secure that this white beam did not fall on *ad*, on which the colour mixture fell. The portion of the beam which was reflected from G^{II} was again reflected by G^{III}, a silvered mirror, on to *ac*, a rod, K_2, being placed in its path to prevent it falling on *ad*. In all three beams, sectors, M^I, M^{II}, and M^{III}, were placed, to allow any or all to be reduced in intensity at pleasure. In the beams X and Y any absorbing medium desired could be placed.

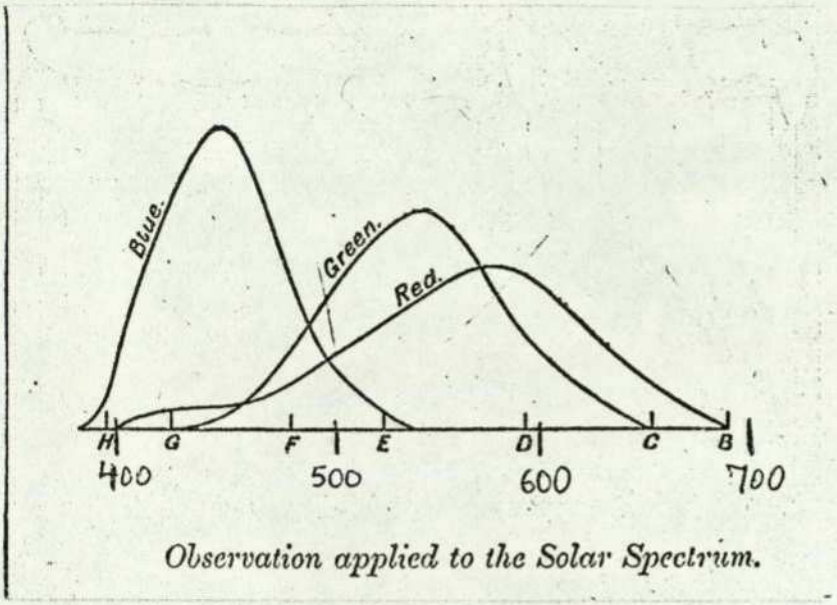


Fig. 3.10 Sensation curves due to Kőnig.

Serious attempts to obtain the sensation curves in large scale experiments using a number of observers were reported by Wright (1929) and Guild (1931).

Using a specially designed colorimeter, Wright obtained results concerned with the determination of the trichromatic coefficients of the spectral colours. Ten observers were used. Wright used the relative amounts of three selected primaries required to match all the spectral colours and the relevant luminosity data, to determine the mixture curves and the sensation curves.

Guild, using another specially designed colorimeter, obtained the sensation curves, employing seven observers.

Both Wright and Guild underestimated the accuracy and significance of their results. They both considered that a much larger number of experimenters was necessary to obtain the true curves for the "typical observer". However, when the two sets of results were compared it was surpising to observe the closeness of these curves.

This similarity led to the average of these two sets of results to be adopted by the Commission Internationale de l'Eclairage (C.I.E.) in 1931 as the "Standard Observer" sensation curves.

This work of Wright and Guild has been of such a great significance in colorimetry, that it is fully described in the next chapter together with all the properties of the C.I.E. system.

CHAPTER 4

THE C.I.E. SYSTEM

4.1 Transformation of Primaries

To enable the comparison of tristimulus colorimeters that are using different primaries, it is essential to derive formulae relating the tristimulus values obtained on one tristimulus colorimeter for any colour with those obtained on another tristimulus colorimeter for any colour with those obtained on another tristimulus colorimeter of known primaries.

A second reason for the transformation from one set of primaries to another is the requirement to eliminate negative numbers for the specification of colour by the tristimulus method. Negative numbers arise due to the impossibility of matching some colours by the mixture of three other spectral colours. To enable such a match, the tested colour is normally desaturated with one of the three primary colours and this mixture is then matched with the other two primary colours. This match is algebraically specified by giving a negative value to the primary mixed with the colour being tested.

A third reason for the transformation of primaries, is the need to convert to (or from, in some isolated cases) a standard system.

4.2 Mathematics of Colorimetric Transformations

The analysis that follows is based on Guild's ideas (Guild, 1924), and the symbolism is the one adopted by the C.I.E. in 1955. (W D Wright 1969).

A match on a colorimeter of a colour C can be represented by the equation

$$q.(C) \equiv u.(R) + v.(G) + w.(B) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 4.1$$

where q,u,v,w are the amounts of (C),(R),(G),(B) at some suitable scale. The brackets around a symbol imply unit quantity; without the brackets the symbols represent numerical quantities referred to as tristimulus values. The identity sign is used to show a colour match.

A useful scale for (R), (G), (B) that has been widely used, is that where equal amounts of (R), (G), (B) match a white of some defined quality. For a trichromatic colorimeter whose matching stimuli are monochromatic radiations with the units of the stimuli adjusted to be equal when matched to a standard white, the quality of the matching stimuli could be defined by their wavelengths, while the standard white could be defined by its colour temperature (assuming a Planckian radiator). On this scale, equation 4.1 could be given by,

$$c.(C) = R.(R) + G.(G) + B.(B) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 4.2$$

On this system, $c = R + G + B$ $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$ 4.3

To separate the colour quality of C as matched by R,G,B from the quantity of light being matched, it is normal to divide R,G,B by (R + G + B), ie

$$1.0\ (C) \equiv r.(R) + g.(G) + b.(B) \quad \ldots\ldots 4.4$$

$$\text{where } r = \frac{R}{R+G+B},\ g = \frac{R}{R+G+B},\ b = \frac{R}{R+G+B} \quad \ldots\ldots 4.5$$

$$\text{from which } r+g+b=1 \quad \ldots\ldots 4.6$$

Equation 4.4 is known as a unit trichromatic equation, and the amount of C in a trichromatic equation a one trichromatic unit, or 1 T-unit.

From equation 4.6 it can be seen that one of r,g,b, (referred to as the chromaticity co-ordinates of C) can be eliminated, and the quality of C can therefore be shown on a two-dimensional diagram (the chromaticity diagram) shown in Fig. 4.1.

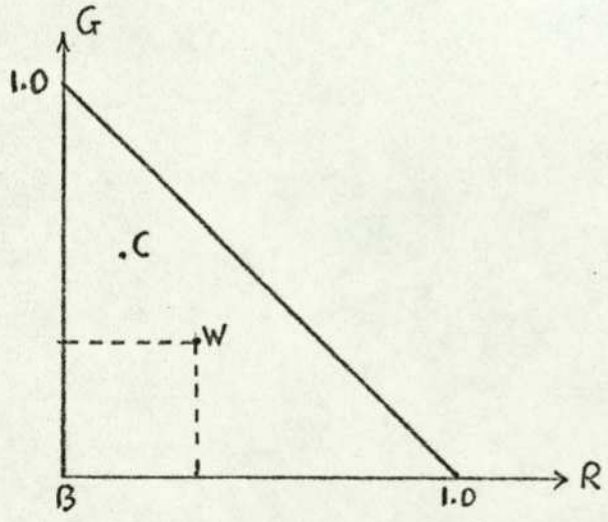


Fig. 4.1 The chromaticity diagram.

In this diagram the standard white used for matching is located at (0.333, 0.333). The orange, yellow and yellow green colours lie along the (R) (G) line, ie r+g=1; the blue-green colours lie along (G) (B) where r=0; the purples lie along (R) (B) where g=0. All the desaturated colours lie within the triangle. Newton's law of colour mixing applies anywhere on the chromaticity diagram.

Consider the transformation (W D Wright, 1969) of a colour C defined by the unit equation 4.7 into a new unit equation 4.8.

$$[C] \equiv r.\ [R] + g.\ [G] + b.\ [B] \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad 4.7$$

$$(C) \equiv x.\ (X) + y.\ (Y) + z.\ (Z) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad 4.8$$

where [R], [G], [B] are the old set of reference stimuli and (X), (Y), (Z) are a new set.

Fig. 4.2 shows [X], [Y], [Z] plotted in the [R],[G],[B], diagram.

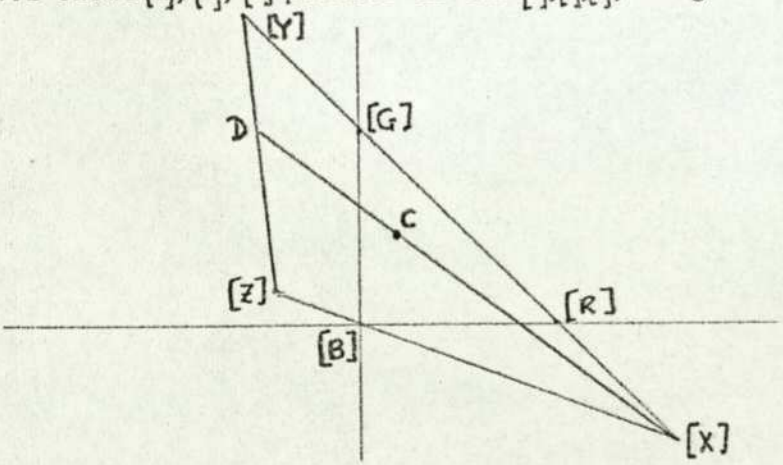


Fig. 4.2 [X],[Y],[Z], plotted in the [R],[G],[B] chromaticity diagram.

Referring to Fig. 4.2, to match colour C the amounts of [R],[G],[B] required are r,g and 1 - (r+g). This colour can also be matched by a mixture of [X] with [D] in the proportion, in T-units, of DC to C[X].

As D can be obtained by a mixture of [Y] and [Z], it follows that C can be matched by a mixture of [X], [Y] and [Z].

Thus, [C] = f [XYZ], where

$$\left.\begin{aligned} [X] &\equiv r_x . [R] + g_x . [G] + b_x . [B] \\ [Y] &\equiv r_y . [R] + g_y . [G] + b_y . [B] \\ [Z] &\equiv r_z . [R] + g_z . [G] + b_z . [B] \end{aligned}\right\} \quad \ldots\ldots\ldots \quad 4.9$$

The unit quantities of (X), (Y) and (Z) in the (XYZ) system are, however, different to those of [X], [Y] and [Z] in the [RGB] system. The units of [R], [G] and [B] are adjusted to satisfy the condition,

$$[W] = 0.333\ [R] + 0.333\ [G] + 0.333\ [B] \ldots\ldots\ldots\ldots\ldots\ 4.10$$

For the new system, the units of (X), (Y) and (Z) have to be adjusted to satisfy,

$$(W) = 0.333\ (X) + 0.333\ (Y) + 0.333\ (Z) \ldots\ldots\ldots 4.11$$

It is possible to solve equation 4.9 to obtain,

$$\left.\begin{aligned} [R] &\equiv x'_r . [X] + y'_r . [Y] + z'_r . [Z] \\ [G] &\equiv x'_g . [X] + y'_g . [Y] + z'_g . [Z] \\ [B] &\equiv x'_b . [X] + y'_b . [Y] + z'_b . [Z] \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots 4.12$$

From equations 4.12, where [R], [G] and [B] each represent one T-unit in the [RGB] system, it follows that, equations 4.12 are unit equations,

$$\left.\begin{aligned} \text{ie, } x'_r + y'_r + z'_r &= 1 \\ x'_g + y'_g + z'_g &= 1 \\ x'_b + y'_b + z'_b &= 1 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 4.13$$

Substituting equation 4.9 into equation 4.10,

$$[W] \equiv x'_e . [X] + y'_e . [Y] + z'_e . [Z] \ldots\ldots\ldots\ldots\ldots\ 4.14$$

Assuming that the two standard whites are identical, ie (W) $\equiv$ [W] , then from equations 4.11 and 4.14 it follows that,

$$\left.\begin{aligned} [X] &= \frac{0.333}{x'_e} \,.\, (X) \\ [Y] &= \frac{0.333}{y'_e} \,.\, (Y) \\ [Z] &= \frac{0.333}{z'_e} \,.\, (Z) \end{aligned}\right\} \qquad \ldots\ldots\ldots\ldots \; 4.15$$

These values of [X], [Y] and [Z] can now be substituted into equations 4.12 to obtain [R], [G] and [B] in terms of (X), (Y) and (Z). These values of [R], [G] and [B] can therefore be substituted into equation 4.7 to obtain [C] in terms of (X), (Y) and (Z), ie

$$[C] \equiv x' . \quad (X) + y' . \quad (Y) + z' . \quad (Z) \qquad \ldots\ldots\ldots\ldots \; 4.16$$

As this equation involves both the old and the new systems, it is not a unit equation. Colour C in the two systems is, however, related by,

$$[C] \equiv (x' + y' + z') . \quad (C)$$

Therefore, $(C) \equiv \frac{x'}{x'+y'+z'} \,.\, (X) + \frac{y'}{x'+y'+z'} \,.\, (Y) + \frac{z'}{x'+y'+z'} \,.\, (Z)$

4.3 The Standard Observer

To avoid the inaccuracy and all the other disadvantages of visual colorimeters, it is possible to specify colour in terms of its spectral composition. To obtain the chromaticity of a colour from its spectral composition it is, however, necessary to use some distribution coefficients related to the response of the human eye.

To establish such a system, the distribution coefficients have to be agreed upon as representing an average human eye. As mentioned in Chapter 3, Konig, Abney, Guild and Wright have tried to obtain an average value of these coefficients.

Agreement on the distribution coefficients only is not, however, enough. The reference stimuli used to match the colours through the spectrum have also to be agreed upon.

Fig. 4.3 shows the chromaticity co-ordinates of the spectrum for seven observers obtained by Guild (1931), as mentioned in Chapter 3.

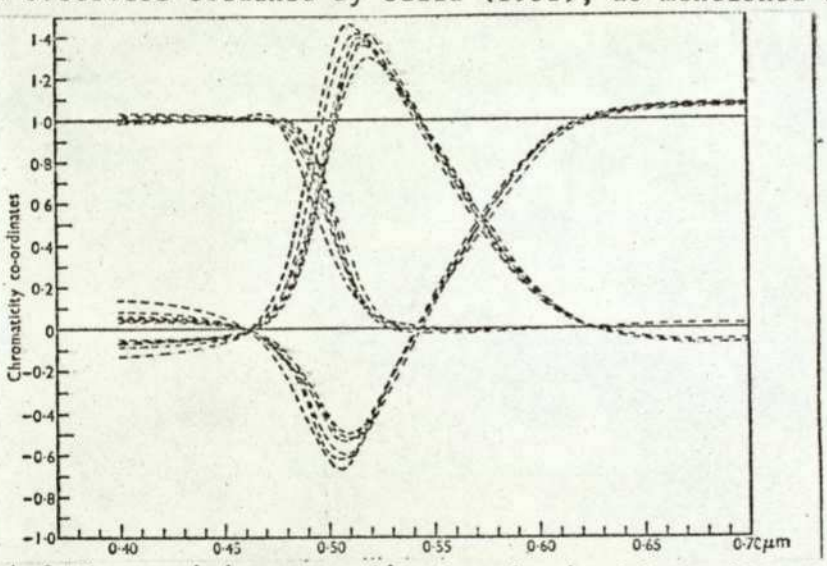


Fig. 4.3 Chromaticity co-ordinates obtained by Guild.

The units of the stimuli (monochromatic radiations of 700, 546.1 and 435.8 nm) used in this investigation were adjusted to be equal when they matched the standard N.P.L. white. Negative values show the necessity of desaturation for some colours to enable a match.

Fig. 4.4 shows the chromaticity co-ordinates obtained by Wright (1928), using ten observers. The units of the stimuli used were adjusted so that equal amounts of two of the monochromatic radiations used (650 and 530 nm), were needed to match a yellow of wavelength 582.5 nm, and also so that equal amounts of 530 nm and 460 nm (the third stimulus used) were needed to match a blue-green of wavelength 494 nm. This system had the advantage that the spectral co-ordinates were unaffected by pigmentation in the eyes of the observers.

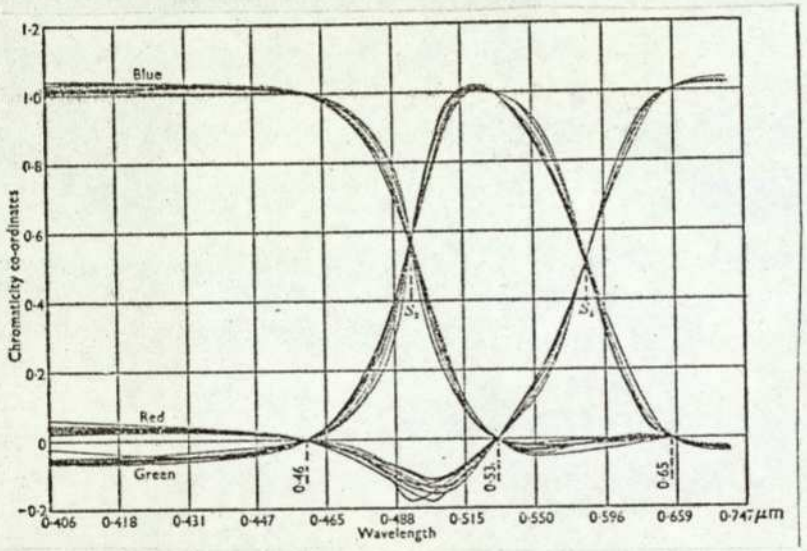


Fig. 4.4 Chromaticity co-ordinates obtained by Wright.

Guild (1931) compared the two sets of data as obtained by Wright and by himself. Fig. 4.5 shows the chromaticity co-ordinates obtained with Wright's results transformed to the same reference stimuli as used by Guild, and also based on the N.P.L. white. The chromaticity diagram also shows the locus obtained by Weaver who worked out the average between Konig's and Abney's results.

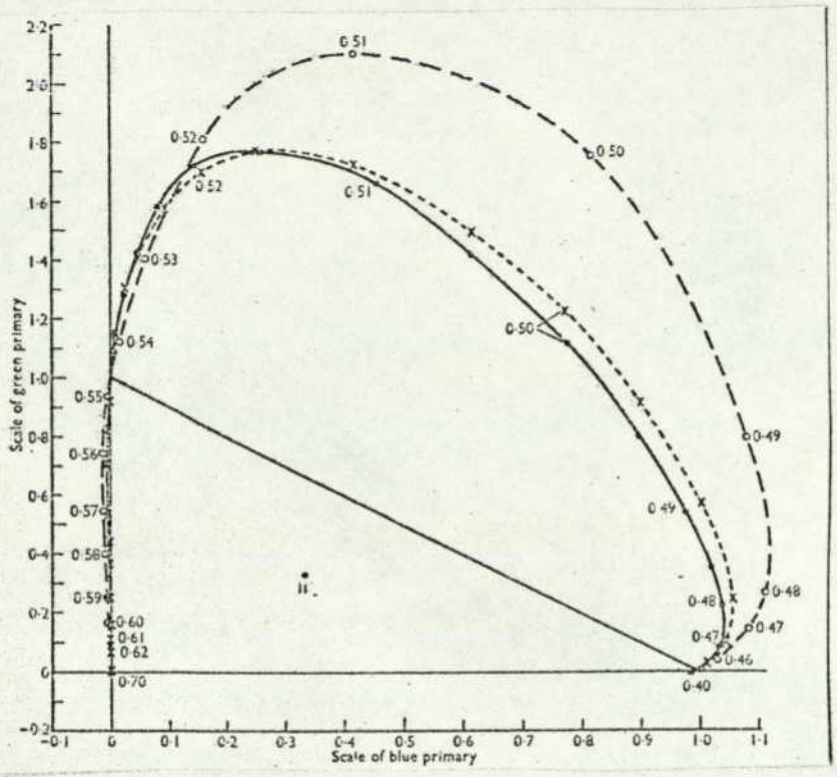


Fig. 4.5 Comparison of the results of the experiments by Guild and Wright.

The high correlation obtained between the two sets of results (Guild's and Wright's), "led to the recommendation and eventual adoption by the C.I.E. of their mean values to define the colour-matching characteristics of the 1931 standard observer".

To completely specify the colour-matching properties of the "standard observer", it is necessary to know the equal-energy distribution coefficients $\bar{r}_\lambda, \bar{g}_\lambda, \bar{b}_\lambda$. To save calculating these coefficients from the number of T-units at each wavelength in the equal-energy spectrum, it was decided to use the standard V_λ curve adopted by the C.I.E. in 1924.

For a colour λ, whose unit equation is given by,

$$[\lambda] \equiv r_\lambda . [R] + g_\lambda . [G] + b_\lambda . [B]$$

The flux in one T-unit of is,

$$V_{[\lambda]} = r_\lambda \cdot V_{[R]} + g_\lambda \cdot V_{[G]} + b_\lambda \cdot V_{[B]}$$

As the flux at wavelength λ in the equal-energy spectrum is proportional to V_λ, the number of T-units at λ in the equal energy spectrum will be proportional to

$$\frac{V_\lambda}{V_{[\lambda]}} \quad \text{or} \quad \frac{V_\lambda}{r_\lambda \cdot V_{[R]} + g_\lambda \cdot V_{[G]} + b_\lambda \cdot V_{[B]}} = m_\lambda$$

The three distribution coefficients are then,

$$\bar{r}_\lambda = m_\lambda \cdot r_\lambda$$

$$\bar{g}_\lambda = m_\lambda \cdot g_\lambda$$

$$\bar{b}_\lambda = m_\lambda \cdot b_\lambda$$

where $r_\lambda, g_\lambda, b_\lambda$, are the distribution coefficients already measured and $\bar{r}_\lambda$, $\bar{g}_\lambda$, $\bar{b}_\lambda$ are the equal-energy spectrum coefficients. The values of $V_{[R]}$, $V_{[G]}$ and $V_{[B]}$ can be obtained experimentally from the three instrument stimuli. For his colorimeter Guild obtained

$$V_{[R]} = 1.00$$

$$V_{[G]} = 4.39$$

$$V_{[B]} = 0.048$$

After modification of the units of [R],[G],[B] to apply to an "equal-energy" white, these new values of $V_{[R]}$, $V_{[G]}$ and $V_{[B]}$ were used by the C.I.E. to derive the distribution coefficients, shown in Fig. 4.6 and thus complete the definition of the standard observer. The spectrum locus resulting in terms of the stimuli 700,546.1 and 435.8 nm with the units based on an equal-energy white E is shown in Fig. 4.7.

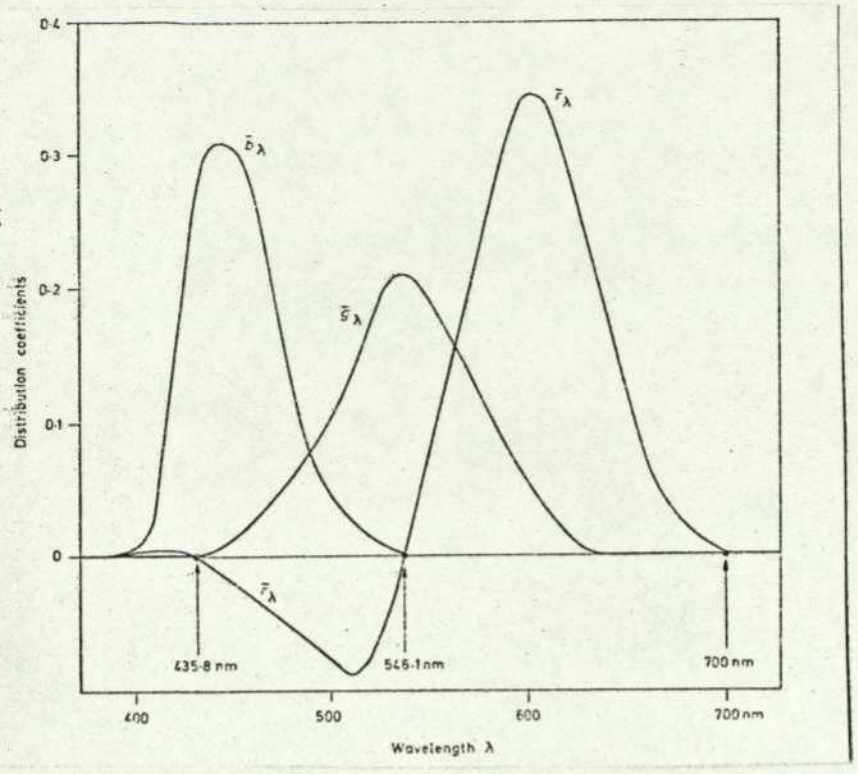


Fig. 4.6 Spectral distribution curves for the equal-energy spectrum for the 1931 C.I.E. standard observer.

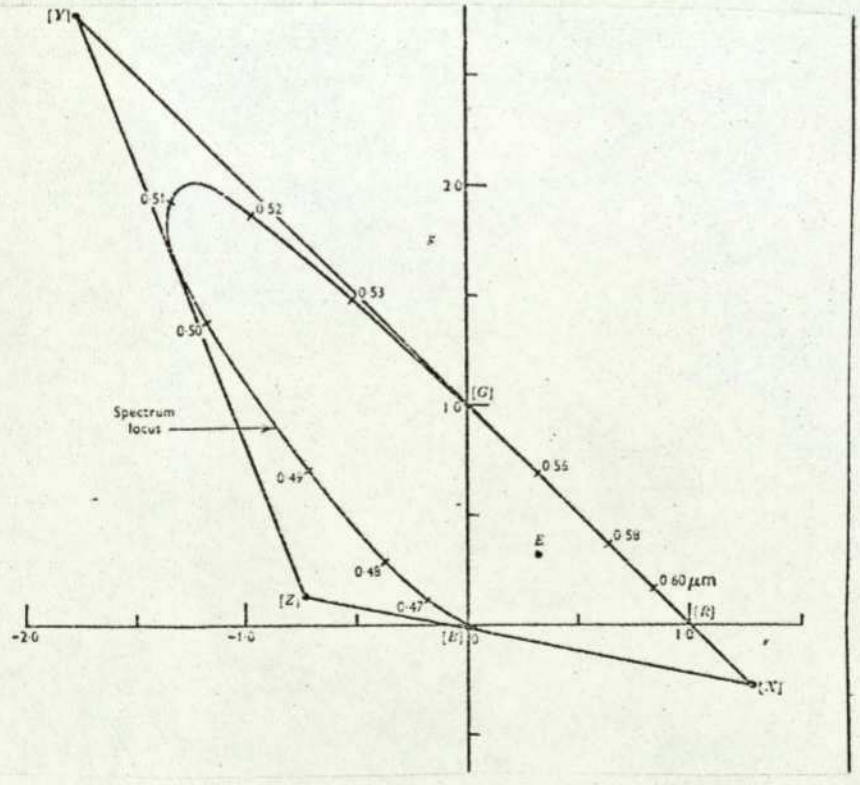


Fig. 4.7 Spectrum locus adopted by C.I.E. in 1931 for standard observer in terms of the stimuli 700,546.1 and 435.8 nm.

4.4 All-Positive Distribution Coefficients

It can be seen from Figs. 4.3, 4.4 and 4.6 that negative quantities are involved, which are highly undesirable in practice for colour specification. It was therefore decided to adopt an all-positive system using the reference stimuli [X], [Y] and [Z] as shown in Fig. 4.7. The triangle formed by these three points encompasses the whole of the area defined by the spectrum locus of the [R],[G],[B] stimuli.

These particular stimuli were chosen for the following reasons:

1. All physical stimuli lie within the triangle [XYZ], Fig. 4.7.
2. The line [XY] lies along the spectrum locus at the red end of the spectrum where the locus is straight. This reduces the amount of data in colour evaluation by eliminating the blue co-ordinates in the red region.
3. The line [X][Z], referred to as the alychne (no light), is the locus of all stimuli for which the flux due to C, $V_{[C]}$ is zero. By the use of this concept, which was enunciated by Schrodinger (1925), the photometric aspects of colorimetry are referred entirely to the third stimulus (Y). The equation of [X][Z] can be derived as follows:

   For a colour C defined by the unit equation 4.7, the relative luminous efficiency of one T-unit is given by,

$$V_{[C]} = r.\ V_{[R]} + g.V_{[G]} + b.\ V_{[B]}$$

Using Guild's corrected figures for V_R, V_G and V_B,

$$V_{[C]} = r + 4.5907\ g + 0.0601\ b$$

But $V_C = 0$ for the alychne, ie

$$0 = r + 4.5907\ g + 0.0601\ b$$

Since $r + g + b = 1$,

$$0.9399r + 4.5306\ g + 0.0601 = 0$$

which is the equation of the alychne.

Thus the photometric value of the light represented by a colour equation is completely defined by the third tristimulus value. Thus for the tristimulus equation,

$$c.\ (C) \equiv X.\ (X) + Y.\ (Y) + Z.\ (Z)$$

the luminous flux, F is

$$F = Y.\ V(Y)$$

It also follows that the $\bar{y}_\lambda$ distribution curve becomes identical to the V_λ curve.

The C.I.E. has produced tables for the chromaticity co-ordinates x_λ, y_λ and z_λ and also of the distribution coefficients $\bar{x}_\lambda$, $\bar{y}_\lambda$, $\bar{z}_\lambda$ for the equal energy spectrum for the wavelength range 380 nm to 780 nm at 5 nm intervals. The chromaticity chart and the distribution coefficients are shown in Figs. 4.8 and 4.9 respectively.

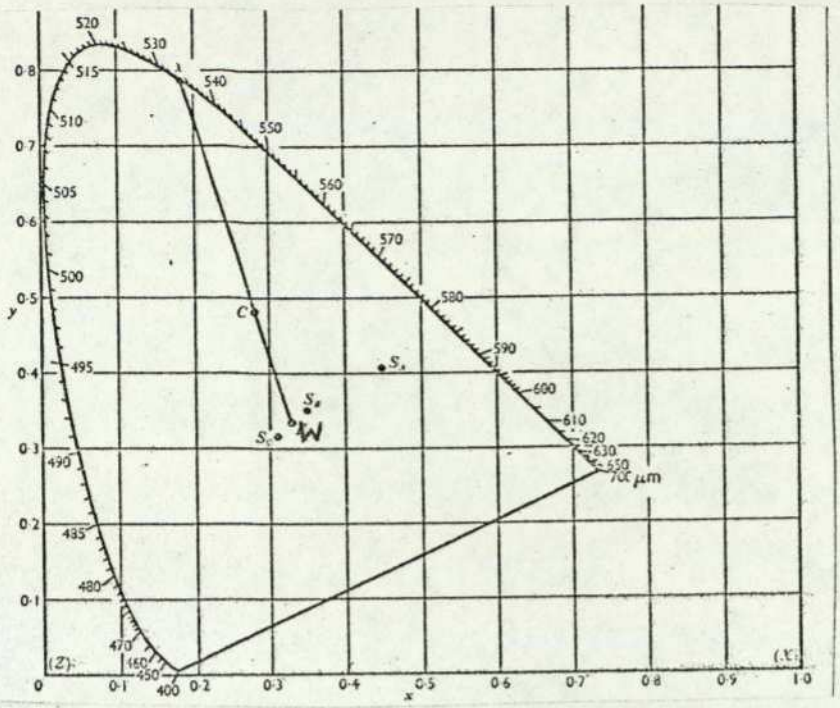


Fig. 4.8 The 1931 C.I.E. chromaticity chart in terms of the reference stimuli (X), (Y), (Z).

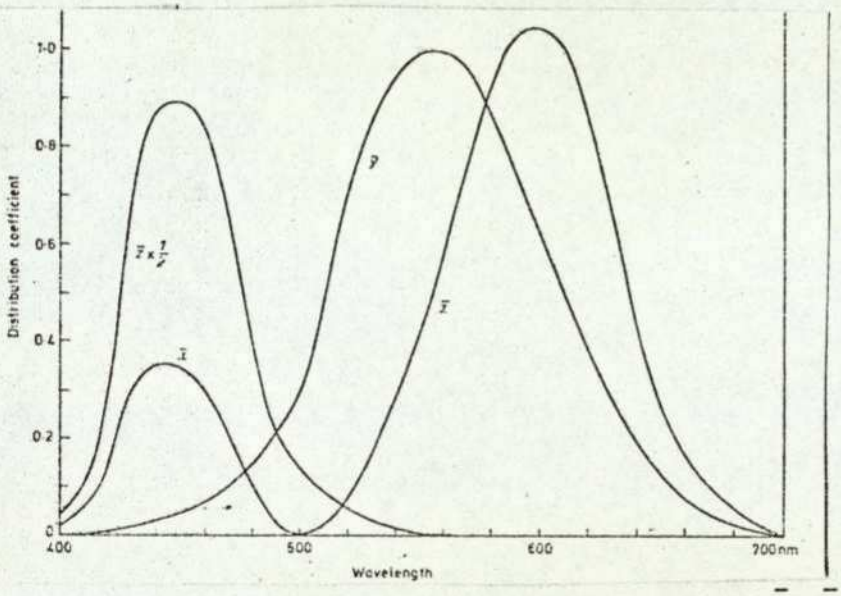


Fig. 4.9 The 1931 C.I.E. distribution coefficients x, y, z.

The results of Guild and Wright and of the V curve were later checked by Stiles (1955) and were found acceptable. In 1964 the C.I.E. defined the 1964 standard observer for large-field colour matching, accounting for rod intrusion, based on data collected by Stiles (1958) and Speranskaya (1959).

Fig. 4.10 shows the chromaticity chart in which a colour C can be specified in terms of either the (x,y) co-ordinates or its dominant wavelength and purity.

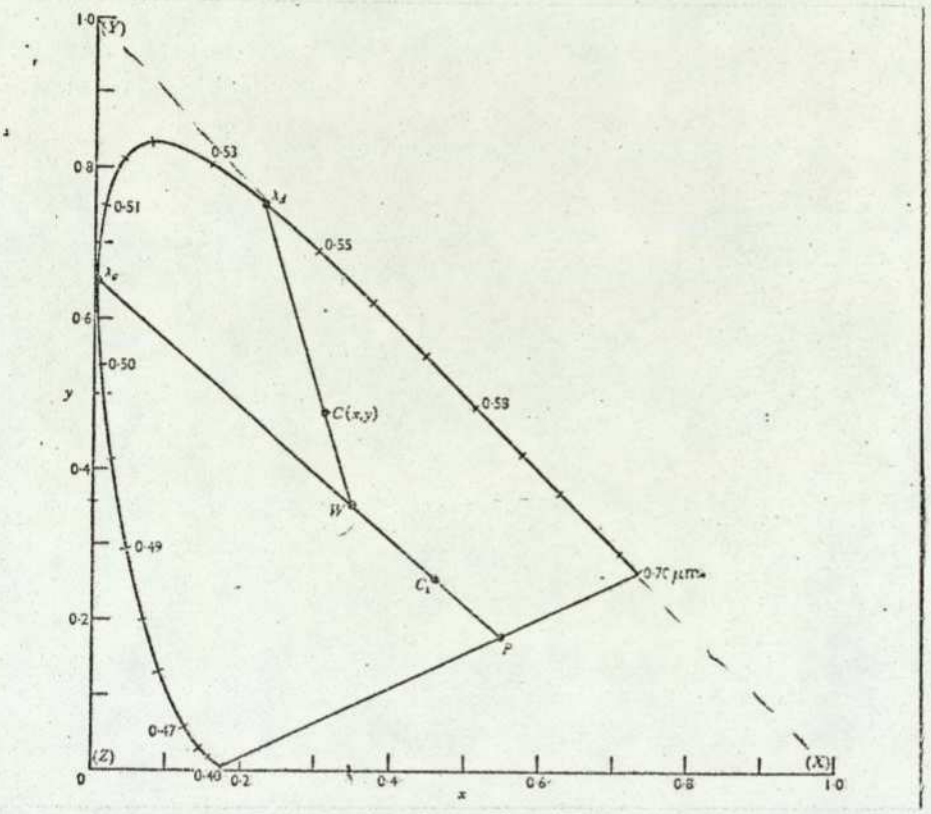


Fig. 4.10 Specification of a colour in terms of dominant wavelength and purity in the chromaticity chart.

In the chromaticity diagram, the purity, p, is given by

$$p = \frac{CW}{\lambda W}$$

If the co-ordinates of the source (white point) are (x_w, y_w) and those of the dominant wavelength are (x_d, y_d) then

$$p = \frac{x - x_w}{x_d - x_w} \qquad \text{or} \qquad p = \frac{y - y_w}{y_d - y_w}$$

CHAPTER 5

METHODS OF COLOUR MEASUREMENT

5.1 Naming

In the English language there is a total of 3500 words which describe colour. The average human eye can, however, distinguish the differences between several million colours and shades. It is thus obvious that a "word" system of complete colour specification is inadequate. Furthermore, the names and terms used to define colour are quite often not only confusing but also misleading. This is because the same words are used with different meanings amongst a number of industries or even amongst the arts and the sciences. Also, words such as "bright", "dark", "light", "rich", "pure", "deep", "dull", "tint", "tone", "shade" etc, present a very vague terminology, with inconsistent and indefinite meanings.

In spite of these limitations, communication can sometimes be established even with the use of such terms as "bright yellow" or "dark red". The usefulness of such terminology in the colour industry is obviously very limited.

Underlying the concept of all colour measurement scales is the science of psychophysics - the branch of scientific thought concerned with the relationship between the physical stimuli (ie the perception of physical magnitudes) and the effects induced by them on the human body.

Colour is a sensory experience defined by the three related attributes: brightness, hue and saturation.

Luminosity (a psychological attribute) has been associated with luminance (a physical stimulus). Thus the grading of luminosity is easy due to this facility of representing it in terms of a physical measurement.

A typical example where a colour classification has been established on the basis of colour naming, was the one adopted by the U.S. Pharmaceutical Convention (Judd and Kelly, 1939). The recommended colour designations originally only applied to powdered drugs and chemicals as viewed in daylight. The Inter-Society Colour Council eventually adopted the system for more general applicability of these designations to colours of opaque, non-metallic surfaces and for liquids and solids viewed by transmission of light.

This designation consists (except for very greyish colours) of a hue name (red, green, blue, purple, etc) preceded by appropriate modifiers (such as weak, moderate, strong, light, dark). The designation for very greyish colours consists of a noun (white, grey or black) with modifiers appropriate to the lightness and hue of the colours (such as dark reddish grey or yellowish white).

To establish the relationships between the names the psychological colour solid (Fig. 5.1) was subdivided into about 300 separate compartments.

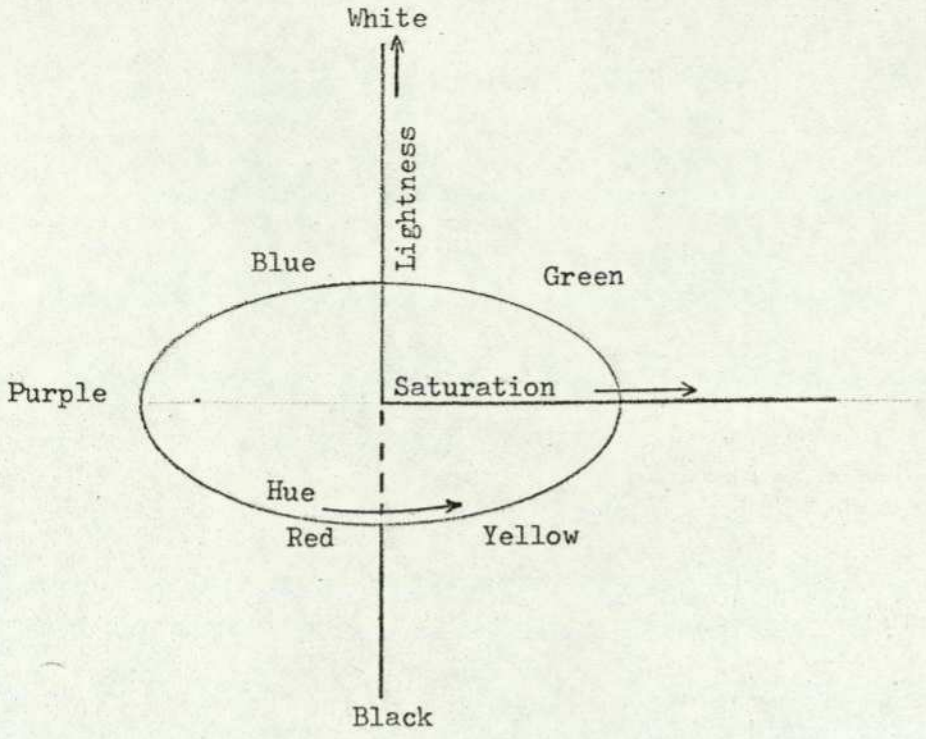


Fig. 5.1 Dimensions of the surface-colour solid.

Fig. 5.2 Shows the system of modifiers adopted.

VERY PALE (VERY LIGHT, WEAK)	VERY LIGHT	VERY BRILLIANT (VERY LIGHT, STRONG)
PALE (LIGHT, WEAK)	LIGHT	BRILLIANT (LIGHT, STRONG)
WEAK	MODERATE	STRONG
DUSKY (DARK WEAK)	DARK	DEEP (DARK, STRONG)
VERY DUSKY (VERY DARK, WEAK)	VERY DARK	VERY DEEP (VERY DARK, STRONG)

Lightness

VIVID (VERY STRONG)

Saturation

Fig. 5.2 System of modifiers.

5.2 Appearance Methods - Colour Atlases

As interest in colour can be both aesthetic as well as functional, colour design would not normally be done purely on the basis of

tristimulus values. It is normal to choose colours from a colour atlas rather than a chromaticity diagram. A colour atlas can thus be used as a bridge between the colour user and the colour scientist.

As colour atlasses are easily transportable or even in some cases individual colour chips sent by post, they are frequently preferred for colour specification to the C.I.E. system. Usually, the nearest chip to the desired colour is found and used as a comparison standard. It is, however, possible to estimate by direct visual interpolation a fairly reliable intermediate value.

Due to the necessity to keep the cost and the size of atlasses down, the number of colours that can be represented by colour chips in a colour atlas is only a small fraction of the total number of colours that can be discriminated by the eye. This sets a limitation to the use of atlasses.

Another severe limitation of the colour atlas is the subjectiveness of the whole process of colour matching.

The difficulty of preserving the colour of the pigments used for the colour chips from deterioration in use by soiling or fading is another disadvantage of the colour atlas.

The two most famous systems used in the tabulation of colour chips to form an atlas are the Ostwald and the Munsell.

5.2.1 The Ostwald System

This system, devised in 1920 by Wilhelm Ostwald, was one of the early successful scientific applications of the colour chart system to the specification of colours. As shown below, his system consisted of twelve diamond-shaped charts. Each chart consists of 64 nearly square colour patches, each side being about 12 mm.

As shown in Fig. 5.3, for a given hue the white is located at the corner W,

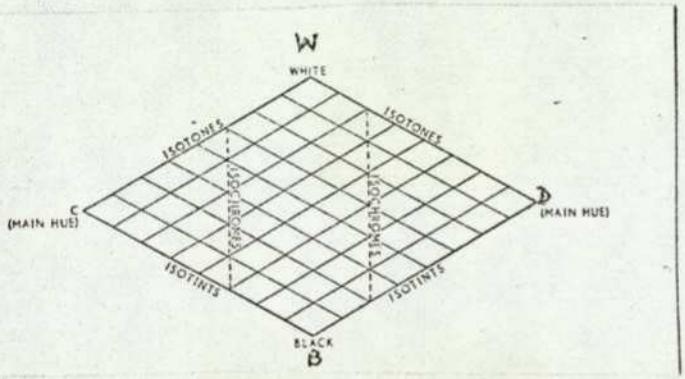


Fig. 5.3 The Ostwald System.

black at the corner B, and full colour pigments at corners C and D. Colours C and D are normally chosen to be complementary. The colours within the chart are derived by suitable mixing of white, black and colour, so that the sum of the white content, black content and colour content is equal to unity, ie W + B + C = 1 or W + B + D = 1.

Colours located along lines parallel to WC in the right half of the chart or along lines parallel to WD in the left half contain equal amounts of black and are called Isotones. To obtain uniform spacing of the colours along these lines, the proportion of the white content is adjusted.

Colours of constant white content (and variable black to make uniform spacings), such as those parallel to BC on the right hand side and parallel to BD on the left hand side, are known as Isotints. Those colours lying on vertical lines parallel to WB are colours of equal full-colour content and are known as Isochromes. If the chart is rotated about the WB axis, the solid formed is the Ostwald Colour Solid. Then colours which have constant white, black and colour content but differ in hue, are known as Isovalent colours.

The complete set of the twelve diamond-shaped charts that make up the colour solid is published in the form of an atlas. It is widely used for colour assessment, especially on the Continent.

5.2.2 The Munsell System

This is another widely used system (especially in the USA) which is usually published in the form of an atlas, for colour assessment. It is more extensive than the Ostwald system containing over 1200 different colour patches against 680 of the Ostwald system.

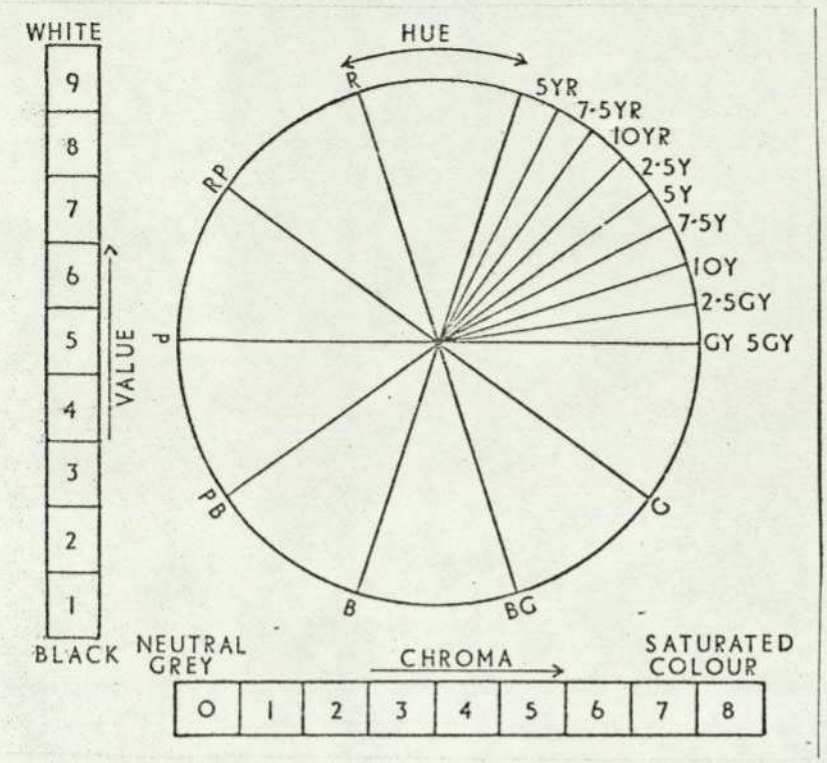


Fig. 5.4 Hue, value and chroma co-ordinates of the Munsell system.

As shown above, the three attributes used to define colour in the Munsell system are Hue, Value and Chroma. Value is the lightness of the sample, and its measure is given by the distance of a sample in the colour solid from black. (The central vertical axis denotes the neutral colours, with black at the bottom and white at the top.) Black is given the value of 0 and white 10. Chroma is the distance of a colour from the neutral axis. With grey having zero chroma, the chroma increases from the neutral axis to the circumference of the chart.

As also shown on the chart, differences in hue are represented by the different planes around the vertical axis. The whole solid is divided into ten equal vertical segments. For closer grading, each segment is divided into 10 further sections.

To completely specify a sample in the Munsell system, one writes the symbols for the hue and value then a dash and then the chroma. Thus for hue = 5 yellow, value = 6 and chroma = 8, one writes 5Y 6/8.

5.3 Colorimeters

5.3.1 Introduction

The systems already described, ie naming and colour atlasses, are basically a means of comparing and classifying colours. Colorimetry is a fundamental method of colour measurement and depends on the principles laid down by Grassmann (Chapter 3) viz. (i) any colour (excluding some spectral colours) can be matched by three selected beams of light, and (ii) two colours when matched in turn by two mixtures will, on mixing them together additively, be matched by the sum of their separate matching combinations.

This process of colour matching expresses an equivalence between the sensation produced by the colour and the sensation produced by the admixture of the selected light rays (the radiation stimuli). Thus, the colorimetric specification of a colour is a physical specification in terms of the light rays, the three stimuli, and not a specification of the physiological colour sensation on the retina of the eye.

Colorimeters are basically either of the additive type or the subtractive type. In the additive type a match is obtained between

the colour being measured and a known mixture of three primary lights. The measure of the colour is thus in terms of the amounts of the three primary lights that match the specimen colour.

The three primaries selected must be able to match white when added together (Chapter 3). Various colorimeters use primaries (normally blue, green and red) that differ slightly in terms of colour. A further criterion of the selection of the primaries is that the mixture of two of them should not match the third.

When a white beam of light is incident on a transparent sample, part of the light is absorbed. The amount transmitted can be matched with three primaries, thus effecting a colour assessment of the transparent sample. This system constitutes the basis of subtractive colorimetry. The Lovibond Tintometer offers the best example of a subtractive colorimeter.

Functionally, colorimeters are best divided into (i) The Visual and (ii) The Photoelectric.

5.3.2 Visual Colorimeters

These colorimeters are subjective in operation. The results obtained do not only vary from observer to observer but also on the psychological and physiological state even of a good observer. They are also very slow in operation. Their subjectiveness has at

times been the cause of argument and friction between supplier and customer of industrial products.

Visual additive colorimeters, such as those of Guild (1925) and Wright (1927) found tremendous use in establishing the chromaticity co-ordinates of the spectrum (Chapter 4). The Donaldson (1935) colorimeter, however, found its way in industry where it was used for a number of years, until reliable photoelectric colorimeters took over.

5.3.3 The Donaldson Colorimeter

Fig. 5.5 shows the colour-mixing system of the Donaldson colorimeter. The three filters (R,G,B) select the red, green and blue radiations. The three shutters T_1, T_2, T_3 control the area of each filter through which light passes, and thus controlling the amount of light in each beam. The positions of T_1, T_2 and T_3 when a match is obtained are related to the measure of the colour specimen.

One particular characteristic of this instrument is the integrating sphere, whose matt white-coated internal surface enables the three beams to undergo repeated diffuse reflections and hence uniform colour mixing.

The light leaving the integrating sphere is then incident onto a photometer prism for comparison with the light being matched.

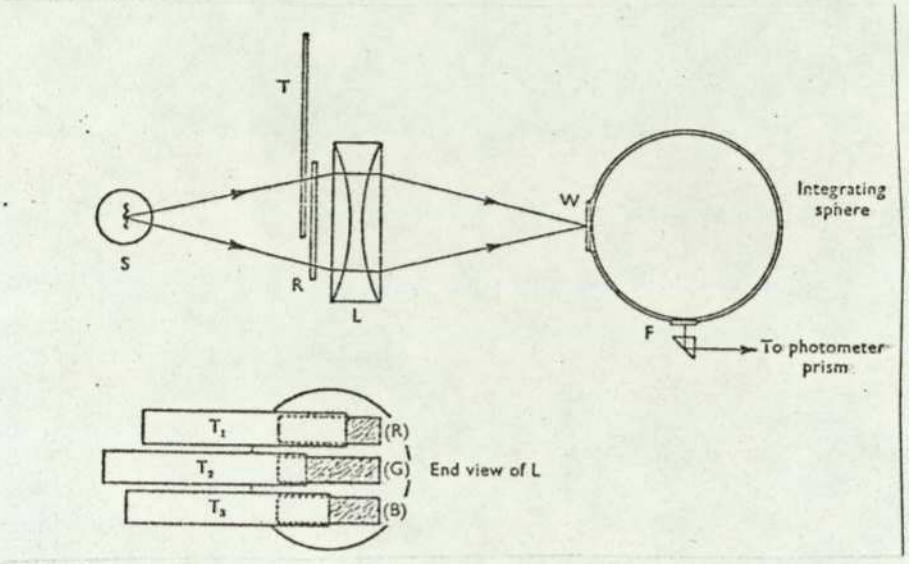


Fig. 5.5 The colour-mixing system of the Donaldson Colorimeter.

5.3.4 The Lovibond Tintometer

This is the most well known colorimeter acting on the subtractive principle. It consists of sets of red, yellow and blue glass slides which provide the primaries. The sample to be matched is placed in the instrument and white light reflected from a Magnesium oxide block is passed through the sample. The colour filter slides (which vary in steps of 0.1 Lovibond Units to normally 10 units of red, 70 units of yellow and 5 units of blue) are moved until a match is obtained. The criterion of these filters is that equal amounts of red, yellow and blue produce that amount of neutral density. Another set of neutral density filters is placed in the light path coming from the sample. This helps to achieve a lightness balance

in the case of bright samples. The Lovibond filters are very carefully prepared and standardised, and they are supposed not to deteriorate with time to any significant extent. This has been one of the reasons of the success of this instrument.

The Lovibond value (unit) has been established as the standard unit of colour specification in the oil industry (chemical and edible) for over one hundred years. In spite of its subjectiveness, the Lovibond value is the primary guide to the price of these oils. To eliminate the human error, the author has designed a photoelectric colorimeter (described in the next section) which measures the colour of refined oils in Lovibond units.

5.3.5 Photoelectric Colorimeters

The early photoelectric colorimeters were designed to measure the amount of light by the use of vacuum or gas photocells. Some even used photomultipliers. However, the most modern ones use selenium or silicon solid state photocells. In most photoelectric colorimeters, the colour sample is illuminated in succession by the three primary lights as obtained through standard filters and the light in each case measured by a single photocell. The calibration of the colorimeter is normally done by reading the photocell output for each filter, using a standard white in place of the sample. The principle of a small number of colorimeters is described below.

5.3.6 The Elrepho Colorimeter

This instrument is manufactured by the German firm Zeiss. The light from an incandescent lamp enters an integrating sphere horizontally, and after multiple reflections the sample which is placed at the base of the sphere is diffusely illuminated. The photocell is lying exactly on the top of the sphere so that the measurement is made in a direction normal to the sample surface. The three filters are manually positioned in turn to obtain the tristimulus values. These values are derived directly from the difference in the readings between the standard white and the sample, by measuring the photocurrent. This instrument has found wide use in the paper and textile industries. It is impossible to measure the tristimulus values in transmission of translucent samples.

5.3.7 The Colormaster

The Colormaster is made by Manufacturers Engineering and Equipment Corporation, Warrington, Pensylvania. A beam from an incandescent lamp illuminates the sample, whilst a second beam from the same source illuminates the standard. Two matched photoelectric cells are used to measure the amount of light reflected from each of the two surfaces. By varying the aperture in front of the reference photocell, it is possible to make the outputs of the two photocells equal. This equality of the two photocurrents can be observed by adjusting them to be in opposition so that when they are equal, a

balance is obtained, with the zero centred instrument meter showing zero. The position of the variable aperture as read on a micrometer scale provides a measure of the tristimulus value. The three tristimulus values can be obtained by making measurements in turn with the red, green and blue filters which are fixed on a rotatable disc controlled by a switch.

If at the position of the reference standard white another sample is placed, then the colour difference between the two colours is obtained. Thus, if in a particular process the aim is to obtain a certain colour identified by a colour standard, then at the end of the process it is necessary to check if there is any difference between the standard and the product. Then the standard and the product can be compared in the Colormaster and the difference in their tristimulus values obtained directly. This type of instrument is referred to as a Differential Colorimeter. The measurements obtained on a Differential Colorimeter are obviously much more accurate than absolute colour measurements.

5.3.8 The Color-Eye

This colorimeter is made by the Development Laboratories, Attleboro, Massachusetts. Its main characteristic is the large integrating sphere (450 mm diameter) by which both the standard and the reference are illuminated simultaneously. The light from the sample and the standard are focused onto a photomultiplier after passing through the selected filter. The two convergent light

beams are "chopped" by a rotating flicker mirror. Thus the two beams are focused onto the photomultiplier in turn. The difference in the photomultiplier currents produced by the two lights can be read on the instrument meter. Additional to the three filters for the tristimulus values, the filter wheel carries a further sixteen spectral filters. These sixteen filters enable the color-eye to be used as an abridged spectrophotometer. It is furthermore possible to obtain transmission values of a translucent sample by inserting this sample between the filters and the photomultiplier and by using a standard white in the normal sample position.

5.3.9 The Refined Oil Colorimeter

This colorimeter was designed by the author and manufactured by Cambridge Consultants for use of Unilever Associated Companies. It measures the redness and yellowness of an oil sample, by the use of a green and a blue filter respectively. As some of the refined oils are very pale indeed, any minute bubbles, striations, small dust particles in the sample, dirt on the side of the cell containing the sample etc, can cause attenuation of the light beam of the same order and even higher than that caused by the absorption of light due to the colour of the sample. To eliminate this error an Infra Red measurement (750 nm) is also made on the sample. As all refined oils transmit more or less equally in this region, any reduction in the beam intensity in this region is not associated with the colour of the oil, but only with the troublesome dirt, striations, bubbles etc. By dividing the transmission in the I.R.

by that in the blue region, the quotient is directly proportional to the yellowness of the sample. Similarly, by dividing the I.R. transmission value by the green transmission value, the redness of the sample can be obtained.

In construction the colorimeter consists of a tungsten hal ogen lamp producing a parallel beam of light which is passed through the sample; the beam is then intercepted by a rotating disc (150 rpm) carrying the three filters (Interference type) and then focused onto a Silicon (PIN type) photodiode. The filter that is crossing the beam at any one time is identified by three electronic gates consisting of light emitting diodes and minute silicon photocells. The output of the photodiode is amplified logarithmically and thus two electronic subtractions can be made to effect the divisions.

This instrument has been calibrated so that the yellowness and redness of the samples are read directly in Lovibond yellow and red values. The accuracy obtained with such a calibration is much better than that normally obtained with average operators. The repeatability of the instrument, however, is excellent.

It has found wide acceptance within a large number of oil refining concerns, both in the laboratory and on the production area. Tintometer Ltd of Salisbury are now marketing this instrument worldwide.

Fig. 5.6 and 5.7 show photographs of the instrument.

Fig. 5.6 The Refined Oil Colorimeter.

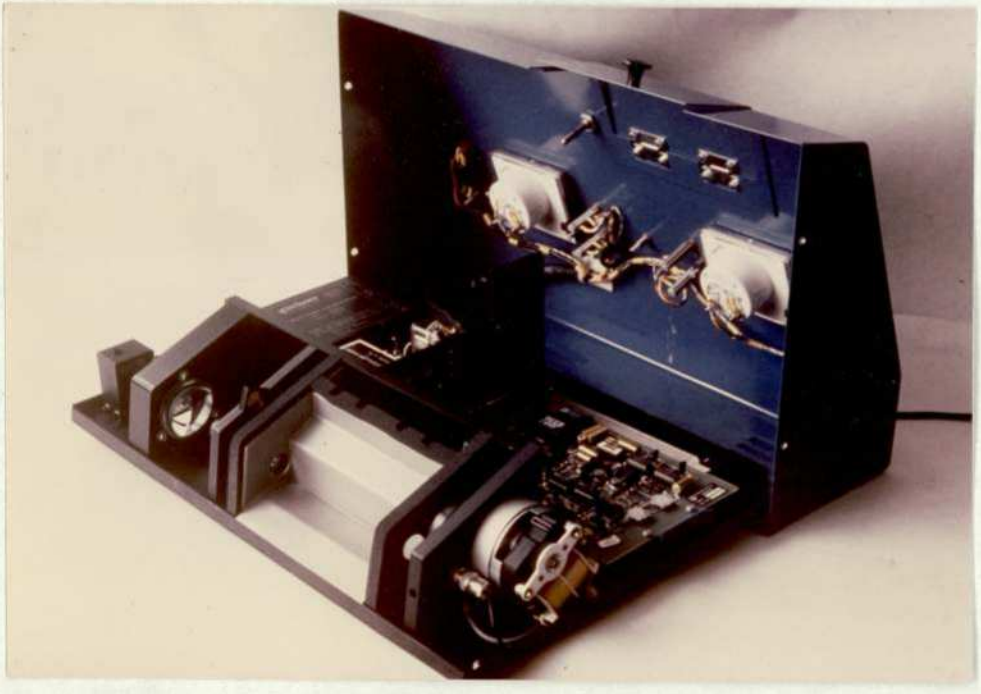


Fig. 5.7 The interior of the Refined Oil Colorimeter.

5.4 Spectrophotometers

The spectrophotometer is another basic instrument that is becoming of increasing use in colour measurement. The measurement is in the form of a graph, the curve of which indicates the proportion of incident light that is reflected or transmitted by the sample throughout the spectrum. This graph provides comprehensive information about the spectral characteristics of the sample and can be used to evaluate the C.I.E. Tristimulus values X, Y and Z. The first spectrophotometers were of the visual type, but these lacked in accuracy, observations were lengthy and fatiguing and have now been replaced by photoelectric instruments. Fig. 5.8 shows in diagrammatic form the principle of a spectrophotometer.

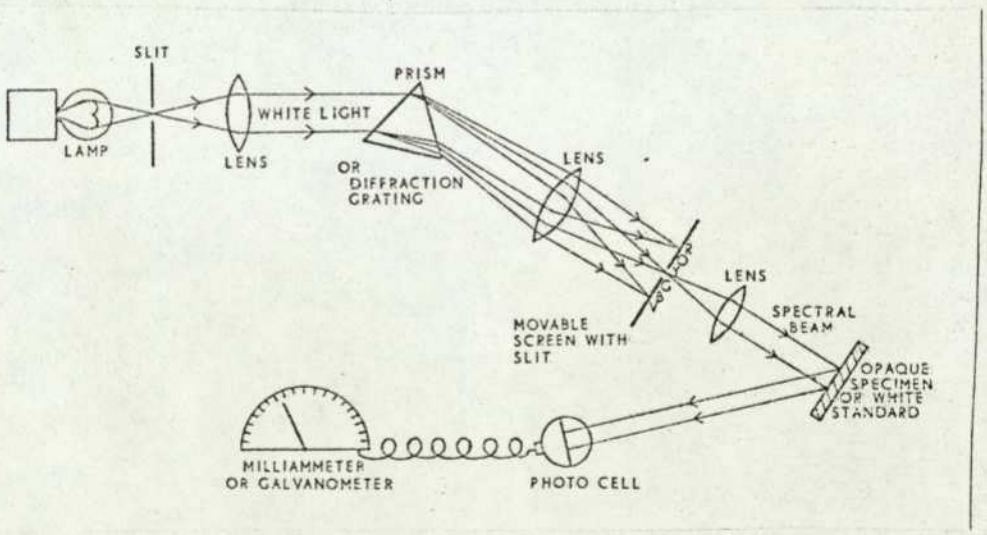


Fig. 5.8 The principle of a spectrophotometer.

Recording spectrophotometers have been designed which automatically record the spectral characteristics of the sample. This is achieved by synchronising the electrical, mechanical and optical systems of the spectrophotometer. Recording spectrophotometers are

rapidly replacing ordinary spectrophotometers, especially for industrial applications, mainly because of speed considerations.

The most common recording spectrophotometer in the United Kingdom is the Unicam SP700. This instrument is primarily intended for Analytical and Chemical research spectrophotometry on transparent solutions and vapours. It can also be used for reflectance measurements by the use of a special accessory into which the opaque sample and white reference surface are inserted. A pure silica prism provides a wide spectral range from 190 nm to 3570 nm. The narrow waveband isolated by means of slits is split into two beams to illuminate or irradiate the sample and standard, and the radiation transmitted or reflected by them is focused by mirrors onto a photomultiplier up to 700 nm and at longer wavelengths onto a lead sulphide photoelectric cell. The signals from the detector are amplified and automatically switched between sample and reference standard, the reference signal being maintained approximately constant. The output signals are fed through a precisely adjusted electrical system, situated between the controls on the front operating panel, to the chart recorder above the panel which plots the ratio of the sample transmission or reflection to that of the standard.

Where an approximate spectrophotometric curve will suffice, an abridged spectrophotometer can be used. These instruments consist of a rotating disc carrying a number of narrow band filters (normally eight to sixteen). The filters are selected so that the

series covers the spectrum in a succession of narrow zones or intervals, each filter transmitting the light over one zone of perhaps 30 nm between 400 nm to 700 nm.

Fig. 5.9 shows the principle of abridged spectrophotometers for reflection and transmission measurements.

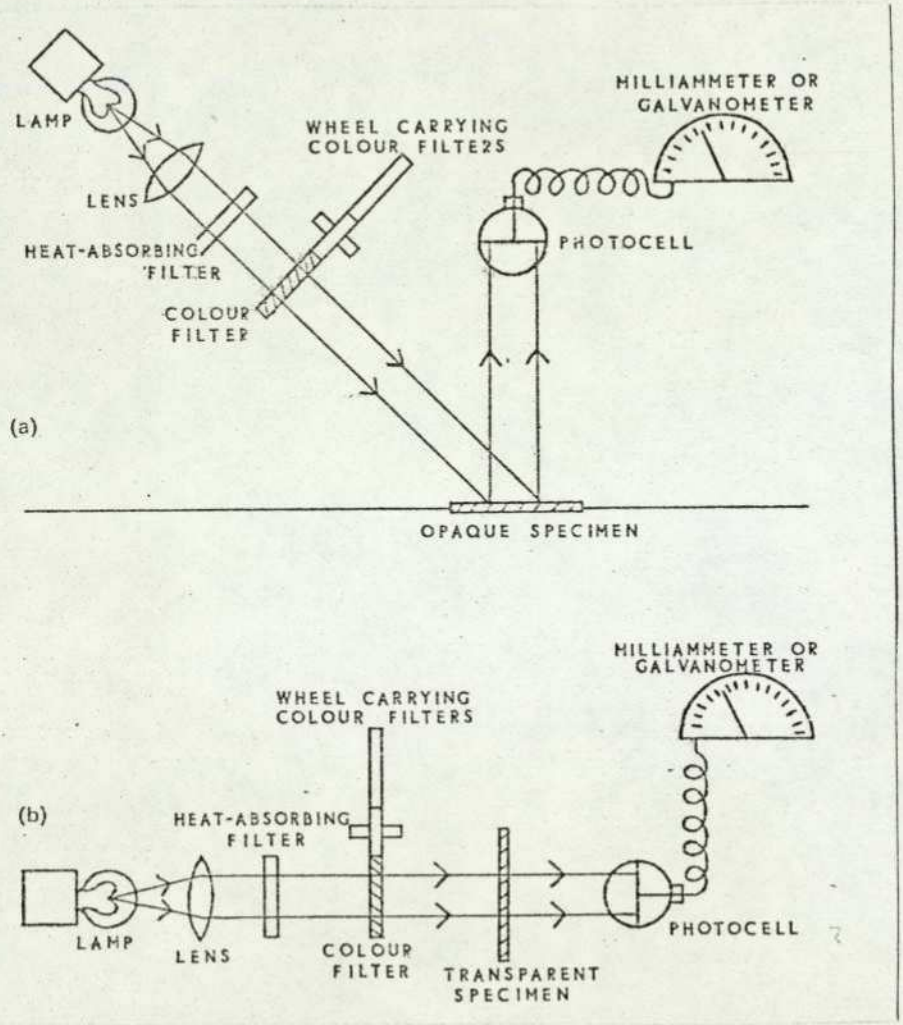


Fig. 5.9 The principle of abridged spectrophotometers.

Measurements are made in turn through each filter. In the simpler instruments a standard white surface is first introduced in place of the specimen and the meter reading adjusted to 100 for the

respective filter, so that a percentage colour reflection or transmission reading is obtained for each measurement on the specimen. The results are then plotted graphically in accordance, on the wavelength axis, with the peak of the transmission band for every filter.

One of the best known abridged spectrophotometers is the Pretema FS-3A Spectromat. It incorporates a large integrating sphere for illuminating the sample and standard. A platform below the sphere holds the sample and the white reference standard against an aperture which can be varied between 5 mm and 50 mm in diameter.

The rotating disc carries 33 narrow-band filters making measurements between 390 and 710 nm at 10 nm intervals. The instrument is fully automatic in operation after selecting by push buttons the reflectance or appropriate C.I.E. X, Y or Z function and the required standard illuminant and pressing a starting button. The Spectromat both displays and prints the reflectance and tristimulus values. It can furthermore be linked with a computer to solve problems concerning colour matching formulations, corrections and colour differences.

Computers are increasingly being used for evaluating colour calculations. A number of spectrophotometers have outputs which can be directly fed into a computer (binary coded decimal, offset binary etc). A large saving in time can thus be effected. These combinations can, however, be prohibitively expensive.

CHAPTER 6

WHY ADD ANOTHER INSTRUMENT TO THE ALREADY LARGE RANGE?

6.1 Replacing Colorimeters

As already discussed in Chapter 5, a photoelectric tristimulus colorimeter is a compromise between a spectrophotometer-computer combination and other methods, eg visual methods. Photoelectric colorimeters normally need to be set up before measurement, they are cumbersome to use and cannot account for metamerism.

Tristimulus colorimeters may be made to be very precise at times, but their accuracy will always be limited by the impossible task of designing a source-filter-detector combination to exactly match the C.I.E. weighting functions.

Thus, consider a tristimulus colorimeter with the wrong combination of source-filter-detector, ie not matching the C.I.E. weighting functions, to be used for the measurement of two colours A and B as shown in Fig. 6.1. This diagram shows the transmission

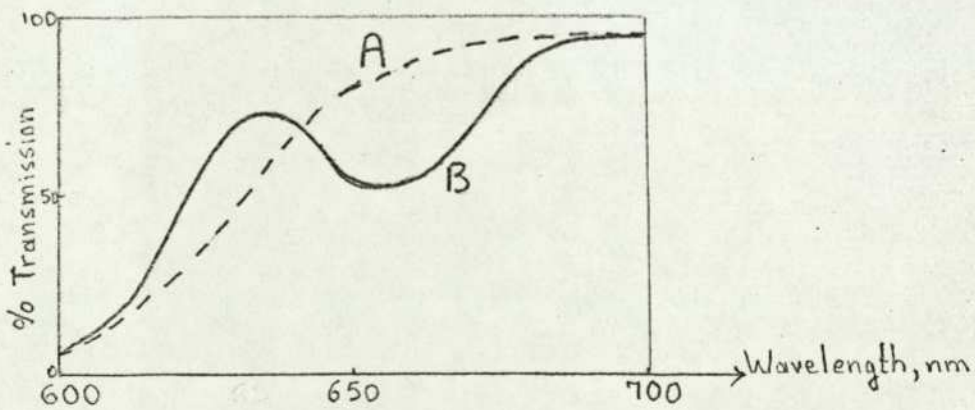


Fig. 6.1 Transmission spectra of typical crude oils.

characteristics of two crude oils (A being a fish oil and B being a palm oil). It is quite possible for two such oils to have very similar transmission spectra between 400 nm and 600 nm. A common deviation in the 600 nm to 700 nm region is as shown. The absorption peak shown in sample B is due to the presence of chlorophyl in the crude palm oil. Sometimes other absorption peaks, such as those due to carrotene, also exist within the 600 nm to 700 nm region.

It is not unusual for the integrated transmission of such two samples, when multiplied by the response function of a tristimulus colorimeter with unmatched filters, to be near enough the same. Thus, the readings obtained on a "C.I.E." tristimulus colorimeter will be the same for both samples.

But the two samples mentioned, appear quite different, A being a bright orange-red and B being a dull brown-red. If the C.I.E. values were worked out by considering the percentage transmission at wavelength intervals of say 10 nm, then the tristimulus values obtained would be quite different.

Thus, we have a "C.I.E." tristimulus colorimeter that cannot measure the C.I.E. tristimulus values. This is a very serious limitation of the tristimulus colorimeter, and this is one of the main reasons why these colorimeters have not been accepted by the oil industry in general.

The oil industry is in desperate need of a colorimeter to replace the presently used visual method (ie the Lovibond Tintometer). As the price of oils is so much dependent on colour, millions of pounds are involved every year in the accurate specification of their colour. One would obviously have thought that the price of such a colorimeter would have been an irrelevant factor. This is probably so for a number of large industrial concerns. But the fact is that many organisations cannot afford to pay high prices for new instruments.

As most of the oil business is conducted between a large industrial concern and a number of smaller ones, there must be a common method of specification. Thus, it is not possible for a large organisation to change over from a Lovibond system to a spectrophotometer-computer combination and start specifying its oil products in an entirely different system such as the C.I.E. If any changes are made, they must be made by all the parties concerned.

Therefore business carries on with a system that was introduced over a hundred years ago, in spite of the subjectiveness of the method and the consequent injustice, dispute and controversy that goes with it.

A colorimeter to replace the Lovibond Tintometer should be (a) much cheaper than a spectrophotometer-computer combination; (b) it should be easy to use; (c) its system of specification should be

universally acceptable; (d) should not need frequent maintenance; (e) preferably with no moving parts.

The present spectrocolorimeter can be used as an instantaneous, continuously reading spectrophotometer whose thirty transmission values can be fed into a computer and a transformation to Lovibond values thus obtained. A program in Basic has been written for this transformation (see Appendix). This facility should prove very useful in a transition stage for large industrial concerns, where some of their suppliers or customers use the C.I.E. and some use the Lovibond system.

Thus to summarise: a tristimulus photoelectric colorimeter cannot be relied upon to give an accurate colour measurement. It is only a compromise between the ideal spectrophotometer-computer combination and other visual methods. Even so, in some industries such as the oil industry the tristimulus colorimeter has been completely rejected.

It is expected that the spectrocolorimeter would be an acceptable alternative. The instantaneous display of the colour values by the spectrocolorimeter is an added bonus.

6.2 Replacing Spectrophotometers for Colour Measurement

It has always been considered that any important colour measurement should involve a spectrophotometer for obtaining the optical

spectrum (either transmission or reflection) of the object concerned. As already mentioned in Chapter 5, a number of reliable, precise, accurate, adaptable, fast spectrophotometers exist on the market. They are using some of the most ingenious mechanisms to produce the necessary scanning, and some very advanced synchronising techniques for signal recovery and recording.

But all these sophisticated designs carry with them a number of unwanted features. First, their price is so high that very few laboratories indeed can justify the acquisition of such expensive an item for colour measurement alone. Most chemical laboratories justify a spectrophotometer mainly for its usefulness in chemical analysis (mainly in the U.V. and I.R.) rather than for colour measurement.

Also, these sophisticated machines require expensive maintenance and repair. This could sometimes be a more serious problem than price.

Furthermore, the spectrophotometric plot on its own is seldom of much use. To get meaningful interpretation of such a trace a long evaluation is necessary. Even with a modern electronic calculator this process is very time-consuming and always subject to human errors in entering the values into the calculator.

A spectrophotometer is therefore nowadays often connected directly to a computer. The colour values can therefore be printed or

displayed onto a Visual Display Unit (VDU) seconds after the completion of the spectrophotometer scanning.

To scan the complete visible spectrum and take readings at 10 nm intervals (a total of thirty-one readings) a very fast modern computer would take about thirty seconds. This time factor is seldom of any signficant consequence in any colour laboratory. It is enough, however, to make it completely unsuitable for most on-line applications.

The optical spectrum (as explained in Chapter 5) can also be used by the computer to test whether the sample under consideration lies within a preset range of colour acceptability. If it does not, then the computer can work out the dye formulation to bring the colour of the sample within the area of acceptability.

To summarise: the spectrophotometer-computer combination is very adaptable, useful and successful.

The spectrophotometer part of the combination does, however, have several disadvantages: it is expensive; difficult to maintain and repair; liable to wear because of moving parts; bulky but delicate; requires mathematical manipulation of the output; though not slow for laboratory use it cannot be used for on-line applications.

Using the thirty outputs of the new spectrocolorimeter to feed into the computer provides an instantaneous and continuous evaluation of the sample being tested. The analogue electronic circuits of the spectrocolorimeter can convert the spectral values to the C.I.E. tristimulus values and/or the chromaticity co-ordinates instantaneously.

Thus, for simple colour evaluation the spectrocolorimeter can replace both the spectrophotometer and the computer; it can produce the answer much quicker.

For more elaborate colour evaluation, such as testing for colour acceptability and dye formulation, the spectrocolorimeter can replace the spectrophotometer, and also obtain the spectrum much quicker and much more simply - a parallel input multiplexer can be used to feed the values into the computer instead of a series input multiplexer which is needed with a normal spectrophotometer.

The new spectrocolorimeter, in addition to being much cheaper, is evidently a more robust instrument and much easier to use than a spectrophotometer.

Considering all these advantages, the spectrocolorimeter should have no problem in replacing the spectrophotometer from nearly all colour measuring applications.

6.3 In-line Colour Control - New Horizons for Industrial Colorimetry

When the spectrocolorimeter is used as a laboratory instrument, the fact that it can measure colour instantaneously is probably not of much significance. This property of the instrument does, however, provide the ability for in-line colour control.

It is normal for in-line densitometers (measuring the absorption of light at one wavelength band only) to be wrongly called colorimeters. This type of instrument is very common for colour control in a number of industries. But as it is well known, such a device cannot measure colour to any degree of accuracy. It can just give an indication of the state of a particular product. The reading of such an instrument is only meaningful for one type of product only. It can be used, for example, to indicate the "yellowness" of beer by a single measurement in the blue region. This "yellowness" value will be meaningless though when applied to another liquid which has dissimilar optical characteristics to those of the beer for which the instrument has been calibrated.

Such "colorimeters" are used in-line in the edible and chemical oil industries to ensure the clarity of refined oils. If the clarity of the oil being refined is below the preset acceptable value a signal is given and either an alarm rings or a valve opens and the oil rerouted. This type of "colorimeter" is not, however, suitable for use for the measurement of colour of crude oils, due to the

large variations that exist in their transmission spectra, as explained in Section 6.1.

The true in-line measurement of the colour of oils in the refining stage, rather than in the laboratory, could be of tremendous advantage to the oil industry in general. The refining processes of oils have been mathematically analysed by Chemical Engineers and for a given set of conditions, it is possible to estimate the change of colour that will occur at a refining process or by the addition of a colour absorbing medium such as Fuller's Earth.

Furthermore, by knowing precisely the colour of a crude or partly refined oil at one stage in the refining process, it is possible to work out the most economical route to follow (ie temperatures, amounts of Fuller's Earth to be added etc) to obtain a refined oil within a predefined colour range.

This problem can be compared with the dye formulation problem. The basic difference is that whilst in the dye formulation problem the main criterion is the appearance of the finished product with the cost of the dyes being of secondary importance. In the case of oils the refining process and hence cost is the predominant factor; as long as the colour of oil is lighter than a certain value, it will sell at the quoted price corresponding to that colour value.

This system will result in improved processing which will save a considerable amount of energy. This is a matter which is becoming increasingly important since energy is becoming a very expensive commodity.

This is only one typical example where the spectrocolorimeter can be applied to previously unexploited fields.

6.4 Other Applications

The versatility of the spectrocolorimeter is such as to enable it to do what most other colorimeters can do, and in most cases more accurately, more quickly and with less complexity.

A typical application is the analysis of the spectral distribution of light sources. All that is necessary to do is to converge a beam of light from the light source onto the entrance slit of the spectrocolorimeter. The spectrum is instantaneously available.

A fibre optics system can be used for this type of application. Optical fibres can also be used to analyse any light emitted from any surface that is not easily accessible. Account has to be taken of the absorption coefficient of the fibres, which is not uniform throughout the visible spectrum.

Another possible application is the incorporation of a telescopic lens to focus the image of large objects at a distance, such as the

walls of buildings, onto the slit to produce a telecolorimeter. Colour evaluation is obviously instantaneous. As no light source is necessary for such measurements, two dry battery cells can supply all the necessary power for the electronics. Thus the spectrocolorimeter can be easily transferred and operated in any desired location. No setting up procedure is necessary and its robustness allows it to be moved around and used in any position (ie, not necessarily horizontal). It could actually be focused on different objects whilst the operator carries it in his own hands, like a cinecamera, allowing quick comparison of the colour of a number of distant objects.

6.5 Conclusions

The spectrocolorimeter fills a definite gap in the colour measurement technology, both from the technical and user's points of view. It has been designed with specific applications in mind, but it should be welcomed by a number of industries.

CHAPTER 7

EVOLUTION OF THE SPECTROCOLORIMETER

7.1 The Light Source

The recent advances made in the tungsten halogen lamp made it more or less the obvious choice as a light source. The most important characteristics of this lamp that make it so appropriate for this application are as follows:

(a) The High Temperature that can be obtained

Normal tungsten lamps are run at around 2700°C. This is a high enough temperature to provide enough visible radiation for most applications. A slight increase in the filament temperature results in a higher rate of evaporation of the tungsten filament (and deposition of the tungsten molecules onto the glass envelope). This causes a large reduction in the useful life of the normal tungsten lamp. The relationship between temperature and lamp life (see Fig. 7.1) follows a power law. The life of a normal lamp at 2700°C is about 2000 hours. At 3000°C it has a life of just a few hours.

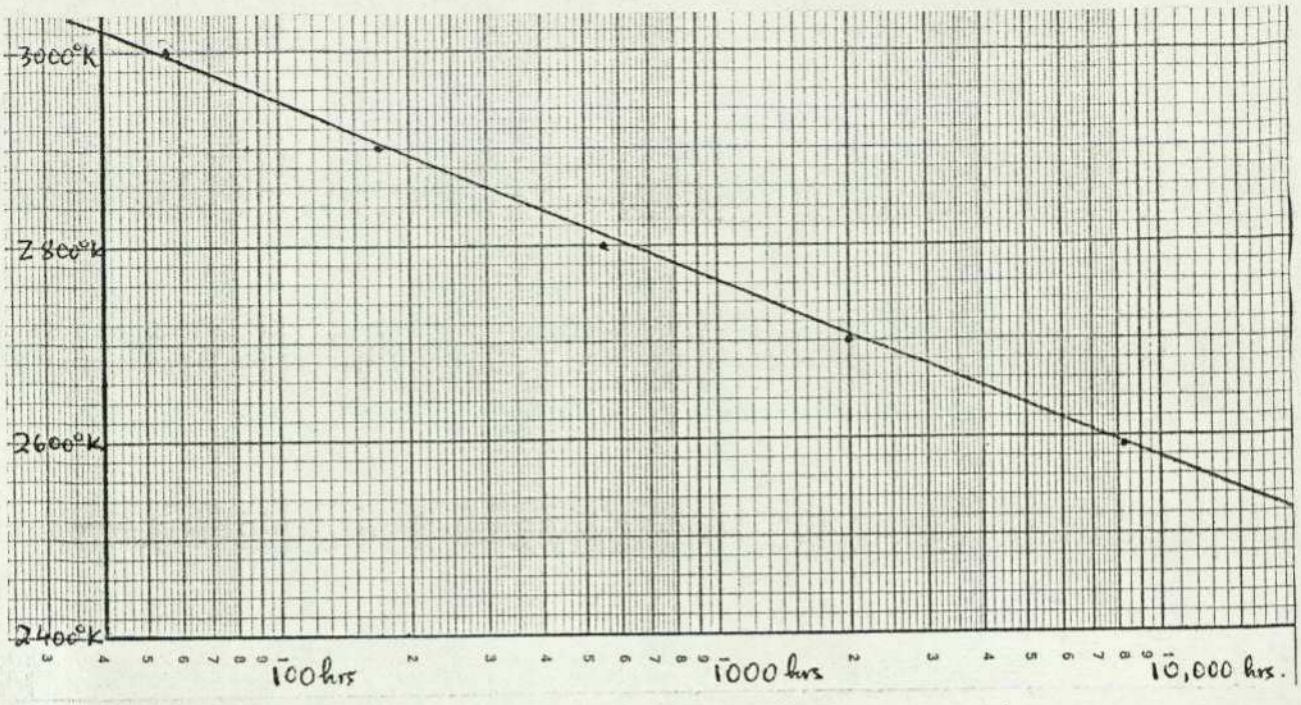


Fig. 7.1 Incandescent lamp characteristics.

The introduction of a Halogen, such as Iodine, in the lamp causes the evaporated tungsten molecules to chemically combine with the iodine molecules to form tungsten iodide. This compound is quite stable below a certain temperature (around 2500°C). The lamp is constructed in such a way that the bulb wall temperature is always above 250°C. As this temperature is above the condensation temperature of the tungsten iodide, it cannot deposit on the bulb wall but instead is carried back to the vicinity of the filament. The tungsten iodide is then broken down into tungsten and iodine by the higher temperature of the filament, so that the tungsten is deposited back onto the filament whilst the iodine is released to repeat the regenerative cycle. This regenerative cycle enables the lamp to run at a temperature as high as 3200°C with a life of 1000 hours.

The high temperature of the lamp is especially useful since it provides a higher output in the blue and violet regions of the spectrum, where most photodetectors have a lower sensitivity, and most light is lost by scattering within the instrument.

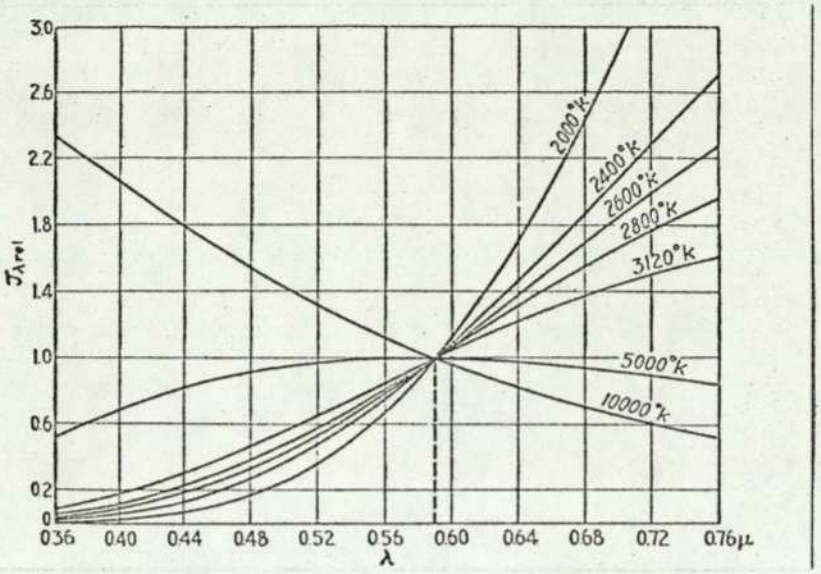


Fig. 7.2 Spectral characteristics of incandescent lamps.

(b) The Small Filament Size

The very small physical size of the filament is especially convenient for concentrating onto a slit or a pinhole. This is very useful for obtaining the parallel beam necessary for the diffraction grating. The small dimensions allow the production of equipment of compact design.

(c) Constant Light Output

As no building up of deposit of tungsten takes place on the glass envelope, the light output of the tungsten halogen lamp remains constant throughout its life.

(d) The quartz bulb gives a high degree of resistance to thermal and mechanical shock. Other light sources have severe limitations either due to their selective emission spectra or due to their large size which makes it very difficult to design optical systems which utilise parallel beams. The lamp eventually chosen and used on the final version is a 100 watt 12 volt (Atlas M28).

7.2 The Diffraction Grating

A diffraction grating was chosen in preference to a prism because of the linearity of the spectrum in terms of wavelength obtained with a diffraction grating.

The angular dispersion of a prism is given by,

$$\frac{d\theta}{d\lambda} = -\frac{2t.}{a} \cdot \frac{B}{\lambda^3}$$

where t = length of the prism base

a = cross section of the emergent beam

B = a constant (from Cauchy's formula)

$d\theta/d\lambda$ = angular dispersion

λ = wavelength

Thus the dispersion is a function of wavelength, being greater at the blue end of the spectrum. The negative sign indicates that θ increases as λ decreases.

The angular dispersion of a diffraction grating is given by

$$\frac{d\theta}{d\lambda} = \frac{p}{d \cos \theta_p} \qquad \text{for order p.}$$

and

$$\frac{d\theta}{d\lambda} = \frac{1}{d \cos \theta} \qquad \text{for the first order}$$

where d = distance between slits

θ = angle of diffraction

Thus the angular dispersion is inversely proportional to the grating spacing, which is constant for any given grating.

Furthermore, the angular separation of the lines of wavelengths λ and $\lambda+\delta\lambda$ is proportional to the order p, ie the separation of neighbouring lines in the second order spectrum is twice that in the first.

For lines lying sufficiently close to the axis, $\cos\theta = 1$ can be justified, and the dispersion is then constant throughout the relevant part of the spectrum. If, in Fig. 7.3, the linear dispersion in the focal plane of the lens L3 is $dx/d\lambda$ then $x = \theta f$ where f is the focal length of L3. Thus, for the first order and for small θ,

$$\frac{dx}{d\lambda} = \frac{f}{d}$$

or

$$x = \frac{f}{d} \lambda$$

ie, the distance from the zero order image to the spectral line for wavelength is directly proportional to f.

The Diffraction Grating chosen is a standard reflection grating manufactured by Diffraction Gratings and Optics Ltd. The ruled area is 25 mm x 25 mm. The ruling is 1200 lines/mm. Thus in the first order it gives a chromatic resolving power ($\lambda/\delta\lambda$) of 25 x 1200 = 60,000. The resolution of this grating is 0.0067 nm in the violet end of the visible spectrum and 0.0167 nm in the red end.

7.3 The Three Photoconductive Cells Spectrometer

It was thought wise that in the original prototype stages as few variables as possible should be involved. Therefore the first array consisted of only three photocells. As the whole idea ultimately depended on the photodetectors to be used, an exhaustive investigation of the market was undertaken to determine the most suitable photodetector to use.

The Cadmium sulphide photoconductive cell type RPY58 made by Mullard was originally chosen because of its extremely high sensitivity and its useful geometrical configuration that enabled the easy mounting in the form of a linear array.

The layout was as shown in Fig. 7.3.

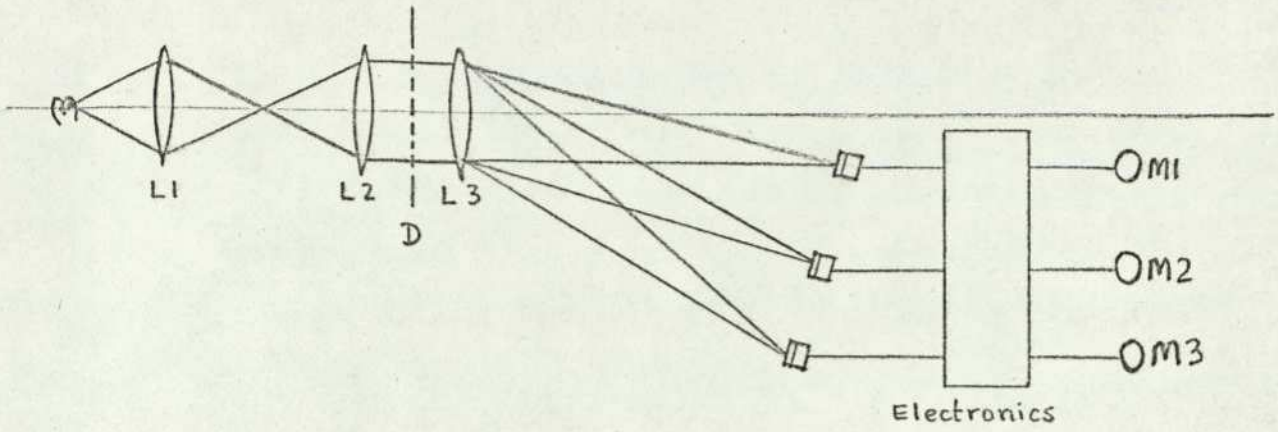


Fig. 7.3 The three cell spectrometer.

The following electronic circuit (Fig. 7.4) was used to amplify and measure the light signal at the three parts of the spectrum:

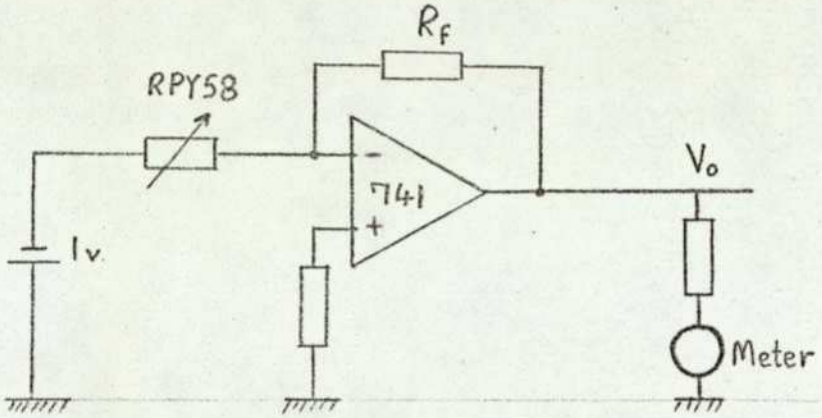
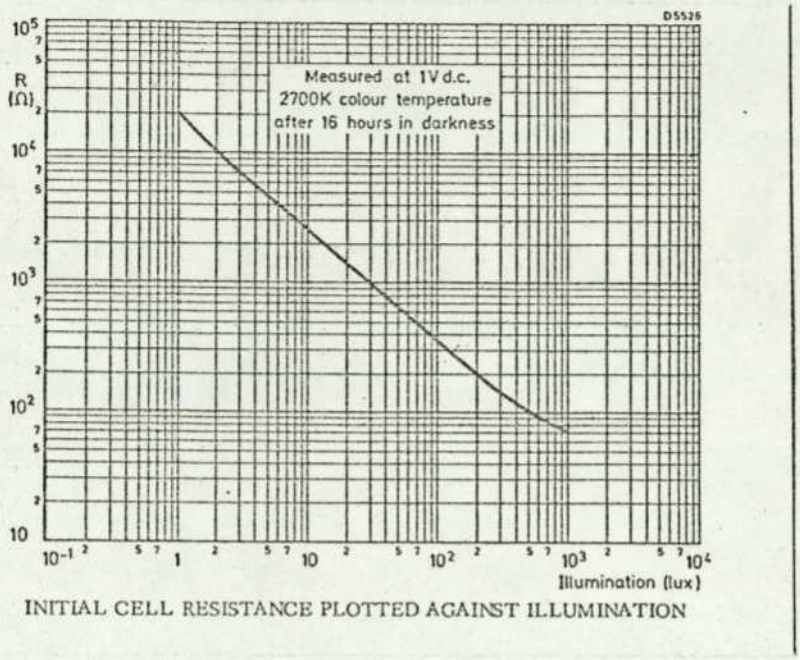


Fig. 7.4 Circuit used with photoconductive cells.

The 741 amplifier is a general purpose operational amplifier. The offset voltage and offset current are guaranteed over the entire common mode range. The amplifier also offers many features which make its application nearly foolproof: overload protection on the input and output, no latch-up when the common mode range is exceeded, as well as freedom from oscillations. Its schematic and connection diagrams are shown in Fig. 7.5. Its electrical characteristics are shown in table 7.1.

Table 7.1. Characteristics of the 741 Amplifier (p.114) has been removed for copyright reasons

The photocell resistance - illumination characteristics for the RPY58 are shown in Fig. 7.6.



INITIAL CELL RESISTANCE PLOTTED AGAINST ILLUMINATION

Fig. 7.6 Response characteristics of the RPY58.

The output voltage of the circuit shown in Fig. 7.2 is given by,

$$V_o = -V_{in} \times \frac{Rf}{Rc}$$

From Fig. 7.4 the illumination I is proportional to $\frac{1}{R_c}$

Thus $V_o = -V_{in} \times R_f \times I$ ie, the output voltage is directly proportional to the illumination. This arrangement is, therefore, suitable for measuring directly the percentage reduction of the intensity after setting the original voltage to be 100%. This configuration was a good starting point for optimising the optical components. The physical layout was mounted on an optical bench in a dark room. This enabled a number of variations to be done quickly.

It was soon realised, however, that the photocell output was drifting with time. After careful reassessment of the situation, which involved long discussions with the photocell manufacturers (Mullard), it was decided to abandon the RPY58 photocells in favour of the RPY71 which had just come on the market.

7.4 The Twelve Photocells Spectrometer

The new RPY71 photocell proved to be much more stable than the RPY58. It also had a more linear response characteristic at the lower illumination levels. But, more important, it had improved sensitivity in the Blue region of the spectrum where the lamp output is also much lower than in the rest of the spectrum. This last characteristic meant lower gains for the amplifiers in the blue region, which was a great advantage. The relative responses are shown in Fig. 7.7.

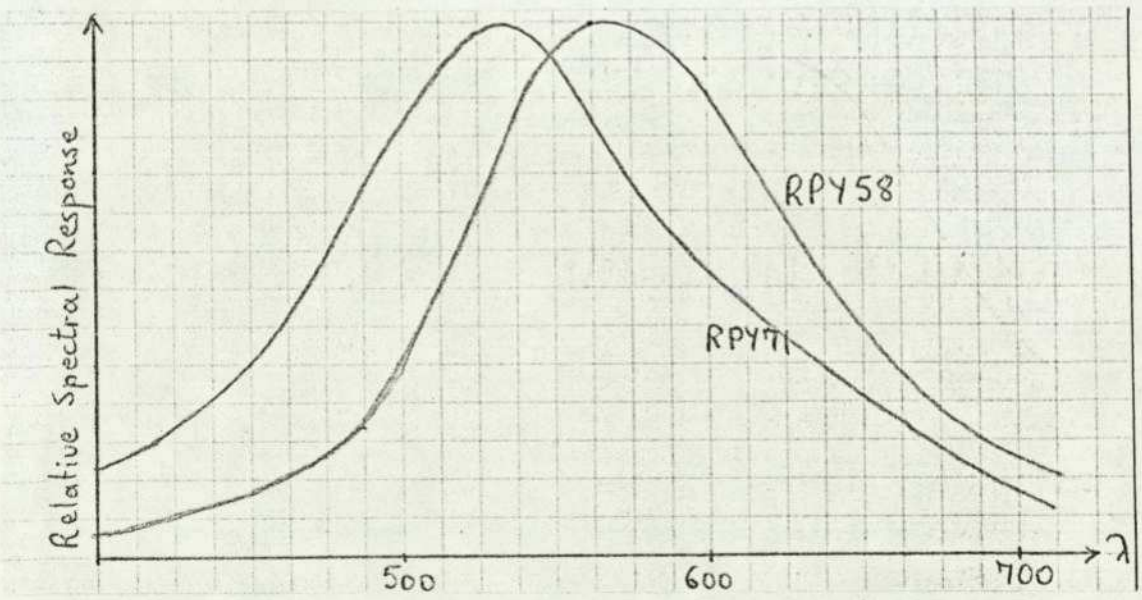


Fig. 7.7 Comparison of the responses of the RPY58 and the RPY71.

It was thought that enough knowledge had been gained with the three photocell model, to proceed to a twelve photocell model. This model would enable comparisons of the relative stabilities of a larger number of photocells.

The same electronic circuit was used as that for the three photocell arrangement. It just comprised of twelve amplifiers the outputs of which were in turn read on a single meter by the use of a rotary switch.

It soon became obvious, however, that even though these photocells were much more reliable than the RPY58, they were not stable enough to use in an instrument such as the one contemplated. The output of the RPY71 photocells was still varying with time, even after the initial fast change (time response). They also showed a hysteresis effect (sometimes called a "history" effect) whereby the reading at any one instant depended on the previous light exposure of the photocell.

The output of the photocells was also strongly influenced by the ambient temperature.

Having reached this stage, it became evident that a reappraisal of the photodetector situation was vital for the continuation of the whole project. Rather than make another arbitrary search for alternative photodetectors, it was considered that a full

theoretical and systematic investigation of all the available types of solid state photodetectors was necessary.

This investigation took six months to complete and proved to be a turning point in the basic design of the proposed colorimeter. See monograph inset at the end.

As a direct result of this investigation, it was decided to use p-n junction silicon photodiodes instead of the very sensitive photoconductive photocells. The Mullard PBX40 type was chosen as the most appropriate to use.

7.5 The Design of the Spectrometer Using Thirty Silicon Photodiodes

The illumination-current characteristics of the pn junction BPX40 silicon photodiode are as shown in Fig. 7.8.

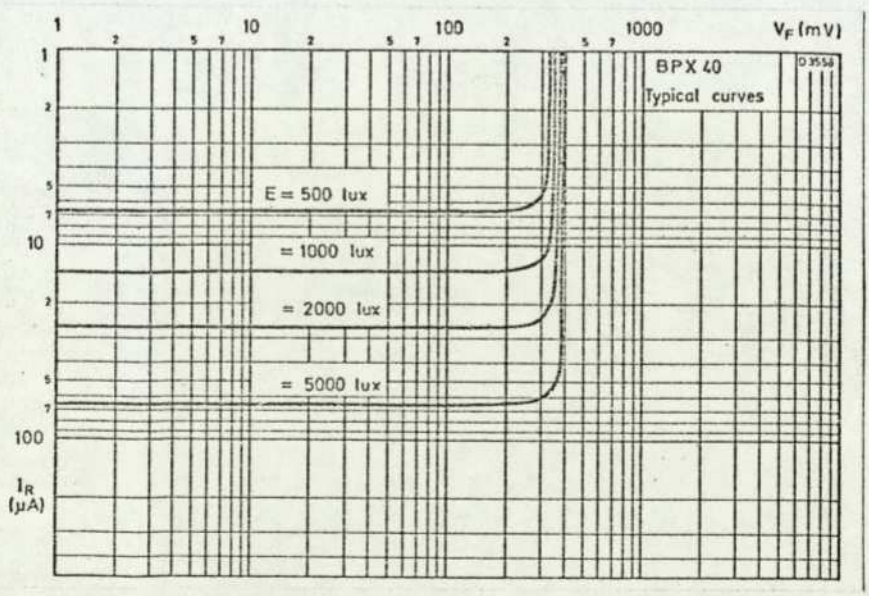


Fig. 7.8 Characteristics of PBX40 used in the photovoltaic mode.

The most stable mode of operation to use is with R=0, ie with zero load, where the current generated at the pn junction is directly proportional to the illumination. The spectral characteristics of the BPX40 are shown in Fig. 7.9. The electronic circuit for this mode of operation is as shown if Fig 7.10.

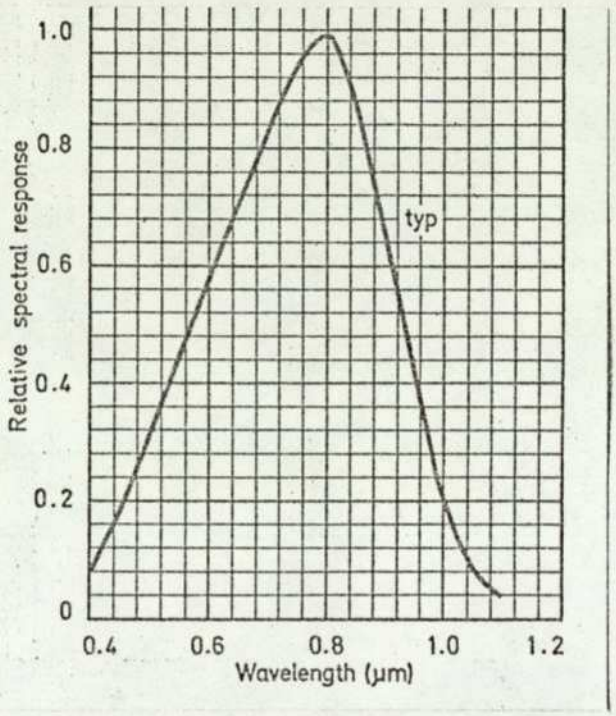


Fig. 7.9 Spectral characteristics of the BPX40.

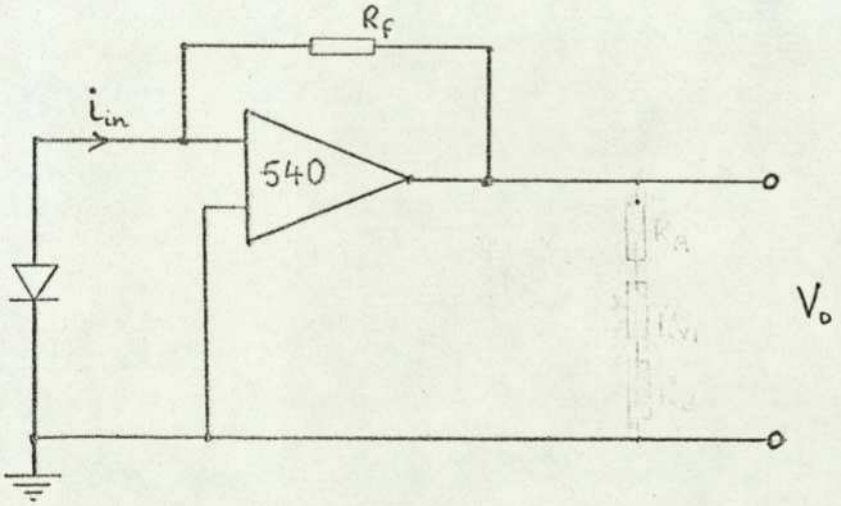


Fig. 7.10 Circuit for current output amplification.

The output voltage, V_o, of this amplifier is given by,

$$V_o = -i_{in} R_f$$

To achieve this condition, ie no load presented to the photocell, the input impedance of the amplifier used must be infinite. In practice an FET (Field Effect Transistor) input amplifier is used whose input impedance is of the order of 10^{12} ohms. For this particular circuit the type AD540J was selected because its characteristics were ideally suited for this application, and also because of its moderate price (a direct result of its popularity). Its electrical characteristics are shown in table 7.2.

The thirty photocell array was then built. Three circuit boards, each carrying ten AD540J amplifiers, were then constructed. Each amplifier was also equipped with a potentiometer to vary its gain and a second potentiometer to adjust for the zero input bias current. The final electronic circuit for each photodiode is as shown in Fig. 7.11.

Table 7.2. Characteristics of the 540 Amplifier (p. 121) has been removed for copyright reasons

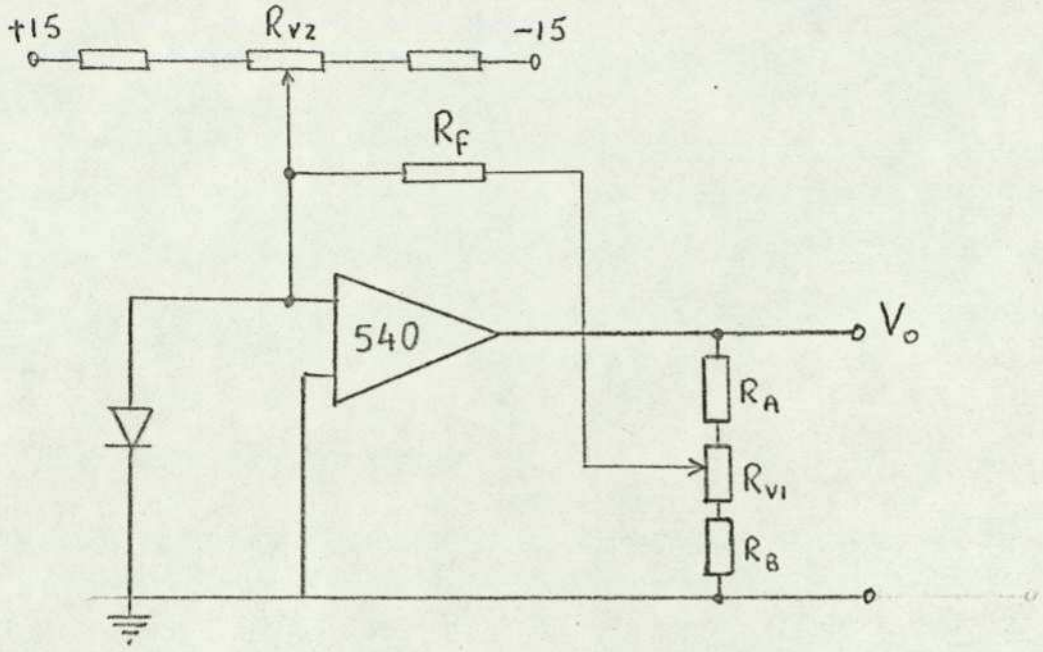


Fig. 7.11 The final design of the photocurrent amplifier.

The value of the gain as decided by R_F, R_A, R_B, and the position of the potentiometer R_{vi}, was arranged so that the output of every amplifier was 10 volts. This obviously meant a much higher gain for the photocells in the blue region than those in the red region. This is because of the lower lamp output in this region and also because of the lower sensitivity of the photocells in this region.

With the output of the amplifier set to 10 volts, a reduction of the light intensity by 1% would cause the output to change by 100mV. As the resolution of the digital panel meters used was 10mV, the instrument resolution when used as a spectrophotometer was 0.1%. The output of each one of the thirty amplifiers was then fed into three further amplifiers as shown in Fig. 7.12, making an array of 90 amplifiers.

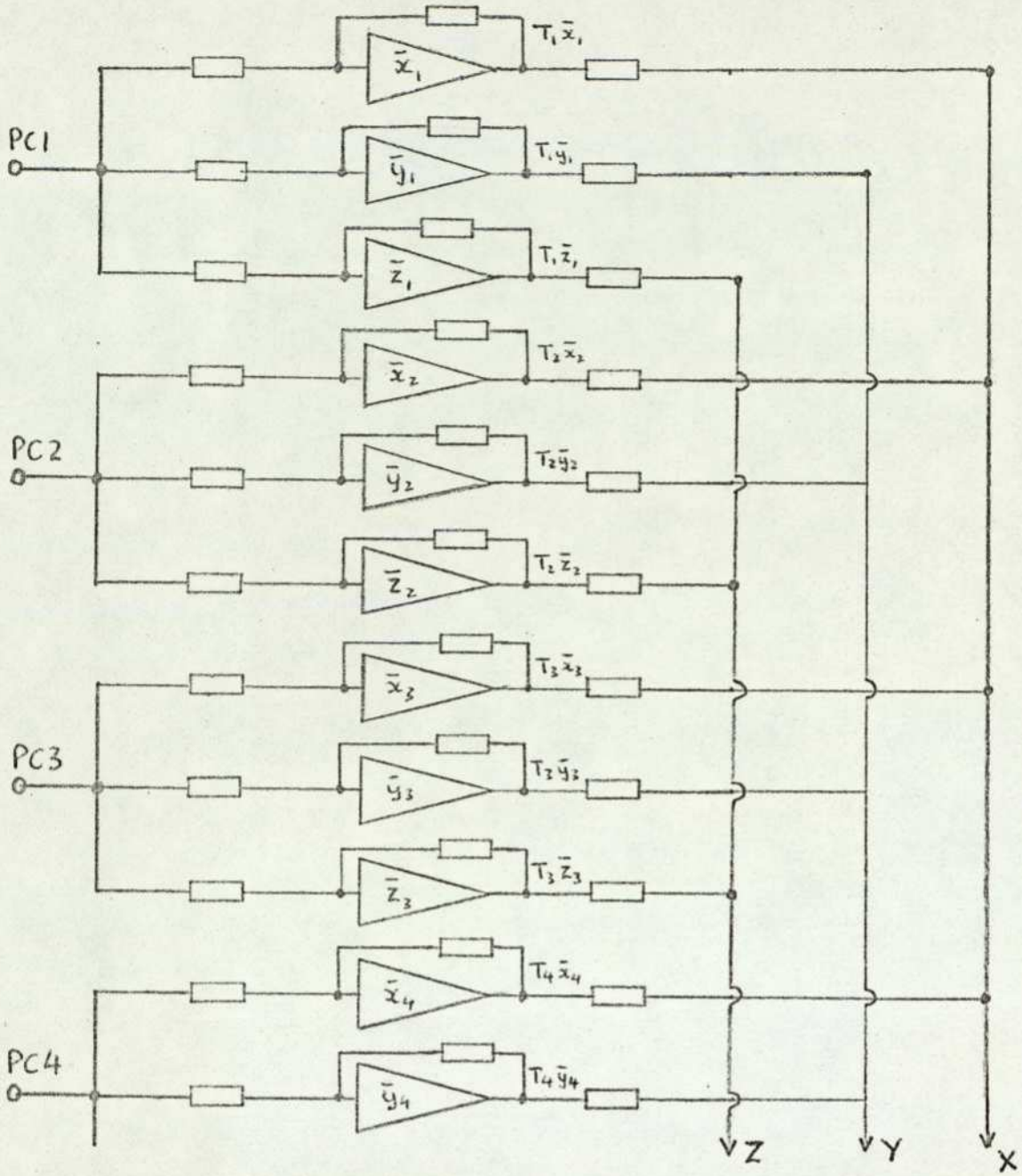


Fig. 7.12 Multiplication by the distribution coefficients.

The gain of each group of amplifiers was arranged to correspond to the distribution coefficients for the equal-energy stimulus. Thus, amplifier x multiplied the output of the first amplifier corresponding to the 400 to 410 nm spectral region, by the distribution coefficient x for the 405 nm wavelength. Amplifier y also multiplied the output of the first amplifier corresponding to the 400 to 410 nm band, by the distribution coefficient y for the 405 nm wavelength. Similarly for the z amplifier.

The choice of combinations of resistors to obtain all the ninety different gains was achieved by the use of a computer. Table 7.3 shows the values of resistors chosen by the computer (both the input resistances and the feedback) as well as the errors involved in such a choice.

The outputs of each x amplifier is summed together to obtain the X tristimulus value. Similarly, for the y and z amplifiers to obtain the Y and Z values.

The X,Y,Z values are then selectively displayed on the digital voltmeter. A selector switch selects the desired parameter to be displayed, ie percentage transmission at a selected wavelength or the tristimulus values.

The same selector switch also enables the display of the x,y chromaticity co-ordinates on the digital meters. To achieve this a divider is used. The circuit for this operation is as shown in Fig. 7.13.

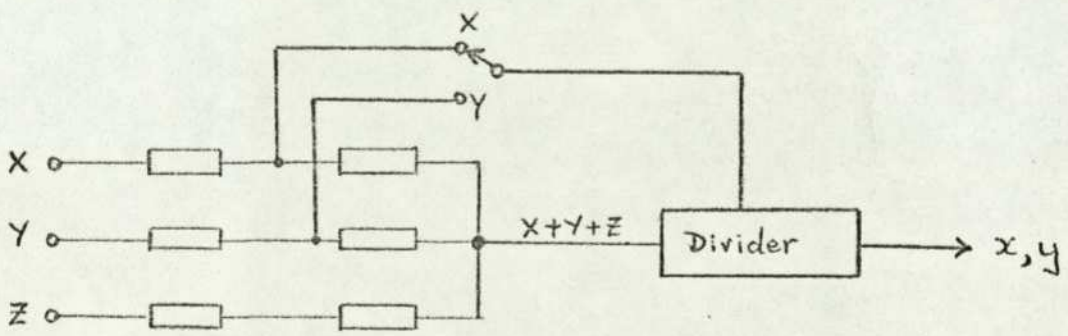


Fig. 7.13 Divider circuit to obtain x,y.

Wavelength Bands	X			Y			Z		
	R_{in} (Kohms)	R_f (Kohms)	% Error	R_{in} (Kohms)	R_f (Kohms)	% Error	R_{in} (Kohms)	R_f (Kohms)	% Error
405	68	1	+ 3.1	-	-	-	100	6.8	0
415	22	1	+ 3.6	-	-	-	27	5.6	0
425	15	1 + 1	- 0.4	-	-	-	6.8 + .47	4.7	0
435	6.8 + 1	2.2	- 0.7	100	1.2	0	4.7 + .22	6.8	- 0.2
445	6.8 + 1	2.7	- 0.6	180 + 22	4.7	0	2.7	4.7	- 0.3
455	6.8	2.2	- 0.6	100	3.9	+ 1.8	2.2	3.9	+ 0.1
465	3.9 + 3.3	1.8	- 0.2	47 + 18	3.9	0	680 + 39	1.2	0
475	10 + 6.8	3.3	+ 0.5	39 + 3.9	3.9	0	3.3 + .33	4.7	+ 0.4
485	56 + 15	6.8	0	27 + 1.2	3.9	0	2.7	2.2	+ 0.2
495	39	1.2	+ 1.5	27	5.6	- 0.2	10 + 1	4.7 + .39	- 0.4
505	100 + 100	1	0	12	3.9	0	10	2.7	- 0.2
515	100 + 100	1.8	0	12 + 1.5	6.8	0	22 + 2.7	3.9	0
525	18 + 1	1.2	0	4.7 + .82	3.9	- 0.4	100	6.8 + 1	0
535	10 + 10	3.3	0	4.7 + 1.8	5.6	0	330 + 27	15	0
545	27 + 1.2	8.2	+ 0.3	2.2 + .1	2.2	0	330 + 12	6.8	0

Table 7.3 Values of input and feedback resistors selected by the computer to provide the multiplication necessary for conversion to X, Y, Z.

Wavelength Bands	X			Y			Z		
	R_{in} (Kohms)	R_f (Kohms)	% Error	R_{in} (Kohms)	R_f (Kohms)	% Error	R_{in} (Kohms)	R_f (Kohms)	% Error
555	12	4.7 + .47	- 0.5	10	10	- 0.5	100 + 100	1.8	0
565	1 + .68	1	0	10	10	- 0.5	-	-	-
575	3.3 + .22	2.7	- 0.6	2.2 + .12	2.2	- 0.4	-	-	-
585	2.2 + .22	2.2	- 0.5	3.3 + .47	3.3	+ 0.6	-	-	-
595	8.2 + 1.5	10	+ 0.4	3.3 + .27	2.7	0	-	-	-
605	5.6 + .82	6.8	0	2.7 + 1.5	1.8	+ 0.2	-	-	-
615	10	10	- 0.3	22 + 1.8	12	+ 0.2	-	-	-
625	3.3 + 2.2	4.7	0	47	18	+ 0.5	-	-	-
635	2.7	2.7 + 1.5	+ 0.1	68	18	0	-	-	-
645	2.2	4.7 + 0.22	0	68	12	+ 0.6	-	-	-
655	1.8 + .12	6.8	0	10 + 6.8	1.8	0	-	-	-
665	3.3	10 + 10	0	33 + 3.3	2.2	0	-	-	-
675	6.8	68 + 10	+ 0.2	100	3.3	+ 2.6	-	-	-
685	4.7	100	0	100	1.8	+ 3.0	-	-	-
695	3.9	120 + 47	+ 0.9	100	.82	- 1.9	-	-	-

Table 7.3 Values of input and feedback resistors selected by the computer to provide the multiplication necessary for conversion to X, Y, Z.

7.6 Optical Changes to the Spectrocolorimeter

The optical layout of Fig. 7.3, even though useful for earlier experimental work, was not very convenient to incorporate into the final version of the spectrocolorimeter. Beam folding by the use of mirrors was necessary in order to keep its size down to reasonable proportions. It was, however, considered that replacing the lenses by spherical mirrors would provide the most suitable answer. The final optical layout is as shown in Fig. 7.14.

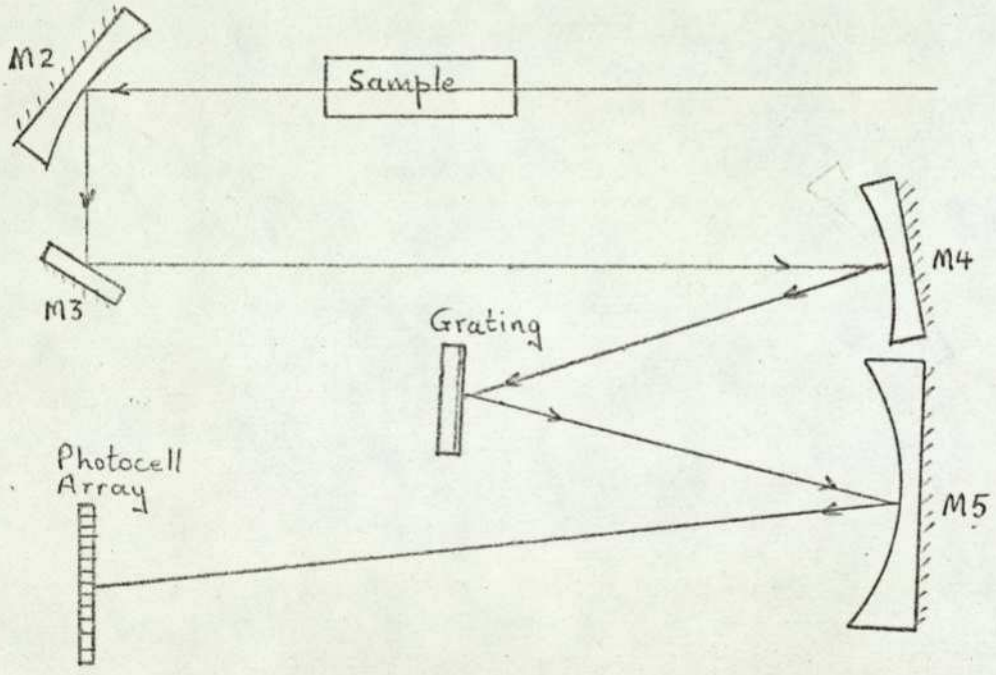


Fig. 7.14 Final optical layout.

In the final version four spherical concave and one plane mirrors have been used. They are all front silvered. M2 is 22.4 mm diameter and of 30 mm radius of curvature. M3 is a plane 20 mm square mirror. M4 is 38 mm diameter and of 300 mm radius of curvature. M5 is 75 mm diameter and of 300 mm radius of curvature.

Three lenses (achromatic doublets) are used to produce the parallel beam transmitted through the sample, as shown in Fig. 7.15. L1 is a plano-convex lens 25 mm diameter, 44.5 mm focal length. L2 and L3 are bi-convex lenses 25 mm diameter, 25 mm focal length.

The beam diameter is actually 10 mm. Thus sperical aberration is highly reduced. The beam divergence obtained with this system is less than 3°.

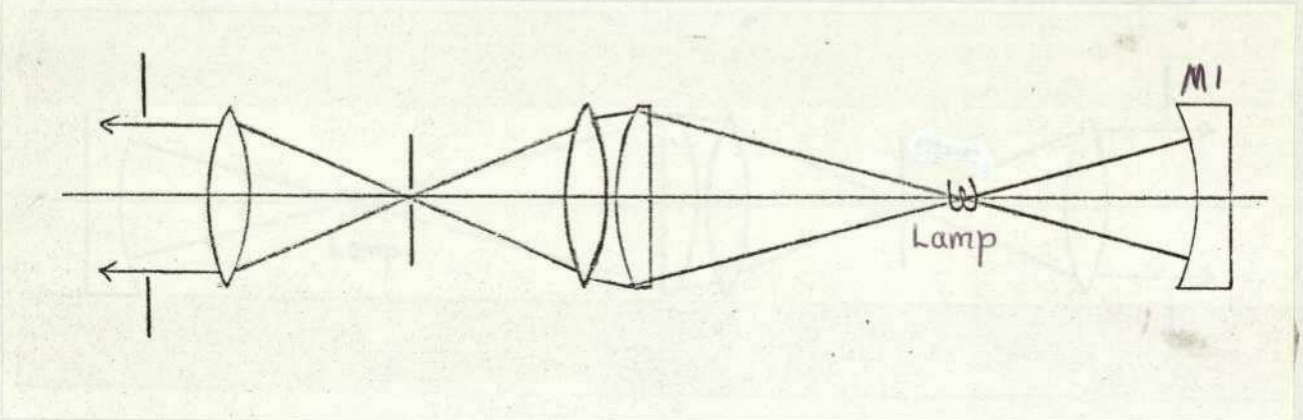


Fig. 7.15 Optical lens system for parallel beam of light.

7.7 Analysis of Some Problems that Arose During the Development Stages

(a) The optical arrangement shown in Fig. 7.16 is very common with transmission spectrophotometers of a number of makes.

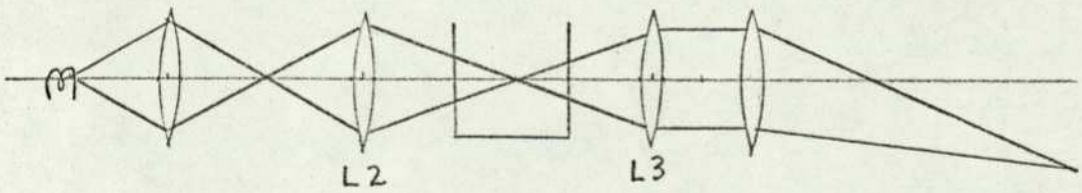


Fig. 7.16 Normal optical layout for transmission spectrophotometers.

Fig. 7.17 shows in more detail what happens when a convergent beam strikes two sample cells of different lengths.

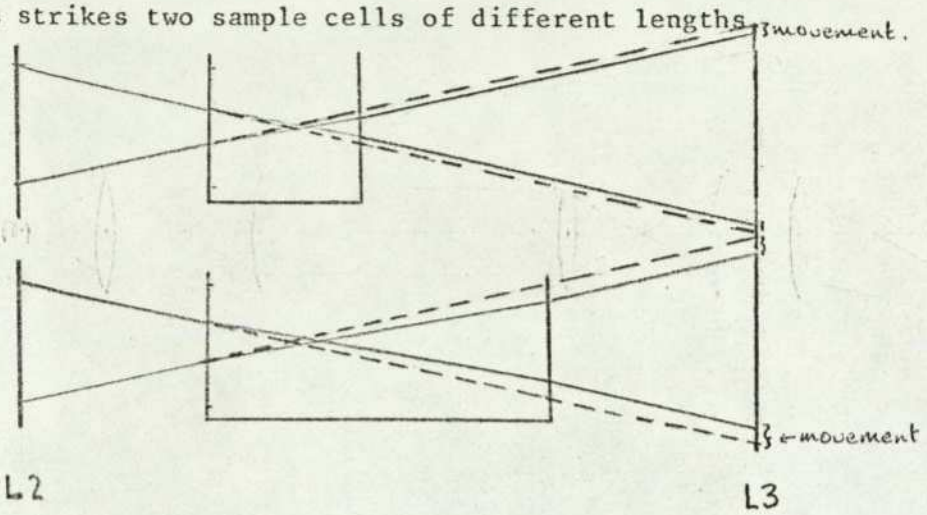


Fig. 7.17 Movement of image depending on length of cell.

It can be seen from Fig. 7.17 that the source image as formed by L2 is shifted by an amount depending on the optical path length of the sample cell and the refractive index of the sample. It follows that a non-parallel beam falls on the diffraction grating with the consequent spectrum problems.

To overcome this effect it was decided to use a parallel beam between L2 and L3 and introduce another lens as shown in Fig. 7.18.

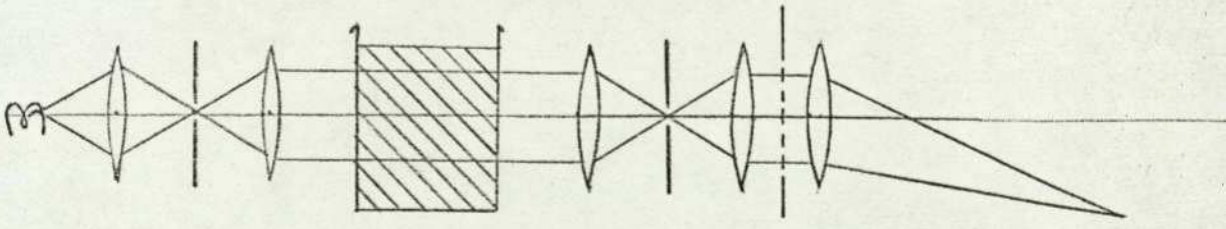


Fig. 7.18 Modified optical layout.

(b) The front surface of the photocells has a very high reflection coefficient. It was found that if the light was falling perpendicularly on the photocells, the reflected light went back along the same optical path and part of it was reflected back again. These multiple reflections could cause large errors, because they were interacting, ie reflections from one photocell could eventually reach another photocell at another part of the spectrum.

To overcome this problem, all the photocells were placed at a small angle of about 5° to the incident beam. Thus any photocell reflections left the optical system and eventually absorbed by the surroundings, which were covered by black velvet cloth.

7.8 Photographs of the Spectrocolorimeter

Various views of the spectrocolorimeter are depicted in Figs. 7.19 to 7.24.

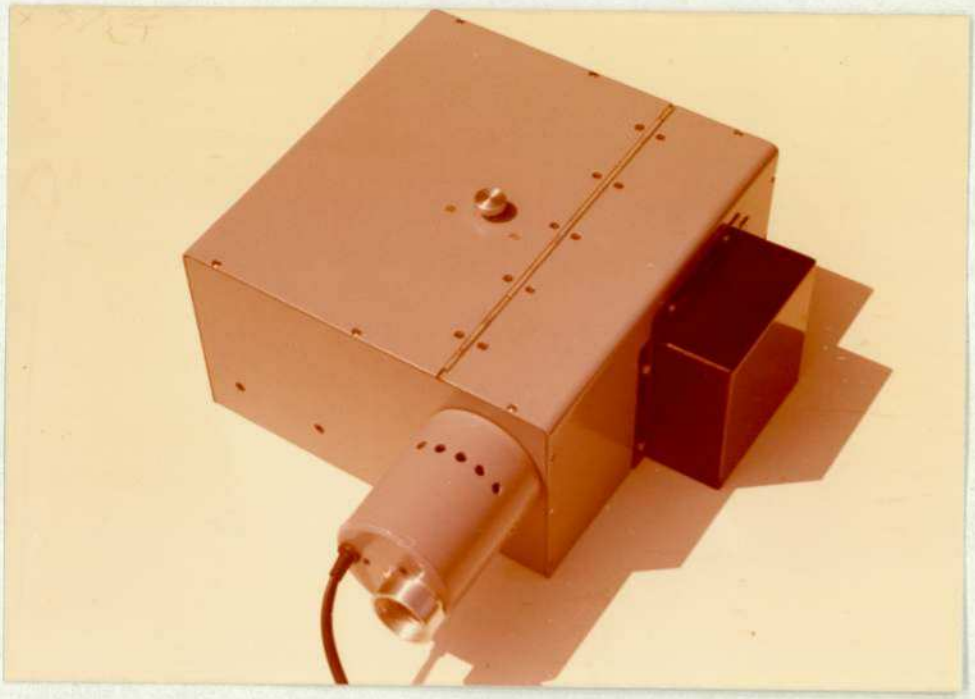


Fig. 7.19. External view of the optical unit.

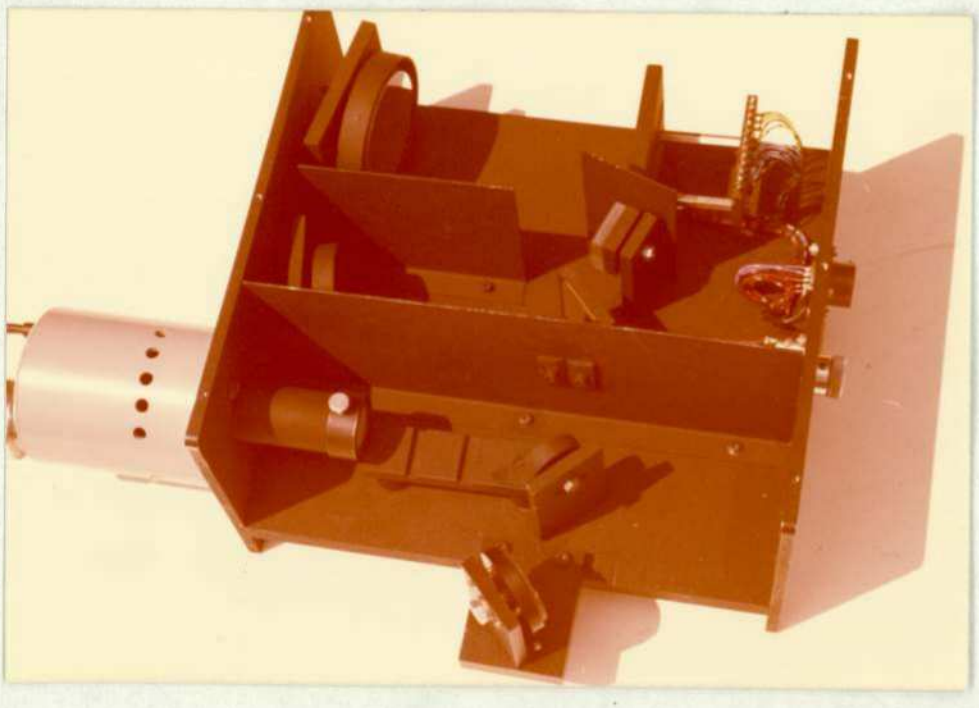


Fig. 7.20. Internal view of the optical unit.

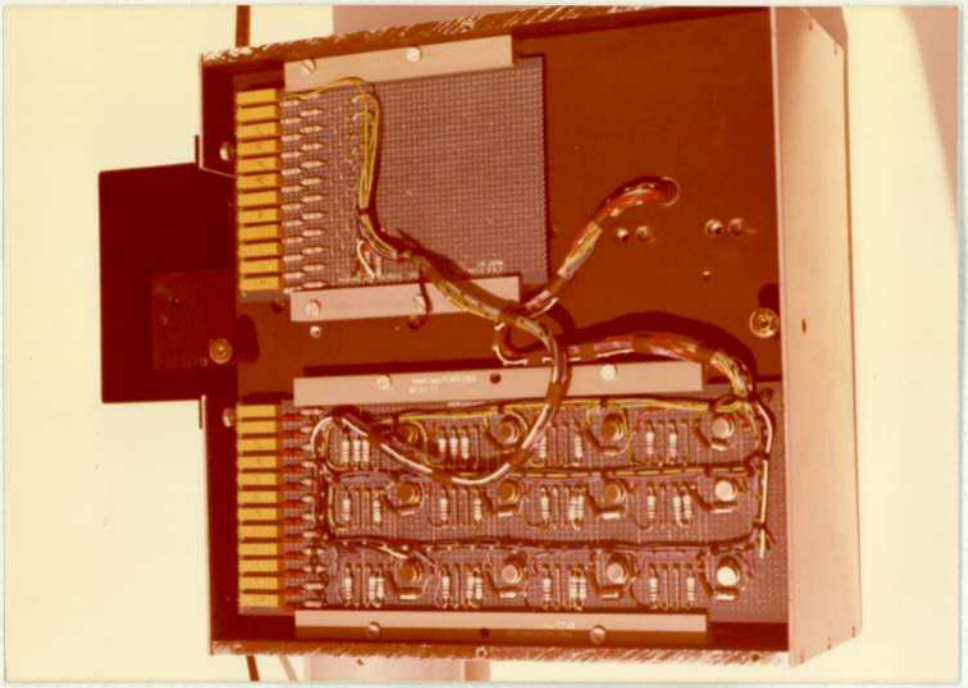


Fig. 7.21. Amplifier boards on the optical unit

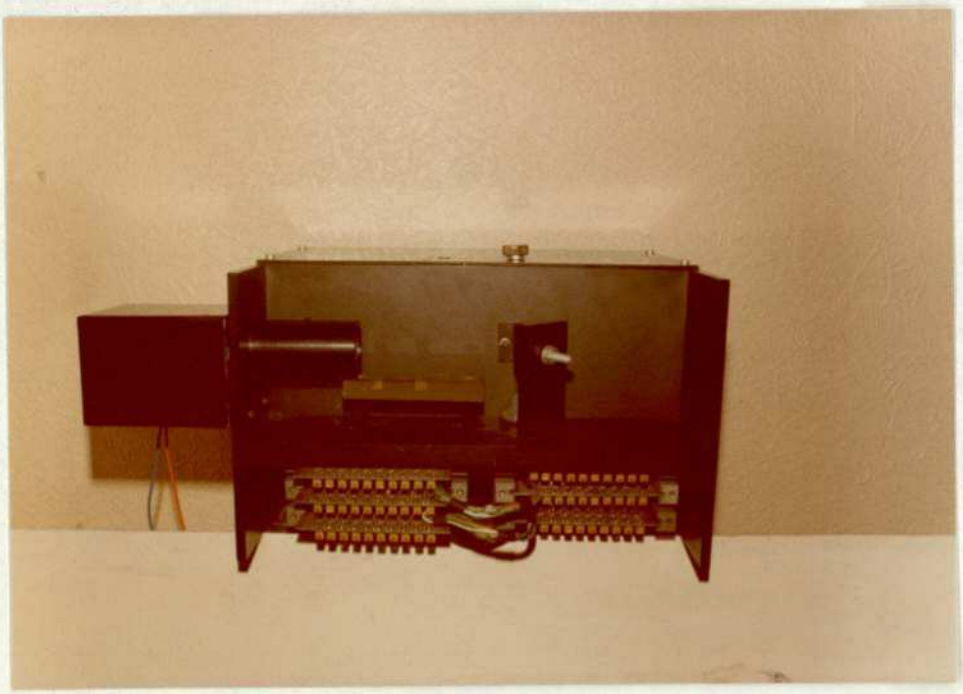


Fig. 7.22. Front view of the optical unit

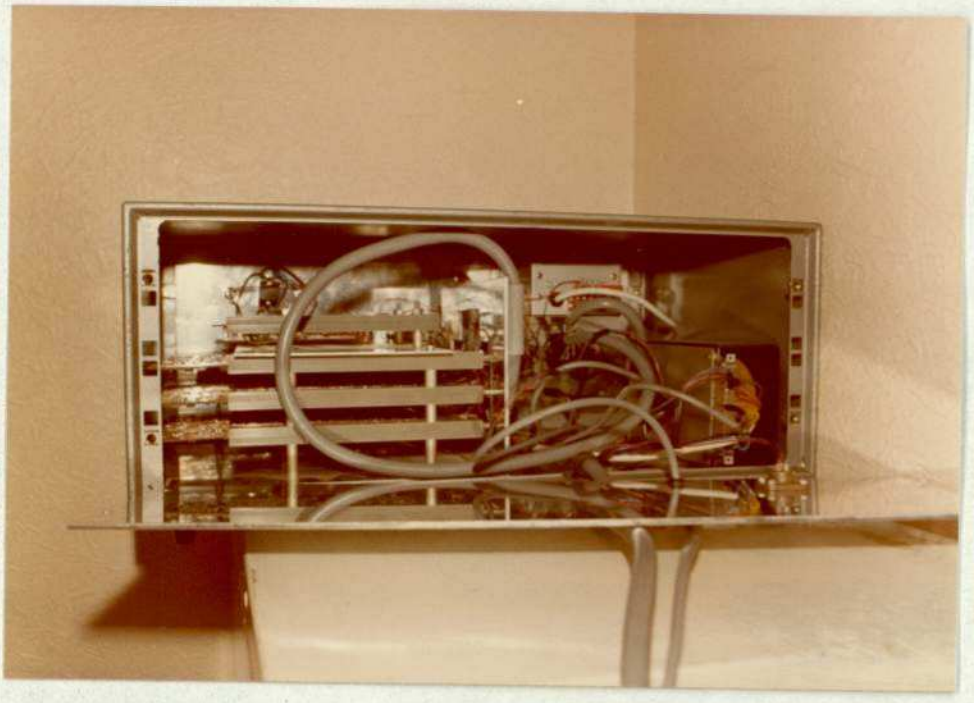


Fig. 7.23. Internal view of the electronics unit.

Fig. 7.24. External view of the electronics unit.

CHAPTER 8

PERFORMANCE OF THE SPECTROCOLORIMETER

8.1 Introduction

All the readings obtained regarding the precision and accuracy of the spectrocolorimeter were through computer printouts. The output from 12 channels was fed via an interface unit and an analogue to digital converter into a PDP-11 mini computer. Thus the signals could be read and printed easily and quickly.

These readings were printed to the nearest 10 microvolts. However, in the Tables presented below the last digit was rounded off in most cases. Thus, for a full scale of 1 volt the last digit corresponds to 0.01% of full scale.

In other cases where comparisons with other instruments or filters with known transmissions were made, the rounding off was sometimes to 0.1% of full scale. This was so because the figures obtainable from the other instruments available from manufacturers' literature was not better than 0.1%.

8.2 Precision and Accuracy as a Spectrophotometer

The spectrocolorimeter was checked against a standard laboratory spectrophotometer and the results for a number of filters were also compared with the figures provided by the filter manufacturers.

Table 8.1. Linearity as shown by four channels. Values given correspond to percentage transmission values.

	0.2 N.D.			0.4 N.D.			0.6 N.D.			0.8 N.D.		
λ	Kodak	Spectroc	Cecil	Kodak	Spectroc	Cecil	Kodak	Spectroc	Cecil	Kodak	Spectroc	Cecil
465	50.3	51.2	51.8	31.9	32.1	32.6	19.5	19.5	19.9	12.78	13.92	13.6
525	56.4	56.9	57.4	35.8	35.9	36.2	22.4	22.8	23.5	14.32	14.64	14.7
585	58.9	59.0	59.6	37.4	37.8	38.1	23.4	23.9	24.2	14.95	15.36	15.6
645	57.3	57.3	58.1	36.4	36.5	36.9	22.7	22.6	22.8	14.55	14.20	14.8

	1.0 N.D.			1.2 N.D.			1.4 N.D.			1.6 N.D.		
λ	Kodak	Spectroc	Cecil	Kodak	Spectroc	Cecil	Kodak	Spectroc	Cecil	Kodak	Spectroc	Cecil
465	7.98	8.06	8.1	5.03	5.12	5.2	3.19	3.25	3.3	1.95	2.02	2.0
525	8.95	9.09	9.2	5.64	5.96	6.0	3.58	3.71	3.6	2.24	2.38	2.4
585	9.36	9.56	9.7	5.89	6.12	6.2	3.74	3.68	3.8	2.34	2.36	2.4
645	9.09	9.22	9.4	5.73	6.04	5.9	3.64	3.62	3.7	2.27	2.27	2.3

Kodak values are the manufacturer's data.

Cecil values are as measured on the Cecil manual spectrophotometer.

Spectroc values are as measured on the spectrocolorimeter.

8.2.1 Linearity of Percentage Transmission Scale

Table 8.1 shows comparisons of results for four channels: 460-470, 520-530, 580-590 and 640-650. These channels are quoted by their mean value in the table. A range of eight neutral density (N.D) filters has been used. These are Kodak Wratten 96 gelatin filters. The Wratten 96 filters are made by suspending colloidal carbon in dyed gelatin. The transmission values quoted by Kodak are only within an accuracy of ±3% of the nominal diffuce opal density value (B.S. 1384:1962) given in Table 8.1.

Thus any correlation obtained cannot be guaranteed to be better than ±3% in terms of absolute values. As a check to these transmission values, further readings were taken on another manual laboratory spectrophotometer, the Cecil SP302. These values are also shown in Table 8.1.

Any conclusions derived from a mathematical evaluation of correlation coefficients or standard deviations can only be very limited and superficial. This is because such values are always in the region of 0.999. Close visual examination of these figures reveals that the precision of the spectrocolorimeter must be at least as good as that of the Cecil or the figures given for the Wratten filters.

The question of the accuracy (as opposed to its precision) of linearity of the instrument can only be solved by the use of standard filters as for example those measured by the National Physical Laboratory. Such filters were not readily available.

The theoretical linearity of each channel of the spectrocolorimeter is dependent on the photocell impedance and the input impedance of the preamplifier (see Inset). As explained there, the combination of the BPX40 photocell and the AD540J FET input amplifier offers a nearly ideal system. An absolutely ideal system should result, in theory, with an FET input amplifier with a much higher input impedance. These amplifiers are, however, very expensive.

Through the use of a programmable calculator (Texas SR52) the standard formulae were used to determine the equations of the best-fit lines relating the various values shown in Table 8.1. Thus the values of a and b were worked out in the equation $y=a + bx$. Values of the correlation coefficients were also worked out. These were always higher than 0.999 and, as explained earlier, any conclusions derived from correlation coefficients can only be very superficial. These figures are therefore not quoted.

The values of a and b do, however, provide the necessary information to evaluate the errors in Range Sensitivity, and the Zero errors. By plotting the points on a graph and drawing the derived line, it is easy to derive the maximum error in linearity.

Fig. 8.1 shows the various errors normally defined in calibrating an instrument against a standard.

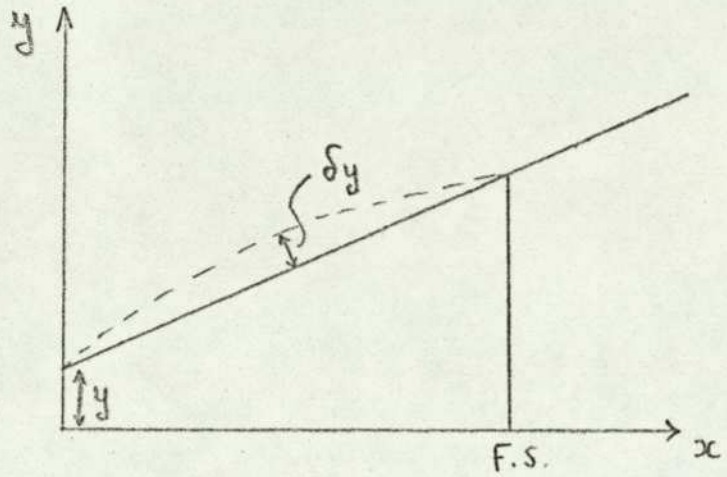


Fig. 8.1 Instrument Errors.

$$\text{Zero Error} = \frac{y}{\text{Full Scale}} \times 100\%$$

$$\text{Range Sensitivity Error} = \frac{\delta(\text{Slope})}{\text{Slope}} \times 100\%$$

$$\text{Non-Linearity Error} = \frac{\delta y}{\text{Full Scale}} \times 100\%$$

The average error values obtained were as follows:

(a) Spectrocolorimeter values referred to the Kodak values:

Zero Error = $\pm$0.1%

Range Sensitivity Error = +1.1%

Non-linearity Error = +1.2%

(b) Spectrocolorimeter values referred to the Cecil values:

Zero Error = $\pm$0.05%

Range Sensitivity Error = +0.8%

Non-linearity Error = +0.9%

8.2.2 Drift with Time

Table 8.2 shows the variation in the full scale of the output of 12 channels over a period of 6 hours. It must be noted that during this period the room temperature did not change by more than 3°C (19°C to 22°C). Half an hour was allowed for the instrument to stabilise with the lamp switched on, before taking the first readings. During the six hours of the test, the lamp voltage did not change by more than ±1mV.

Table 8.3 shows the variation in the zero output of the 12 channels over a different period of 6 hours. During this test the room temperature varied by about 8°C (20°C to 28°C). The lamp was switched on throughout the test. The light was stopped from reaching the photocells by a metal disc in front of the slit.

The variations obtained for both the above tests are too small for graphical representation.

In theory, there is no reason why a drift with time should occur. However, the author knows of no instrument that does not drift with time. Nevertheless, in this test the spectrocolorimeter probably surpassed any other known similar type of instrument. The maximum drift obtained in full scale was 0.23%, whilst the maximum zero drift was 0.26% of full scale.

Table 8.2 Full scale drift with time at room temperature

Channel Number	Original Reading	After 5 min.	After 30 min.	After 2 hours	After 6 hours
1	1.0000	0.9991	0.9981	1.0011	1.0009
2	1.0009	1.0000	0.9983	0.9988	1.0016
3	1.0000	1.0001	0.9991	1.0012	1.0014
4	1.0010	0.9991	1.0009	1.0003	1.9992
5	0.9992	0.9992	0.9991	1.0004	1.0013
6	1.0009	1.0009	1.0009	1.0012	1.0008
7	1.0000	0.9996	1.0009	1.0005	1.0000
8	1.0000	1.0002	1.0011	1.0000	1.0009
9	0.9990	1.0001	1.0011	1.0013	1.0011
10	1.0000	0.9993	1.0008	1.0013	1.0010
11	1.0000	1.0004	1.0002	1.0018	1.0016
12	1.0000	0.9995	1.0002	1.0005	1.0002

Table 8.3 Zero drift with time at room temperature

Channel Number	Original Reading	After 5 min.	After 30 min.	After 2 hours	After 6 hours
1	0.0023	0.0034	0.0050	0.0038	0.0041
2	-0.0035	-0.0032	-0.0031	-0.0030	-0.0020
3	-0.0022	-0.0023	-0.0018	-0.0017	-0.0015
4	-0.0036	-0.0028	-0.0034	-0.0032	-0.0030
5	0.0048	0.0043	0.0035	0.0038	0.0042
6	-0.0012	-0.0008	-0.0011	-0.0006	-0.0005
7	0.0000	0.0004	0.0010	0.0008	0.0011
8	0.0059	0.0047	0.0046	0.0049	0.0052
9	-0.0012	-0.0023	-0.0023	-0.0021	-0.0019
10	0.0022	0.0022	0.0018	0.0019	0.0023
11	0.0071	0.0047	0.0047	0.0051	0.0052
12	0.0014	0.0023	0.0012	0.0016	0.0017

Table 8.4 Full scale drift with ambient temperature

Channel Number	10°C	20°C	30°C	40°C
1	1.0012	1.0000	0.9938	0.9880
2	1.0006	1.0000	0.9942	0.9891
3	1.0008	0.9998	0.9956	0.9905
4	1.0005	0.9996	0.9954	0.9912
5	1.0022	1.0011	0.9961	0.9921
6	1.0014	1.0004	0.9947	0.9913
7	1.0007	1.0000	0.9936	0.9905
8	1.0011	1.0006	0.9942	0.9907
9	1.0021	1.0017	0.9953	0.9914
10	1.0014	1.0009	0.9941	0.9896
11	0.9991	0.9986	0.9925	0.9877
12	0.9996	0.9994	0.9915	0.9894

Table 8.5 Zero drift with ambient temperature

Channel Number	10°C	20°C	30°C	40°C
1	0.0034	0.0028	0.0006	-0.0019
2	-0.0021	-0.0034	-0.0058	-0.0076
3	-0.0013	-0.0018	-0.0047	-0.0081
4	-0.0024	-0.0032	-0.0063	-0.0096
5	0.0046	0.0039	0.0011	-0.0037
6	0.0006	-0.0002	-0.0034	-0.0055
7	0.0008	0.0003	-0.0029	-0.0054
8	0.0076	0.0064	0.0028	-0.0013
9	-0.0002	-0.0008	-0.0042	-0.0085
10	0.0021	0.0013	-0.0030	-0.0082
11	0.0083	0.0076	0.0036	0.0001
12	0.0026	0.0019	-0.0033	-0.0067

8.2.3 Drift with Temperature

To obtain the steady temperature variations, the instrument was placed in a climatic cabinet whose temperature could be controlled to within $\pm 1^{\circ}C$.

Half an hour was normally allowed after the set temperature was reached by the climatic cabinet, before any readings were taken. This was to ensure that all parts of the instrument were at the same temperature. Tables 8.4 and 8.5 show the results obtained. The variations obtained are too small to represent graphically.

The variation of around 1% obtained in the output, for a change of $30^{\circ}C$, is about three times less than that to be expected by considering the photocells on their own (see Inset). It appears though as if some compensating mechanism exists in the system.

The most obvious sources of such a compensation are: the amplifiers, the feedback resistors, the lamp (see Section 8.3).

8.2.4 Accuracy of Wavelength Calibration

To check that each one of the photocells was reading the correct wavelength of light, three narrow band filters and five wide band filters were used. The results obtained on the Spectrocolorimeter

and the Cecil Spectrophotometer, as well as those quoted by the manufacturers, are tabulated in Tables 8.6 to 8.9. For the wide band filters (Kodak Wratten), only ten percentage transmission measurements are shown at intervals of 30 nm. For the narrow band filters only relevant wavelengths are indicated, as the transmission in the remaining part of the spectrum is zero. The narrow band filters used were manufactured by Spectrum Systems Inc. (USA).

For the wide band filters the average error values obtained were as follows:

(a) Spectrocolorimeter values referred to the Kodak values:

Zero Error = -0.05%

Range Sensitivity Error = +1.8%

Non-linearity Error = +1.4%

(b) Spectrocolorimeter values referred to the Cecil values:

Zero Error = +0.13%

Range Sensitivity Error = -0.9%

Non-linearity = +1.1%

For the narrow band filters, the average error values (Spectrocolorimeter values referred to the Spectrum Systems values) were as follows:

(a) Filter SS590:

Zero Error = +3.4%

Range Sensitivity Error = +2.6%

Non-linearity Error = $\pm$1.3%

	45 (Blue-green)				38 (Light blue)				38A (Blue)				59A (Light green)				96(N.D.1.0)		
λ^*	Kodak	Cecil	Col.		Kodak	Cecil	Col.		Kodak	Cecil	Col.		Kodak	Cecil	Col.		Kodak	Cecil	Col.
405	0	0	0		63.5	63.9	63.8		38.7	39.0	38.8		0.14	0.1	0.12		4.60.	4.6	4.58
435	0	0.1	0.18		75.8	76.2	76.1		57.8	58.1	58.0		0.67	0.6	0.68		6.54	6.6	6.86
465	32.8	31.6	31.2		74	74.5	74.5		52.6	53.2.	53.1		11.2	11.0	11.2		7.98	7.9	8.06
495	33.1	31.8	31.5		68.1	68.7	68.5		41.7	42.3	42.3		59.1	58.9	59.5		8.66	8.5	8.73
525	5.8	4.9	5.0		59	59.4	59.3		29.7	30.3	30.2		74.5	73.8	75.5		8.95	8.8	9.09
555	0	0	0.12		45.5	46.2	46.1		15.9	16.2	16.4		59	58.7	59.4		9.25	9.3	9.48
585	0	0	0.06		28.1	28.5	28.4		5.1	5.4	5.5		28.8	28.6	29.2		9.36	9.4	9.56
615	0	0	0		14.9	15.4	15.4		1.3	1.5	1.6		10.3	10.2	10.6		9.65	9.7	9.72
645	0	0	0		7.2	7.6	7.7		0.1	0.2	0.24		1.18	1.3	1.62		9.09	9.2	9.22
675	0	0	0		5.7	5.9	5.9		0	0.1	0.16		7.5	8,0	8.73		9.10	9.0	9.24

Table 8.6 Wavelength calibration check, using wide band transmission Filters (Kodak Wratten). Values given correspond to percentage transmission as on Table 8.1.

Table 8.7 Percentage Transmission and C.I.E. Values of narrow band red filter SS590

	Wavelength	SS Value	Spectrocol Value	% Age Difference
	525	0	0.09	-
	535	0	0.10	-
	545	0.2	0.28	2.8
	550	0.5	-	-
	555	1.5	1.8	1.6
	560	3.7	-	-
	565	9.2	9.3	1.07
	570	22.5	-	-
	575	45.0	46.2	2.6
	580	62.5	-	-
	585	70.0	71.4	1.9
	590	71.7	-	-
	595	70.7	72.6	2.7
	600	67.7	-	-
	605	63.5	63.0	0.8
	610	58.0	-	-
	615	49.5	50.5	1.9
	620	31.2	-	-
	625	15.2	15.5	1.9
	630	5.7	-	-
	635	2.7	2.9	6.9
	640	1.0	-	-
	645	0.5	0.82	3.9
	650	0.2	-	-
	655	0	0.23	-
	665	0	0.16	-
	675	0	0.09	-
	685	0	0.02	-
C.I.E. 10 nm interval	X	31.13	31.64	1.6
	Y	21.47	22.41	4.2
	Z	0.06	0.06	-
	x	0.5859	0.5846	-
	y	0.4135	0.4142	-
	z	0.0006	0.0013	-
C.I.E. 5 nm interval	X	31.29		
	Y	22.08		
	Z	0.07		
	x	0.5856		
	y	0.4132		
	z	0.0012		

SS values correspond to the values provided by the manufacturer (Spectrum Systems).

Table 8.8. Percentage Transmission and C.I.E. Values of narrow band green filter SS540.

	Wavelength	SS Value	Spectrocol Value	Percentage Difference
	475	0	0.01	-
	485	0	0.08	-
	495	0.2	0.27	2.6
	500	0.7	-	-
	505	1.2	1.20	0
	510	3.0	-	-
	515	7.7	8.3	7.2
	520	19.7	-	-
	525	49	48.6	0.8
	530	67.2	-	-
	535	71.7	73.4	2.3
	540	70.7	-	-
	545	65	66	1.5
	550	57	-	-
	555	48	47.4	1.3
	560	35	-	-
	565	21.5	21.8	1.4
	570	18.7	-	-
	575	3.7	4.1	9.7
	580	1.7	-	-
	585	0.7	0.85	17.6
	590	0.5	-	-
	595	0.2	0.24	16.6
	605	0	0.09	-
C.I.E. 10nm interval	X	9.48	9.26	2.3
	Y	25.23	25.01	0.8
	Z	0.71	0.74	4.1
	x	0.2675	0.2633	
	y	0.7123	0.7155	
	x	0.0202	0.0212	
C.I.E. 5nm interval	X	9.16		
	Y	24.96		
	Z	0.72		
	x	0.2629		
	y	0.7163		
	z	0.0208		

SS values correspond to the values provided by the manufacturer (Spectrum Systems).

Table 8.9 Percentage Transmission and C.I.E. Values of narrow band blue filter SS460.

	Wavelength	SS Value	Spectrocol Value	Percentage Difference
	415	0.0	0.07	-
	425	0.2	0.25	20
	430	0.7	-	-
	435	2.7	2.9	6.9
	440	7.5	-	-
	445	21.2	21.3	0.5
	450	54	-	-
	455	69		4.9
	460	70.5	-	-
	465	63.5	61.5	3.2
	460	54	-	-
	475	43.2	43.8	1.3
	480	24	-	-
	485	8.2	8.9	7.9
	490	3	-	-
	495	1	1.4	28
	500	0.2	-	-
	505	0	0.18	-
	515	0	0.04	-
	525	0	0.01	-
C.I.E. 10nm interval	X	5.44	5.38	1.12
	Y	1.54	1.55	0.64
	Z	31.79	31.54	0.79
	x	0.1403	0.1400	
	y	0.0397	0.0405	
	z	0.8200	0.8196	
C.I.E. 5 nm interval	X	5.35		
	Y	1.53		
	Z	31.34		
	x	0.1400		
	y	0.0400		
	z	0.8200		

(b) Filter SS540:

Zero Error = +3.1%

Range Sensitivity Error = +3.2%

Non-linearity Error = ±1.5%

(c) Filter SS460:

Zero Error = +2.2%

Range Sensitivity Error = +1.8%

Non-linearity Error = ±1.4%

It can be seen that the errors involved in the measurement of transmission values of narrow band filters are higher than those obtained in the measurements of the Kodak Wratten filters. This is obviously due to the low resolution of the spectrocolorimeter which is dependent upon the finite width of the measuring photocells. Thus very abrupt changes in the spectrum cannot be accurately defined by the spectrocolorimeter. The spectrum systems spectrophotometer used had a slit width of only two nanometers, and thus had a much higher resolution.

In spite of this limitation, the spectrocolorimeter is comparable to most common spectrophotometers as far as resolution and accuracy of wavelength calibration are concerned. It certainly is at least as good as the Cecil spectrophotometer used in these tests.

8.3 Accuracy as a Colorimeter

The C.I.E. values for a number of filters, as measured by the spectrocolorimeter, were compared with the values obtained by straight evaluation of their transmission values, as provided by the manufacturers' data using a computer. Evaluations and comparisons were also made on filters at 5 nm intervals.

8.3.1 The Use of 30 Points for the Evaluation of the C.I.E. Tristimulus Values.

At the bottom of Tables 8.7, 8.8 and 8.9 the C.I.E. Tristimulus Values evaluated at 5 and 10 nm intervals, as well as those measured on the spectrocolorimeter are shown. These filters have been deliberately chosen as they should show high error values in colour evaluations due to the relatively high errors obtained in their spectral transmissions, as discussed in Section 8.2.4.

Considering, however, the Spectrum Systems values only (which are assumed to be much more precise), it can be seen that the two sets of X,Y,Z values in each of the three Tables agree very closely between them.

Thus in Table 8.7 for the red filter the percentage differences in the X,Y,Z values, as obtained by evaluating firstly at 10 nm intervals and then at 5 nm intervals, are:

% difference in X = 0.51%

% difference in Y = 2.8%

% difference in Z = 14.2%

The last value above should not be given any special significance. It is superficially high due to the small numbers involved. Furthermore, the X,Y,Z values, as they stand, do not provide sufficient detail as to how important these differences really are. The Y value, which defines the lightness of the sample, is the most important of the three values. A difference of 2.8% in the lightness value is, however, extremely hard to perceive, even under the best possible conditions of observation.

The differences in the chromaticities, as shown in Table 8.7 for the red filter, are:

Difference in x = 0.0003

Difference in y = 0.0003

Difference in z = 0.0006

Again, the difference in Z should not be given any special significance due to the absolute values involved. The above differences in x and y are extremely close to the threshold perception values. Wright (Wright, 1959) quotes tolerance values of ± 0.0005 in x and y for near neutral colours for master panels in the automobile industry.

The percentage differences in the Y values obtained in Tables 8.8 and 8.9 for the green and blue filters are 1.1% and 0.6% respectively. In Table 8.8 the chromaticity differences were 0.0046 and 0.0040 in x and y respectively. In Table 8.9 the differences are 0.0003 for both x and y.

The chromaticity differences obtained in Table 8.8 are much higher than those in the other two tables. These values do, however, refer to a green colour. MacAdam (MacAdam, 1942) and Wright (Wright, 1941) have shown that subjectively equal colour steps in the "green" region of the chromaticity diagram could correspond to ten or even twenty times as high chromaticity difference values as those in the "red" or "blue" regions of the chromaticity diagram. The chromaticity co-ordinates for the red and the green filters are shown in Fig. 8.3 where direct comparison with the MacAdam ellipses is also possible.

The difference in values for the SS460 Blue filter are too close to show on a chromaticity diagram of these dimensions.

It can thus be concluded that all the above differences between the two sets are very close to the absolute perception threshold, and therefore the 10 nm interval is perfectly acceptable. It can be assumed that the above differences are even smaller for wide band filters.

Therefore the 10 nm interval should be acceptable for any sample.

Strocka (Strocka, 1973) concluded that the 20 nm interval normally used for computer colorant formulations is quite acceptable for the vast majority of cases. Strocka does, however, stress that "samples with high chroma and curves with steep slopes and sharp curvatures will be most critical". De Kerf (De Kerf, 1958) concluded that for three interference filters and a didymium filter, the 10 nm interval was not sufficient. The author (unpublished data) showed that for a large number of industrial applications a 25 nm interval is also acceptable.

8.3.2 The Use of "Bands" Rather than Narrow Wavelength Intervals

It can be assumed that for samples with slowly changing spectral curves, there can be very little difference as to whether one integrates through the use of intervals or bands. Any differences would necessarily be associated with just samples with steep slopes and sharp curvatures.

Fig. 8.2 shows the transmission curve of a didymium filter over the region 500 to 540 nm, at 1 nm intervals.

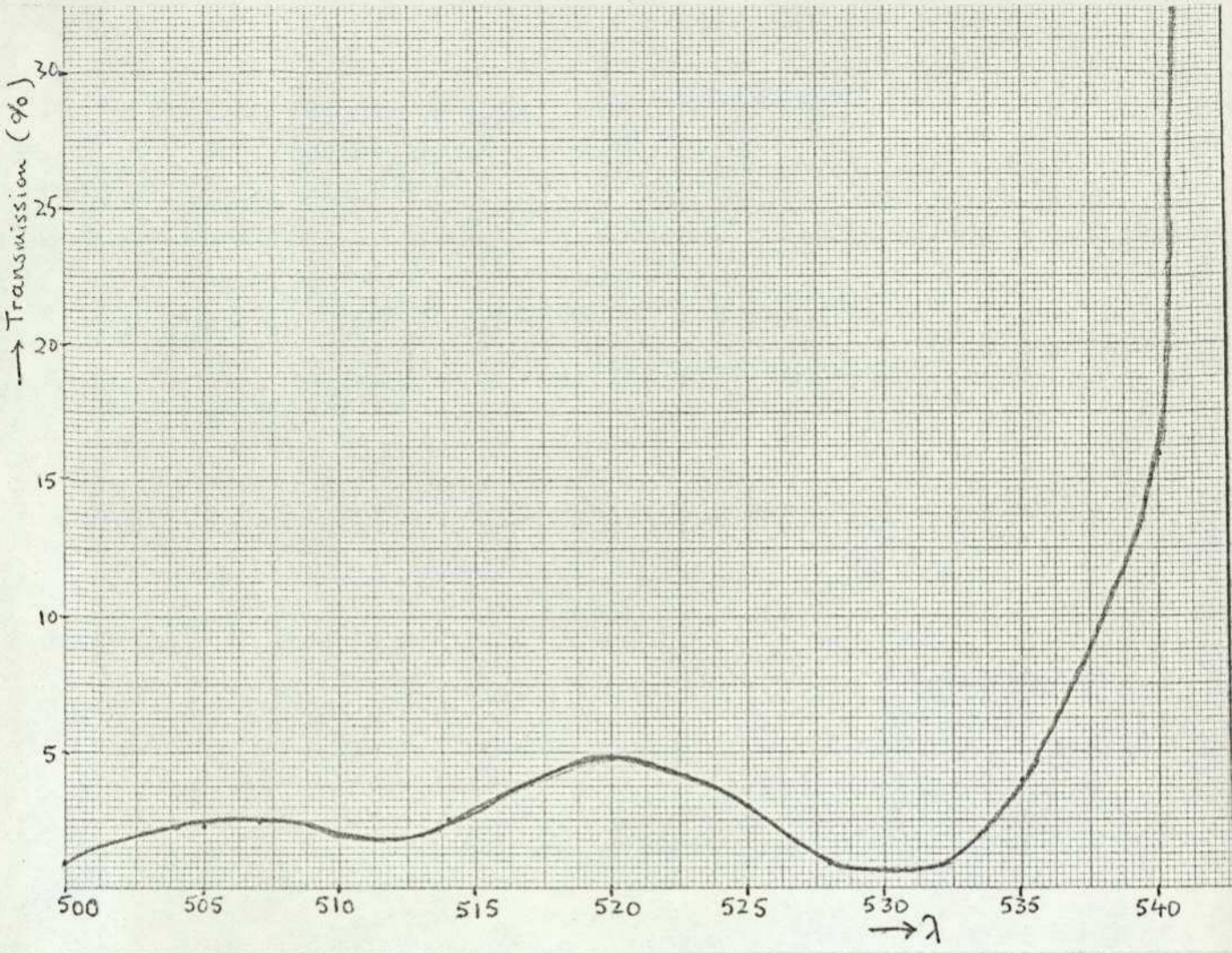


Fig. 8.2 Transmission curve of a didymium filter.

It can be seen that the transmission at 530 nm is 0.7%. If, however, one was considering a 10 nm band, ie between 525 and 535 nm, then the integrated transmission over this band would be 1.7%. Thus a large difference would be observed. Similarly, the transmission at 535 nm is 4%, whilst the integrated transmission between 530 and 540 nm is 5.85%.

It should, however, be borne in mind that the distribution coefficients refer to spot measurements rather than to band intervals. It is not, therefore, necessarily true that by

calculating the tristimulus values with the use of bands rather than with the use of spot measurements, one would obtain increased accuracy.

One could define the ideal tristimulus values as those obtained by multiplying spot measurements at 1 nm intervals with the distribution coefficients obtained at 1 nm intervals. Tristimulus values obtained (a) by 10 nm interval spot measurements, and (b) by integrated transmissions over 10 nm bands, could then be compared with the "ideal" tristimulus values. When this is repeated over a large number of samples with steep and sharp transmission curves, it will be found that neither of the two "inferior" methods would be significantly better than the other.

8.4 Further Tests on the Spectrocolorimeter

Table 8.10 shows the C.I.E. Tristimulus values of 14 gelatin filters (Kodak Wratten) as computed from Kodak transmission values and those measured on the Harrison colorimeter and the Spectrocolorimeter. Table 8.11 shows the computed x,y values for these measuremnts. All these figures show a much higher correlation between the Kodak and the Spectrocolorimeter than between the Harrison and any of the other two.

The average percentage difference of the tristimulus values between the Kodak and the Spectrocolorimeter were as follows:

Percentage difference in X = ±1.28%

Percentage difference in Y = ±1.91%

Percentage difference in Z = ±2.32%

The average differences between the Kodak and the Harrison were:

Percentage difference in X = ±2.47%

Percentage difference in Y = ±4.10%

Percentage difference in Z = ±8.64%

Fig. 8.3 shows the chromaticity co-ordinates of the 14 filters on the C.I.E. chromaticity diagram. Also shown on this diagram are six MacAdam ellipses for comparison. Even though these ellipses have been disputed as to their validity (Crawford, 1969), they nevertheless provide a first order approximation, especially for comparison purposes. It can be seen from this diagram that the differences between the Kodak and the Spectrocolorimeter values are in the majority of cases either imperceptible or just about perceptible. The main exceptions being filters 34, 74 and 75. All these three filters have very low tristimulus values, so that spectrophotometric errors are very high.

Thus, these results show that the spectrocolorimeter is much more accurate than a common filter colorimeter and comparable to a good spectrophotometer-computer combination.

Kodak Filter Number	C.I.E. Values Calculated from Kodak Data			C.I.E. Values as Measured by the Harrison Colorimeter			C.I.E. Values as Measured on the Spectrocolorimeter		
	X	Y	Z	X	Y	Z	X	Y	Z
3	84.46	95.02	33.74	82.5	92.3	35.6	85.51	94.36	34.60
11	24.66	42.33	10.09	25.7	40.4	9.2	25.14	41.84	9.83
30	62.28	32.32	34.78	59.8	35.1	37.2	63.05	32.07	35.12
32	35.87	14.01	53.87	36.2	13.4	52.1	36.10	13.76	54.54
33	16.16	6.16	6.56	14.8	7.3	5.9	17.33	6.49	7.08
34	9.21	1.25	39.67	8.6	3.4	38.2	8.67	2.38	40.84
38	35.41	44.11	78.38	37.5	42.6	81.4	36.37	43.31	79.62
38 A	14.93	17.93	56.32	13.6	19.2	58.3	15.36	18.74	55.91
44	5.82	15.06	28.72	7.4	16.1	26.8	6.34	14.77	27.69
45	2.94	4.58	19.44	4.1	5.0	19.1	3.25	4.82	20.06
47 B	6.04	0.69	31.39	6.5	2.6	30.5	6.43	1.30	32.08
58	8.71	24.92	2.25	10.2	23.2	4.3	9.07	25.36	2.64
74	1.25	4.25	0.17	3.0	5.5	2.1	1.82	3.98	0.86
75	0.53	1.82	4.37	0.1	1.2	6.1	0.99	1.44	3.95

Table 8.10 C.I.E. Tristimulus values of 14 Gelatin Filters as calculated from Kodak transmission data, and as measured on the two instruments.

The Harrison is a manual three Filter C.I.E. equal energy distribution coefficients colorimeter.

Kodak Filter Number	Kodak Data		Harrison Colorimeter		Spectrocolorimeter	
	X	Y	X	Y	X	Y
3	.3961	.4456	.3921	.4387	.3987	.4400
11	.3199	.5491	.3413	.5365	.3273	.5447
30	.4813	.2498	.4527	.2657	.4841	.2462
32	.3457	.1351	.3559	.1318	.3458	.1318
33	.5598	.2131	.5286	.2607	.5608	.2100
34	.1836	.0249	.1713	.0677	.1671	.0459
38	.2242	.2794	.2322	.2638	.2283	.2719
38 A	.1674	.2011	.1493	.2108	.1706	.2082
44	.1173	.3037	.1471	.3201	.1299	.3027
45	.1092	.1697	.1454	.1773	.1155	.1713
47 B	.1585	.0181	.1641	.0657	.1615	.0327
58	.2427	.6947	.2706	.6154	.2448	.6841
74	.2207	.7488	.2830	.5189	.2733	.5976
75	.0786	.2711	.0135	.1622	.1552	.2257

Table 8.11 C.I.E. Chromaticity co-ordinates detained from the values in Table 8.10.

Fig 8.3 Measured and computed chromaticity values. Their differences compared against the MacAdam Ellipses.

Numbers followed by 'S' correspond to filters measured by the spectrocolorimeter. Those numbers without 'S' were computed from manufacturers data.

8.5 Conclusions

The final instrument is to a very large extent what was originally envisaged. This would apply both to the design and performance of the instrument.

The fact that a version of this instrument is already used industrially probably demonstrates its success more than anything else.

The manufacturing cost of such an instrument is estimated to be in the region of £1500-£2000. Thus the market price would be in the region of £4000. This is comparable to 3 filter colorimeters and, furthermore, it compares very favourably with spectrophotometer-computer combinations whose market price is in the region of £20,000.

The possibility of commercial exploitation is being considered by Unilever at the present moment.

REFERENCES

1. Alney W. de W., Proc. Royal Society London, A, CXCIII, 250 (1899)

2. Alney W. de W., Phil. Trans. of the Royal Society 193, 259 (1900)

3. Aston S.M et al: Trans. Illum. Eng. London, Vol. I, No.4, 1969

4. Band W. and Wright W.D., J. Opt. Soc. Am. 1930, 20, p. 381.

5. Birren F., "Psychological implications of color and illumination", Illuminating Engineering, May, 1962.

6. Birren F., Abstr. 2nd Congress Inter. Col. Ass., Colour 73, p. 179, London 1973

7. Brown P.K. and Wald G., Science, 144, p.45, 1964

8. Cheskin L., Secrets of Marketing Success, New York, 1967.

9. Crawford B.H., Tagungsbericht Internationale Farbtagung COLOR 69, Stockholm, 1969.

10. Da Vinci Leonardo, "Treatise on Painting", English Translation, London, 1877.

11. De Kerf, J.L.F., Accuracy of tristimulus computations, J.Opt. Soc. Am., 48., 334, 1958.

12. De Valois R.L. et al, J.Opt. Soc. Am., 56 P.966, 1966.

13. Donaldson R., Proc, Phys. Soc., 1935, 47, p.1068 and 1947, 59, p.554

14. Faure J.P., "Colour sells your package", Zurich, 1969.

15. Ferree, C.E. and Rand G., "Lighting and the hygiene of the eye", Archives of Ophthalmology, July, 1929.

16. Fick A., Hanbuch der Physiologie, Leipzig, Vogel, 1879.

17. Gautier, Paris, 1730.

18. Ghoete J.W.von, Farbenlehre, Stuttgart, 1810,

19. Grassmann H.,: Phil. Mag (4), 7, p.254, 1853.

20. Guild J.: Trans. Opt. Soc., 26, p.139, 1924.

21. Guild J.: Trans. Opt Soc., 27, p.106, 1925.

22. Guild J.: Phil. Trans. Royal Society, A, 230, p.149, 1931.

23. Helmholtz, H.V., Philos. Mag. (4), 4, 519, 1852.

24. Hering, E., "Outline of a theory of the light sense", 1878, English Translation, Harvard U.P., 1964.

25. Ives H.E., J.Frank. Inst., 180, p.673, 1915.

26. Ives H.E., J.Frank. Inst., 195, p.23, 1923.

27. Jameson D. and Hurwich L.M., J.Opt. Soc. Am., 51(1), p.46, Jan 1961.

28. Judd D.B., J.Research, Nat. Bur. Std., 42, 1, 1949.

29. Judd D.B. and Wyszecki G., "Colour in Business, Science and Industry", J.Wiley & Sons, N.Y. 1963.

30. Judd D.B. and Kelly, J.Research Natl. Bur., Std., 23, 355 (1939), RP 1239.

31. King G.F., The Analyst, 77, p.742, 1952.

32. König A., Z. Psychol. Physiol. Sinnesorgane, 4, 231 (1892).

33. Le Blond, Frankfurt, 1730.

34. MacAdam D.L., J.Opt. Soc. Am. 32, 247 (1942).

35. Marks W.B. et al, Science, 1964, 143, p.1181.

36. Maxwell J.C., Proc. Royal Society London, 10, 404, 484 (1860).

37. Middleton, A.R., "Visibility in Meteorology", University Press, Toronto, 1941.

38. Motokawa A., J.Opt. Soc. Am., 43, 1953.

39. Müller G.E., Z.Psychol. Erganzungsb., 17 and 18, 1930.

40. Newton, Sir Isaac, "Opticks", London 1704.

41. Padgham C.A., Vision Research, 8, p.939, 1968.

42. Padgham C.A. and Saunders, J.E., "The Perception of light and colour". G.Bell & Sons, London, 1975.

43. Pilditch J., "The Silent Salesman", London, 1961.

44. Rowe S.C.H., Ph.D Thesis, City University, London, 1973.

45. Schrodinger E., Sity Akad. Wiss. Wien, Abt. II a, 1925, 134, p.471.

46. Sprariskaya N.I., Optics and Spectroscopy (Translated by Opt. Soc. Am) 1959, of, p.424.

47. Stiles W.S., Phys. Soc. Year Book, p.44 1955.

48. Stiles W.S., N.P.L. Symposium on Visual Problems of colour (HM Stationary Office, London), 1958, p.211.

49. Strocka D., Abstracts 2nd Congress Int. Col. Assocn, Colour 73, p.453, London 1973

50. Svaetichin G.,Acta Physical. Scandi 39, 1953 Suppl. 134.

51. Wright W.D., Trans. Opt. Soc. London, 29, 225, 1927.

52. Wright W.D., Trans. Opt. Soc., 30, p.141, 1928.

53. Wright W.D., Proc. Roy. Soc. London, B115, 49 (1934).

54. Wright W.D., Proc. Phys. Soc. London, 53, 93, 1941.

55. Wright W.D., J.Opt. Soc. Am., 49, p.384, 1959.

56. Wright W.D., "The measurement of colour", Adam Hilger Ltd, London, 1969.

57. Wurtman, R.J., "Biological implications of atificial illumination", Illuminating Engineering, October, 1968.

58. Young, Lectures on Natural Philosophy, Kelland, 1802.

APPENDIX I

DEFINITION OF BASIC COLORIMETRIC CONCEPTS

A. Psychological Concepts

Hue

Hue is the attribute of a colour perception denoted by blue, green, yellow, red, purple, and so on.

Saturation

Saturation (also referred to objectively as Purity) is the attribute of a colour perception determining the degree of its difference from the achromatic colour perception most resembling it.

Chromaticness

Chromaticness is the attribute of a colour perception composed of the attributes hue and saturation.

Luminosity

Luminosity (of an area perceived as self-luminous) is the attribute of a colour perception permitting it to be classed as equivalent to some member of the series of achromatic colour perceptions (those not possessing a hue) ranging from very dim to very bright or dazzling.

Lightness

Lightness (of an object perceived as nonself-luminous) is the attribute of a colour perception permitting it to be classed as equivalent to some member of the series of achromatic object-colour perceptions ranging for light-diffusion objects from black to white, and ranging for regularly transmitting objects from black to perfectly clear and colourless.

B. Psychophysical Concepts

Primary Colours

Primary colours are the colours of three reference lights by whose additive mixture nearly all other colours may be produced.

Note 1: These colours are often chosen to be either red, green, and blue, or red, green, and violet.

Note 2: In accordance with the laws of additive colour mixture, non real primaries can be defined which have the useful property that any real colour can be represented by an additive mixture of positive amounts of the primaries (linear combination with positive coefficients).

Chromaticity Co-ordinates

The chromaticity co-ordinates of a colour are the ratios of each tristimulus value of the colour to their sum.

Note: The chromaticity of a colour is the colour quality of a light definable by its chromaticity co-ordinates.

Dominant Wavelength

The dominant wavelength of a colour is the wavelength of the spectrum colour that, when addively mixed in suitable proportions with a specific achromatic colour, yields a match with the colour considered.

Complementary Wavelength

The complementary wavelength of a colour is the wavelength of the spectrum colour that, when addively mixed in suitable proportions with a specific achromatic colour, yields a match with a specified achromatic colour.

Note: Every colour has either a complementary wavelength or a dominant wavelength. Some, but not all, colours have both.

Metameric Colours

Metameric colours are colour stimuli of identical tristimulus values but different spectral energy distributions.

Isomeric Colours

Isomeric colours are colour stimuli of identical spectral energy distributions (and tristimulus values).

Appendix II

Computer Programme for Colour Correction - Dye Formulation

POL111 BASIC V01-05

```
90 DIM P(12),Q(12),R(12),I(12),S(12),U(12),N(12),M(12)
95 DIM T(12),H(12),L(12)
96 DIM W(12)
100 U0=.200888\V0=.30726\N=0
110 E=5
120 P(1)=1.3035\P(2)=9.0825\P(3)=8.4055\P(4)=1.674
121 P(5)=.2235\P(6)=5.2965\P(7)=14.6995\P(8)=21.9295
122 P(9)=20.811\P(10)=11.179\P(11)=3.4035\P(12)=.635
125 Q(1)=.0225\Q(2)=.4645\Q(3)=1.7345\Q(4)=4.8865
126 Q(5)=12.0825\Q(6)=21.4705\Q(7)=24.602\Q(8)=18.2925
127 Q(9)=10.411\Q(10)=4.4765\Q(11)=1.259\Q(12)=.2295
130 R(1)=3.925\R(2)=44.8735\R(3)=48.2485\R(4)=17.7855
131 R(5)=3.8\R(6)=.6995\R(7)=.965\R(8)=.0315
132 R(9)=6.50000E-03
135 PRINT "ENTER LENGTH OF SAMPLE CELL"\INPUT N1
145 PRINT "ENTER SAMPLE TRANSMISSIONS"
150 INPUT W(1),W(2),W(3),W(4),W(5),W(6)
151 INPUT W(7),W(8),W(9),W(10),W(11),W(12)
153 FOR O=0 TO 5
154 Q=0
155 FOR I=1 TO 12
160 T(I)=W(I)^(1/N1)\NEXT I
170 F=0
175 L(1)=0\L(2)=0\L(3)=0\L(4)=0\L(5)=0\L(6)=0
176 L(7)=0\L(8)=0\L(9)=0\L(10)=0\L(11)=0\L(12)=0
180 A(O)=0\B(O)=0\C(O)=0\D(O)=0\G(O)=0\Z(O)=0\J(O)=0
185 X1=0\Y1=0\Z1=0
190 FOR I=1 TO 12
195 X1=X1+P(I)*T(I)*EXP(-L(I))
200 Y1=Y1+Q(I)*T(I)*EXP(-L(I))
205 Z1=Z1+R(I)*T(I)*EXP(-L(I))
210 NEXT I
214 PRINT "X1",X1,Y1,Z1
215 L1=4*X1/(X1+(15*Y1)+(3*Z1))
220 M1=6*Y1/(X1+(15*Y1)+(3*Z1))
225 W1=25*(Y1^(1/3))-17
230 U1=13*W1*(L1-U0)
235 V1=13*W1*(M1-V0)
260 N=N+1\IF N>1GO TO 400
275 PRINT "ENTER SAMPLE NO."\INPUT M
285 IF M=1GO TO 305 \IF M=2GO TO 315
295 IF M=3GO TO 325 \IF M=4GO TO 335
305 S(1)=.15\S(2)=.18\S(3)=.45\S(4)=.85\S(5)=.88\S(6)=.9
306 S(7)=.1\S(8)=.2\S(9)=.2
307 S(10)=.1\S(11)=.3\S(12)=.1
310 GO TO 340
315 S(1)=.15\S(2)=.15\S(3)=.15\S(4)=.35\S(5)=.85\S(6)=.9
320 GO TO 340
325 S(1)=.28\S(2)=.35\S(3)=.54\S(4)=.8\S(5)=.4\S(6)=.2
330 GO TO 340
335 S(1)=.2\S(2)=.3\S(3)=.4\S(4)=.5\S(5)=.6\S(6)=.7
340 X2=0\Y2=0\Z2=0
345 FOR I=1 TO 12
350 X2=X2+P(I)*S(I)
```

```
353 Y2=Y2+Q(I)*S(I)
356 Z2=Z2+R(I)*S(I)
359 NEXT I
363 PRINT X2,Y2,Z2
365 L2=4*X2/(X2+(15*Y2)+(3*Z2))
370 M2=6*Y2/(X2+(15*Y2)+(3*Z2))
375 W2=25*(Y2^(1/3))-17
380 U2=13*W2*(L2-U0)
385 V2=13*W2*(M2-V0)
400 IF (((U1-U2)^2+(V1-V2)^2+(W1-W2)^2)^.5)>EGO TO 435
415 IF 0=0GO TO 421 \GO TO 757
421 IF F=0GO TO 425 \GO TO 757
425 PRINT "MIX ACCEPTABLE"\GO TO 999
435 F=F+1
440 IF F<>1GO TO 442 \Q=Q+1
441 IF Q=8GO TO 760 \GO TO 460
442 IF F<>2GO TO 444 \Q=Q+1
443 IF Q=8GO TO 760 \GO TO 633
444 IF F<>3GO TO 446 \Q=Q+1
445 IF Q=8GO TO 760 \GO TO 648
446 IF F<>4GO TO 448 \Q=Q+1
447 IF Q=8GO TO 760 \GO TO 663
448 IF F<>5GO TO 450 \Q=Q+1
449 IF Q=8GO TO 760 \GO TO 678
450 IF F<>6GO TO 452 \Q=Q+1
451 IF Q=8GO TO 760 \GO TO 693
452 IF F<>7GO TO 454 \Q=Q+1
453 IF Q=8GO TO 760 \GO TO 703
454 IF F=8GO TO 455
455 F=1\GO TO 440
460 IF 0>0GO TO 630
465 GO TO 502
480 PRINT "ENTER REWORK TRANSMISSIONS"
485 INPUT U(1),U(2),U(3),U(4),U(5),U(6)
486 INPUT U(7),U(8),U(9),U(10),U(11),U(12)
490 FOR I=1 TO 12
495 N(I)=U(I)^(1/N1)
500 NEXT I
503 N(7)=.81\N(8)=.89\N(9)=.9\N(10)=.9\N(11)=.9\N(12)=.9
505 N2=1
508 X3=0\Y3=0\Z3=0\N1=1
509 FOR I=1 TO 12
511 H(I)=(LOG(1/N(I)))/N2
513 W9=1.00000E-05
514 M(I)=EXP(-H(I)*W9*N1)
517 X3=X3+P(I)*(T(I)^(1-W9))*M(I)*EXP(-L(I))
519 Y3=Y3+Q(I)*(T(I)^(1-W9))*M(I)*EXP(-L(I))
521 Z3=Z3+R(I)*(T(I)^(1-W9))*M(I)*EXP(-L(I))
523 NEXT I
526 D1=ABS((X2-X1)*W9*N1/(X3-X1))
529 D2=ABS((Y2-Y1)*W9*N1/(Y3-Y1))
532 D3=ABS((Z2-Z1)*W9*N1/(Z3-Z1))
534 PRINT "D1",D1,D2,D3
535 IF D1<D2 THEN 540 \D1=D2
540 IF D1<D3 THEN 541 \D1=D3
541 IF X1>X2GO TO 542 \GO TO 543
542 IF X3>X1GO TO 435 \GO TO 544
543 IF X3<X1GO TO 435
544 IF Y1>Y2GO TO 545 \GO TO 546
545 IF Y3>Y1GO TO 435 \GO TO 547
546 IF Y3<Y1GO TO 435
547 IF Z1>Z2GO TO 548 \GO TO 549
548 IF Z3>Z1GO TO 435 \GO TO 550
549 IF Z3<Z1GO TO 435
550 GO TO 890
554 FOR I=1 TO 12
555 L(I)=L(I)+((D1/N2)*LOG(1/N(I)))
```

```
560 L(I)=L(I)*(N1/(N1+D1))
565 T(I)=T(I)^(N1/(N1+D1))
570 NEXT I
610 GO TO 720
620 GO TO 635
630 IF 0>1GO TO 645
633 P=2
900 J=9

READY
```

THE PRINCIPLES OF OPERATION AND PERFORMANCE OF COMMON SOLID STATE PHOTOELECTRONIC DEVICES

By A. POLYDOROU

Introduction

In this report, an introduction to the theory of operation, limiting factors and the performance of the most common solid state photoelectronic devices is given. The treatment given is by no means exhaustive. It will be of use to potential users of these devices who cannot afford the time to do a thorough study before deciding which device to choose, but would like to have a rough idea of the performance of the most common devices. It is hoped that such a user will, more likely than not, belong to that group of users normally referred to as Industrial Control Engineers who have to design so many specialised systems. It is not intended for photoelectronic device designers and will only be of limited use to users who would like to choose the most suitable device to be incorporated in mass produced instruments.

The first three chapters give an exposition of the fundamental physics necessary to understand the operating mechanism of the devices described in the next two chapters. Chapter 6 gives a review of methods used for relative performance comparisons, whilst chapter 7 deals with actual performance figures. Chapter 8 covers the applicability of solid state photoelectronic devices, with main emphasis on some industrial instruments designed by the author.

The use of books and reviews have been extensive. A list of references is given.

Index

GLOSSARY (UNLESS OTHERWISE STATED)

δE = Energy gap, in eV.

n_i = Equilibrium concentration of electrons in intrinsic semiconductor, cm^{-3}.

p_i = Equilibrium concentration of holes in intrinsic semiconductor, cm^{-3}.

T = Absolute temperature, deg K.

h = Planck's constant = 6.62×10^{-34} joule sec.

K = Boltzman's constant = 1.38×10^{-23} joule deg K^{-1}.

m_e = Effective mass of free electron.

m_p = " " " " hole.

E_a = Energy specifying the position of acceptor levels in the forbidden band.

E_d = " " " bottom of the conduction band.

E_F = Fermin level.

N_a = Concentration of acceptor atoms, cm^{-3}.

N_d = " " donor " "

N_{au} = " " thermally unionized acceptor atoms, cm^{-3}.

N_{du} = " " " " donor " "

n = Free electron concentration, cm^{-3}.

p = Hole concentration, cm^{-3}.

q = Electronic charge, coulomb.

μ_e = Electron mobility, cm^2 volt $^{-1}$ sec $^{-1}$.

μ_p = Hole mobility, cm^2 volt $^{-1}$ sec $^{-1}$.

σ = Conductivity, ohm cm.

D_e = Electron diffusion constant, cm^2 sec $^{-1}$.

D_h = Hole diffusion constant cm^2 sec $^{-1}$.

S = Surface recombination velocity cm sec $^{-1}$.

δN_{sf} = Excess carrier concentration at the surface.

α = Absorption coefficient cm^{-1}.

Q = Photon flux density, photons cm^{-2} sec $^{-1}$.

ω = Angular frequency.

ε_o = Dielectric constant

L_p = Hole diffusion length

L_e = Electron diffusion length

1. ELECTRICAL PROPERTIES OF SEMICONDUCTORS FUNDAMENTAL TO PHOTOELECTRONIC DEVICES:

1.1 Electrical Conductivity

The electrical conductivity of a semiconductor can be described in terms of the energy band model:

The highest energy band, the valence band, is completely occupied by electrons at absolute zero. Electrons are agitated into the conduction band (completely empty at absolute zero) at higher temperatures. The vacancies created by the removal of the electrons are called holes.

The two bands are separated by an energy gap δE. Conduction can take place due to the generation of electron-hole pairs. This type of conduction and also that due to the passage of electrons between bands is called intrinsic conduction.

Electromagnetic radiation of suitable energy can also create electron-hole pairs. This is known as the internal photo-electric effect.

1.2 Factors Important to Photoconductors

1.2.1 The Energy Gap

This is the minimum energy required by an impinging photon to move an electron from the valence into the conduction band.

The semiconductor lattice is affected by temperature, which in turn causes variations in the energy gap.

1.2.2 Thermal-Equilibrium Free Carrier Concentration

Raising the temperature, raises the number of electrons in the conduction band, and thus the conductivity increases. The concentration of thermally generated carriers must be kept to a minimum, to maximize the effect of the excess carriers due to the impinging radiation.

The concentration of electron-hole pairs in an intrinisc semiconductor is given by:

$$ni = pi = 2\left(\frac{2\pi kT}{h^2}\right)^{3/2} (m_e m_h)^{3/4} \exp\left(-\frac{\delta E}{2kT}\right)$$

Thus, Silicon (δE = 1.1 eV) is a more suitable photoelectric intrinsic material than Germanium (δE = 0.68 eV), because of the lower concentration of thermally generated carriers.

1.2.3 Impurity Levels

Impurities might be of two types, ie. donor impurities (supplying electrons to the conduction band) or acceptor impurities (creating holes in the valence band). Either or both types of impurity could be present and they can affect the photoelectric functions.

The ionization energies of the impurities as well as their concentration could drastically affect conductivity. At high concentrations, the impurity level could be split into a band (just as the discrete energy levels of isolated atoms of the semiconductor split when the atoms are brought together to form a crystal lattice). The conductivity could increase to a very high value, if sufficient splitting of the impurity levels could cause them to join with the conduction band.

The concentrations of impurity atoms which are not ionized thermally are given by the relations

$$N_{du} = N_d \left[1 + 0.5 \exp\left(\frac{E_d - E_F}{kT}\right)\right]^{-1}$$

$$N_{au} = N_a \left[1 + 0.5 \exp\left(\frac{E_F - E_a}{kT}\right)\right]^{-1}$$

For the largest possible impurity photoconductivity N_d as well as N_a must be as high as possible so that the incident radiation can generate as many electrons or holes as possible. Also in the absence of incident radiation, the thermal ionization of carriers from the impurity levels should be negligible, ie the concentration of the un-ionized levels N_{du} and N_{au} should be as close as possible to N_d and N_a.

1.3 Conductivity and Carrier Mobility

The specific conductivity of a semiconductor can be expressed in the form

$$\sigma = q\,(n\mu_e + p\mu_h)$$

Carrier mobility is mainly limited by scattering due to collisions with the lattice atoms and with un-ionized impurity atoms. Mobility decreases with impurity concentration. It also decreases with temperature increase, as the thermal vibration of the lattice atoms increases, causing more intense carrier scattering.

1.4 Drift and Diffusion Currents

Drift current is produced by applying an external voltage E — producing a current density, i given by

$$i = \sigma E = qn\mu E$$

Diffusion current is due to variations in the carrier concentration, and is given by

$$i_e = qD_e \frac{du}{dx}$$

The diffusion current flows towards higher electron concentration — as the electrons move in the reverse direction. For holes, the motion and the diffusion current are in the same direction. Therefore the hole diffusion current density is given by

$$i_h = - qD_h \frac{dp}{dx}$$

The Einsteim relationship $Dq = \mu kT$, relates the diffusion constants to the mobilities.

1.5 Carrier Recombination - Trapping

Electrons and holes can recombine together - liberating their energy to the crystal. Dynamic equilibrium is set up, where the generation rate of hole-electron pairs equals the recombination rate. As momentum has to be

conserved in combining a hole and an electron, and as it is highly unlikely that a hole and an electron would have nearly equal and opposite momenta, direct recombination is very unlikely.

Recombination instead normally occurs at a "trap" level in the energy-band gap. A free electron leaves the conduction band, and enters the trap. The electron then releases some of its energy to the trap, and stays in the trap for a period, until it fills a passing hole.

These traps are associated with imperfections in the crystal. Specifically, metallic impurities in the semiconductor can introduce them in the forbidden gap.

Gold is extensively used as a recombination agent. Thus, desired carrier lifetimes may be obtained by introducing gold into CdS, say, under controlled conditions.

1.6 Lifetime and Diffusion Length

The electron lifetime, T_e, is the average time an electron spends in the conduction band before recombining with a hole.

The mean length to which carriers can diffuse from the point of generation is known as the diffusion length, L.

For electrons $L_e = \sqrt{T_e D_e}$

For holes $L_h = \sqrt{T_h D_h}$

1.7 Surface Recombination

Recombination is affected not only by volume impurities but also by surface imperfections in the crystal.

The number of carriers recombining per unit time per unit surface area is proportional to the excess carrier concentration at the surface $S \delta N_{SF}$. S is known as the surface recombination velocity ($cm\,sec^{-1}$).

At a free surface, in the absence of electrodes the net current flow is zero. Equal numbers of electrons and holes from the bulk of the semiconductor can reach the surface and recombine there. This number of electrons arriving at the surface per unit area per second is $\frac{i_n}{q}$ where i_n is the current density normal to the surface. Thus,

$$\frac{i_n}{q} = S\delta N_{sf}$$

When radiation is incident on the semiconductor surface, electron-hole pairs are generated in a very thin surface layer (1 micron). The flux $-\frac{i_n}{q}$ of electrons moving from the surface into the semiconductor bulk is equal to the number of pairs produced by the radiation minus the number of pairs produced at the surface. If g is the generation rate of electron-hole pairs

$$-\frac{i_n}{q} = -D_e \frac{dn}{dx} = g - S\delta N_{sf}$$

1.8 The Equation of Continuity

This equation describes the concentration of carriers in a certain region of a semiconductor. The carrier concentration might be increased by the entrance of carriers from neighbouring regions or due to generation of electron-hole pairs by incident radiation. The decrease in carrier concentration might be due to recombination or because of the carriers leaving the region due to an applied field or due to diffusion.

The continuity equation can be given in the form

$$\frac{\partial n}{\partial t} = g + \frac{1}{q}\frac{\partial i_e}{\partial n} - \frac{\delta n}{t}$$

Where $\frac{n}{t}$ is the rate of change of carriers in the volume under consideration, g is the rate of generation due to the incident radiation, i_e is the total current density, and $\frac{n}{t}$ is the rate of recombination.

1.9 Quantitative Theory of the p-n Diode Currents

The object is to derive an expression for the total current as a function of the applied voltage (the volt-ampere characteristic). The depletion layer thickness is considered negligible.

A forward bias causes holes to be injected from the p side into the n side. The concentration pn of holes in the n side therefore increases above its thermal equilibrium value pno and is given by

$$pn\ (X) = pno + P_n\ (0) \exp\left\{-\frac{X}{Lp}\right\} \qquad (1.9.1)$$

where $P_n\ (0)$ = the injected concentration at x = o, ie., $P_n\ (0) = pn\ (0) - pno$ (1.9.1a)

From paragraph 1.4 the diffusion hole current in the N side is given by

$$ipn = - AqDp \frac{dpn}{dx} \qquad (1.9.2)$$

for $\frac{dpn}{dx}$ from equation (1.9.1)

$$ipn\ (X) = \frac{AqDpPn\ (0)}{Lp} \exp\left\{- \frac{X}{Lp}\right\} \qquad (1.9.3)$$

This equation shows that the hole current decreases exponentially with distance. Since Pn (0) is a function of voltage, it follows that ipn also depends upon the applied voltage.

If Pp and pn are the hole concentrations at the edges of the space-charge region in the p and n materials respectively, then, by the Boltzmoun relationship,

$$P_p = P_n \exp\left\{\frac{qVB}{KT}\right\} \qquad (1.9.4)$$

where V_B = carrier potential across the depletion layer.

For an open-circuited p-n junction, $P_p = P_{po}$, $P_n = P_{no}$ and $V_B = V_O$ the contact potential, ie.

$$P_{po} = P_{no} \exp\left\{\frac{qVo}{KT}\right\} \qquad (1.9.5)$$

For a p-n junction forward biased by an applied voltage V, (at x = o),

$$P_{po} = P_n\ (0) \exp\left[\frac{q(Vo-V)}{KT}\right] \qquad (1.9.6)$$

From equations (1.9.5) and (1.9.6)

$$P_n(0) - P_{no} \exp\left\{\frac{qV}{KT}\right\} \qquad (1.9.7)$$

This boundary condition is called the law of the junction. It indicates that for a forward bias ($V>0$), the hole concentration $P_n(0)$ at the junction is greater than the thermal-equilibrium value P_{no}. A similar result, valid for electrons is obtained by interchanging p and n, in equation (1.9.7). The hole concentration $P_n(0)$ injected into the n side at the junction is obtained by substituting equation (1.9.7) into equation (1.9.1a), ie.

$$P_n(0) = p_{no}\left[\exp\left(\frac{qV}{KT}\right) - 1\right] \qquad (1.9.8)$$

Using equations (1.9.3) and (1.9.8), the hole current $i_{pn}(0)$ crossing the junction into the n side is given by,

$$I_{pn}(0) = \frac{AqD_pP_{no}}{L_p}\left[\exp\left(\frac{qV}{KT}\right) - 1\right] \qquad (1.9.9)$$

Similarly, the electron current $I_{np}(0)$ crossing the junction into the p side, is given by,

$$I_{np}(0) = \frac{AqD_nN_{po}}{L_n}\left[\exp\left(\frac{qV}{KT}\right) - 1\right] \qquad (1.9.10)$$

The total diode current, is therefore given by,

$$I = I_o\left[\exp\left(\frac{qV}{KT}\right) - 1\right] \qquad (1.9.11)$$

where $I_o = Aq\left[\frac{D_pP_{no}}{L_p} + \frac{D_nN_{po}}{L_n}\right]$

2. OPTICAL PROPERTIES OF SEMICONDUCTORS FUNDAMENTAL TO PHOTOELECTRONIC DEVICES:

2.1 Interaction of E.M. Radiation with a Semiconductor:

When E.M. radiation is incident upon a semiconductor a fraction of it is reflected, a fraction is transmitted and the remainder absorbed. Photoelectric devices converting E.M. energy to electrical, must be able to absorb as much incident ener as possible.

Photons are absorbed in a semiconductor by one of two processes:

a) A non-ionizing process whereby all or part of the energy of the incident photon is converted into another form of energy. In the case whereby a photon changes its frequency to a lower frequency (photoluminescence), the difference in the energies is given off as a phonon - a quantum of energy associated with the thermal vibrations of the crystal lattice. The entire energy of the incident photon could sometimes be converted into thermal energy as the result of the generation of a large number of phonons.

A photon can sometimes be absorbed by a semiconductor, to produce an exciton - a bound electron-hole pair which can move in the crystal transferring energy but no charge. Free carriers in the semiconductor can also absorb photons.

b) In the second type of process, the interaction between photons and the crystal lattice results in the ionization of lattice atoms and the generation of free carriers. Photo-emission can sometimes result if the transferred energy is high enough. If the photon interacts with an atom of the lattice an electron-hole pair is produced; if the photon interacts with a donor atom, an electron is released; if with an acceptor then a hole is released.

These processes describe the internal photoelectric effect.

In a homogeneous semiconductor these processes cause an increase in conductivity - photoconductivity.

In the presence of potential barriers, they cause potential difference to arise - photovoltaic effect.

2.2 Intrinsic Absorption:

The absorption of radiation by the valence electrons of the atoms constitutes the basic principle of most semiconductor photoelectric devices. As the minimum energy required by a photon to activate an electron from the valence into the conduction band must equal the energy gap δE, then this energy defines the threshold wavelength of the radiation that can be detected by a particular device.

The absorption coefficient, α, can be defined by the relation

$$Q(x) = Q_o \exp(-\alpha x)$$

Fan[1] has worked out the theoretical absorption coefficient of a semiconductor, and gives the formula

$$\alpha = \frac{2q^2 (2m)^{5/2}}{3m^2 h^2 cn_r} \quad \frac{(h\nu - \delta E)^{3/2}}{h\nu}$$

Photons with energy less than δE, can penetrate the semiconductor without attenuation ($\alpha = 0$). Normally photons are absorbed within a layer of about 1 micron thick.

As the energy gap is affected by temperature (see 1.2.1) the intrinsic absorption edge is also affected by temperature.

2.3 Extrinsic Absorption:

Impurity atoms with a lower energy gap than that of a semiconductor, can be ionized by photons with lower energy. If the ionization energy of the impurities is below their thermal vibration energy at room temperature, then thermal ionization can take place and thus cooling of the semiconductor would be required for efficient operation as a photoelectric device.

2.4 Free Carrier Absorption:

As free charge carriers are always present in a semiconductor at any temperature above absolute zero, incident photons can interact with these carriers, thus increasing their kinetic energy.

The absorption coefficient, , is given by

$$\alpha = \frac{\sigma}{E_o C n_r} \left[1 + \left(\frac{\omega \text{time}}{q} \right)^2 \right]^{-1}$$

Thus, α, may be modified by a change in the free charge carrier concentration or by a change in their mobility.

2.5 Exciton Absorption:

Absorption of incident photons may in some cases lead to the separation of the positive and negative charges, thus resulting in an electron and a hole bound together by their own fields - excitons. An exciton moves through a crystal in a random manner, by transmitting its excitation to a similar neighbouring atom.

The production of excitons by incident radiation, reduces the efficiency of photoelectric devices, as the energy absorbed has no effect on photoconductivity.

2.6 Absorption by Lattice Vibrations:

The vibrations of the ions in the lattice of ionic crystals cause a strong absorption band in the far infra-red region. The interatonic binding forces of ionic crystal lattices are comparable with the electronic forces which act within the atoms. As the ionic mass is of the order of 10^4 times larger than the electronic mass, the absorption associated with lattice vibration lies in the far infra-red. This type of absorption, also reduces the efficiency of photoelectric devices, as the energy absorbed has no effect on photoconductivity.

2.7 Quantum Yield:

This is the number of electron-hole pairs produced in a semiconductor per incident (or, in some cases, absorbed) photon. To be able to generate electron-hole pairs a photon must possess a minimum energy, δE, which is the energy required to free an electron from its valence bond.

3. NOISE IN PHOTOELECTRONIC DEVICES: [22,27]

There are several sources of noise that might affect the output of a photoelectronic device. The ultimate aim is to reduce noise which is generated in the detector itself. It is impossible though to improve on the noise which originates in the incident radiation.

3.1 Radiation Noise

Considering a statistical variation in the number of photons incident on the photodetector per unit time, then the mean deviation is equal to the square root of that number of photons. This photon fluctuation causes a proportional fluctuation in the number of carriers generated in the semiconductor, thus affecting the photoelectric current.

The mean square fluctuation $\left(\overline{\delta n^2}\right)$ per frequency interval of the emitted number of photons, (n) per unit area of a radiator is given by[2]

$$\left(\overline{\delta n^2}\right) = \frac{2n}{c^2} \frac{\nu^2 \exp\left(\frac{h\nu}{KT}\right)}{\left[\exp\left(\frac{h\nu}{KT}\right) - 1\right]^2} \qquad (3.1)$$

which may be derived from the fluctuations of radiations of a black body at a temperature T and volume V in the spectral region ν_1 to ν_2. Fluctuations in black body radiation are given by[2]

$$\left(\overline{\delta s^2}\right) = \frac{s_n V}{c^2} \int_{\nu_1}^{\nu_2} \frac{h^2\nu^4 \exp\left(\frac{h\nu}{KT}\right)}{\exp\left(\frac{h\nu}{KT}\right) - 1} \, d\nu \qquad (3.2)$$

Integrating over the total range of frequencies, then the mean square fluctuation of the incident power per unit bandwidth $\left[\overline{P^2}\,(F)\right]$ is given by 3.2

$$\overline{P^2}\,(F) = 8KTP_m = 8A\sigma KT^5 \qquad (3.3)$$

Where σ = the Stefan-Boltzmann constant

K = Boltzmann's constant

A = detector area.

Thus the photon noise spectrum has a constant amplitude which is independent of the frequency. At the detector output, P (F) should be multiplied by the square of the detector responsivity R (F), ie,

$$Pn(F) = R^2(F)\ P(F)$$

This causes a dependence of the output photon noise power spectrum, on the frequency response of the detector.

3.2 Shot Noise

Shot noise is generated due to the fluctuation of the general pattern of the flow of electrons (or holes) in a semiconductor. This corresponds to the random arrival of electrons at the anode of a vacuum tube.

The root mean square value of the current fluctuations $(\delta i^2)^{\frac{1}{2}}$, is given by[3]

$$(\delta i^2)^{\frac{1}{2}} = (2ei\delta f)^{\frac{1}{2}}$$

where F is the bandwidth of the measuring circuit at the output of the detector. The root mean square fluctuations of the voltage across the load resistor, can be given by,

$$(\delta v^2)^{\frac{1}{2}} = R\ (2ei\delta F)^{\frac{1}{2}}$$

There is a conceptual as well as numerical convenience to reduce the variety of all types of noise, to the same formalism used in shot noise.[3]

At the output of a detector with a responsivity R (F), the r.m.s. shot noise current should be multiplied by R (F).

3.3 Thermal or Johnson-Nuquist Noise

This noise is due to the fluctuations of the thermal velocities of free charge carriers. The r.m.s. noise voltage fluctuations, $(\delta v^2)^{\frac{1}{2}}$, across a resistor of R ohms is given by[3]

$$(\delta v^2)^{\frac{1}{2}} = (4KTR\delta F)^{\frac{1}{2}}$$

When the detector is loaded with a matched resistor, then the fluctuations voltage is reduced by a factor of four. Thus the noise power, would then be given by,

$$P_n = KT\delta f$$

Johnson noise power is therefore dependent on temperature but not resistance or frequency. For frequencies exceeding 10^{13} Hz quantum theory yields,

$$P_n = \frac{h\nu\delta f}{\exp\left\{\frac{h\nu}{KT}\right\} - 1}$$

3.4 Noise Due to Generation and Recombination

The number of thermally generated carriers shows a statistical fluctuation. Their lifetimes show a similar pattern. These fluctuations cause an additional noise at the photodevice output. A typical formula for the power spectrum is[2]

$$P_n(f) = \frac{4\bar{I}^2 T}{N_o\left[1 + (2\pi fT)^2\right]}$$

Where T is the carrier mean lifetime and N_o the mean number of carriers.

3.5 Modulation Noise

This type of noise is believed to be due to imperfections of the semiconductor surface. It has been observed that the statistical distribution of carriers can be affected by non-thermal causes. Thus the semiconductor conductivity is modulated by these additional fluctuations. The spectrum of modulation noise power is given by,

$$P_n(F) = \frac{CI^2}{f^a}$$

Where C is a constant, and a is a number close to unity. This relationship has been found to be obeyed from frequencies of the order of 2×10^{-4} Hz to 4×10^{6} Hz

3.6 Contact Noise

This is the noise generated at the metal-semiconductor contacts. It has a similar spectrum to modulation noise:

$$P_n(f) = \frac{CI^b}{f^a}$$

where a is again close to unity whilst b is close to two.

4. PHOTOCONDUCTIVITY

4.1 Lifetime τ :

The most general relation which characterizes photoconductivity is

$$N = F\tau$$

where N is the steady state increase in the density of free carriers in a given specimen and F is the number of excitations per unit volume per second occurring in the specimen. Alternatively this equation can be put in the form

$$\frac{N}{\tau} = F \qquad (4.1)$$

Where $\frac{N}{\tau}$ is the rate of disappearance of free carriers, and F their rate of generation, at steady state.

The conductivity, of the sample can then be given by

$$\sigma = nq\mu \qquad (4.2)$$

Thus, it is quite obvious that the lifetime of a free carrier describes the essential performance and contains the essential physics of a photoconductor.

The lifetime is mainly a function of the rate of recombination of carriers with their ground states. Recombination takes place via bound states in the forbidden zone. These bound states are composed of impurities, vacant lattice cites and crystal defects. Thus lifetime is not a constant of the photoconductor material, but depends on the way of preparation, impurity content etc.

The lifetime usually varies inversely as the light intensity. Sometimes though, it increases with ligh intensity, so that the photocurrent varies as a power greater than unity of the incident radiation - superlinearity.

It is similarly common to find photocurrents increasing, decreasing or remaining invariant with temperature.

Lifetimes in some commonly known sensitive photoconductors range from 10^{-3} - 10^{-5} seconds in PbS, to 10^{-2} - 10^{-3} seconds in CdS, and CdSe.

4.2 Photoconductive Gain and Sensitivity

The current observed between two points is equal to the rate of generating carriers multiplied by the ratio of their statistical lifetime in the region between the two points to their transit time between these points.

$$I = qF \frac{\tau}{Tr} \text{ amperes} \qquad (4.3)$$

where Tr is the transit time of a carrier between electrodes and is given by

$$Tr = \frac{L}{E\mu} = \frac{L^2}{V\mu} \qquad (4.4)$$

where L is the electrode spacing in cm.

A gain factor, G, may be defined as the number of electron charges passing through the photoconductor per absorbed photon. Thus,

$$G = \frac{I}{qF} = \frac{\tau}{T_r} \qquad (4.5)$$

Photoconductive gain is similar in principle to quantrum efficiency, though the latter term is normally reserved for primary excitation processes.

Alternatively a photoconductive sensitivity Sp, may be defined by the relationship[1].

$$I = I_d + S_p P \qquad (4.6)$$

where I is the total current in amp, I_d is the dark current, P the incident radiation power in watts and S_p the sensitivity in amp $watt^{-1}$.

The dark current can be obtained from

$$I_d = \sigma_o \frac{wd}{l} V \qquad (4.7)$$

Where σ_o = dark conductivity in ohm^{-1} cm^{-1} and w the thickness, d the width and l the length of the photoconductor.

Similarly for the photoelectric current

$$I_{ph} = \delta\sigma \frac{wd}{l} \quad V \qquad (4.8)$$

Where $\delta\sigma$ is the increase in conductivity due to the incident radiation.

In the range where the photoelectric current is proportional to the incident radiation intensity, the sensitivity is given by

$$S_p = \frac{I_{ph}}{P} = \frac{\delta\sigma}{P} \frac{wd}{l} \quad V \qquad (4.9)$$

Thus, S_p is proportional to the increase in the semiconductor conductivity per unit increase in the radiation power and also to the supply voltage. This is a very important relationship, as, unlike other photodetectors, a photoconductive detector can increase its sensitivity by increasing the applied voltage.

The sensitivity in the visible region only can sometimes be given in amp per lumen.

Equations 4.7, 4.8 and 4.9 are valid provided the diffusion length for carriers is much larger than the thickness of the photoconductor. Where this is not the case, S_p reaches a limiting value as the thickness is increased. But the dark current continues to be proportional to the thickness of the semiconductor. Therefore, as the semiconductor thickness increases, the ratio of the photoelectric current to dark current (signal to noise ratio) decreases.

Typical values of sensitivities are 10A/W for Cadnium Sulphide and Cadmium Selenide. (At very low intensities Cadmium Sulphide is much more sensitive than Cadmium Selenide).

4.3 Charge Amplification

Intrinsic photoconductors have very limited use in practice. Carefully chosen added impurities can have drastic effects on the properties of photoconductors. Their spectral response can be controlled by the

addition of impurities. Their sensitivity as well as their response time can also change.

When an electron-hole pair is produced in a photoconductor, across which a potential is applied, the electron and the hole may travel relatively unhindered to the electrodes. There is a high probability though that they may be captured in a trap before they travel very far. They can stay in a trap until they are either neutralized by an opposite charge passing near the trap, or else until they are released by thermal energy.

In n-type Cadmium Sulphide the capture cross-section for the electron is much lower than that of the hole. Thus whilst the electron can move easily to the anode, the hole might be trapped, thus leaving a positive charge in the photoconductor material. This positive charge can now attract another electron from the cathode to enter the material. This new electron can now either recombine with the trapped hole, or else pass directly to the anode. A good photoconductor should offer a very low capture cross section for the recombination process of trapped holes and electrons. Indeed, Cadmium Sulphide is a good photoconductor in this respect. The probability that the new electron from the cathode will pass directly to the anode is quite high. Thus, a single trapped hole can cause a large number of electrons to cross the semiconductor material producing a current (secondary current) flow. This will continue until the charged hole is neutralized, by a passing electron. Since the original electron-hole pair was produced by a single incident photon, it is obvious that a quantum efficiency of greater than one can easily be obtained.

The photoconductive gain can thus be defined as the ratio of the mean time during which the holes are trapped, to the mean transit time of the electrons through the semiconductor. This confirms equation 4.3.

Photoconductive gains of the order of 10^5 are common, and even 10^6 have been reported. Thus the effective gain of a sensitive photoconductor is of the same order as that of an average photomultiplier.[4]

4.4 Simplified Mathematical Treatment of Photoconductivity[5]

(a) In the Absence of Traps

For simplicity consider only one type of current carrier, say electrons which are thermally created at a constant rate A. Their number, N, at any moment would decay at a rate proportional to N, ie.

$$\frac{dN}{dt} = A - \frac{N}{T} \qquad (4.10)$$

T = constant.

At equilibrium, $N = N_o = AT$

For a departure δN from N,

$$\frac{d(\delta N)}{dt} = -\frac{\delta N}{T}$$

Such a departure δN will decay exponentially with a time-constant T.

Consider now, incident radiation of intensity I, creating carriers at a rate γI. We then have,

$$\frac{d(\delta N)}{dt} = \gamma I - \frac{\delta N}{T}$$

If at $t = 0$, $N = N_o$, then

$$\delta N = \gamma IT \left[1 - \exp\left(-\frac{t}{T}\right)\right] \qquad (4.11)$$

The equilibrium value δN_f of δN is given by

$$\delta N_f = \gamma IT \qquad (4.12)$$

If after equilibrium the radiation is cut off,

$$\delta N = \delta N_f \exp\left(\frac{-t}{T}\right) \qquad (4.13)$$

From (4.11) and (4.13), it can be seen that the number of current carriers rises and decays exponentially with a time constant T.

It can also be seen from (4.12), that the equilibrium value δN_f is proportional to the time constant T. From basic considerations, it can be assumed that the change in resistance δR is proportional to δN, ie. the change in resistance of the photoconductor is proportional to T. This indicates that a very sensitive photoconductor, would have a long time constant.

In the above treatment, the effect of holes which might have different mobilities, has been neglected. The effect of traps, which greatly increase the time constant, has also been neglected.

(b) Trapping Effects

Consider, a p-type semiconductor. Suppose that incident radiation, produces δN electrons in the conduction band and that δN_t electrons are tapped.

Suppose that electrons in the conduction band, decay with a time constant T_1 by recombination with holes.

Suppose that electrons in the conduction band, decay with a time constant T_2 into traps, and that they are re-excited from the traps into the conduction band by a time constant T_3.

As stated earlier on, for a sensitive photoconductor, the electron should stay considerably longer in a trap, than it takes to be caught, ie. $T_3 \gg T_2$. In this treatment, direct recombination of electrons and holes, whilst the electrons are trapped, is excluded.

We therefore have, when radiation is incident on the photoconductor,

$$\frac{d(\delta N)}{dt} = \gamma I - \frac{\delta N}{T_1} - \frac{\delta N}{T_2} + \frac{\delta N}{T_3} \qquad (4.14)$$

$$\text{and, } \frac{d(\delta N_t)}{dt} = \frac{\delta N}{T_2} - \frac{\delta N_t}{T_3} \qquad (4.15)$$

At equilibrium,

$$\delta N_{tf} = \frac{T_3}{T_2} \delta N_f \qquad (4.16)$$

For the condition $T_3 \gg T_2$, there will be many more electrons in traps, than there will be in the conduction band.

It can be seen, that also in the steady state,

$$\delta N_f = \gamma I T, \qquad (4.17)$$

This shows that in the steady state, the number of electrons in the conduction band, is not affected by the presence of traps.

In order to preserve electrical neutrality, the number of "extra" holes in the valence band must equal $\delta N + \delta N_t$, so that in the steady state,

$$\delta N_{hf} = \delta N_f + \delta N_{tf} \qquad (4.18)$$

For the condition $T_3 \gg T_2$ to apply, we must have $\delta N_{hf} \gg \delta N_f$. Thus the change in conductivity, will be mainly a function of δN_{hf}.

From equations (4.16) and (4.18),

$$\frac{\delta N_{hf}}{\delta N_f} = 1 + \frac{T_3}{T_2}$$

and since $T_3 \gg T_2$,

$$\frac{\delta N_{hf}}{\delta N_f} \simeq \frac{T_3}{T_2} \qquad (4.19)$$

This is a very important result. It shows that the ratio of the conductivities in the presence and in the absence of traps is equal to the ratio $\frac{T_3}{T_2}$. Thus, the traps have the effect of increasing the response of the photoconductor by a factor $\frac{T_3}{T_2}$.

4.5 Speed of Response

A close examination of equations (4.14) and (4.15) can show what happens when the incident radiation is switched off. As T_1 is much smaller than T_2, δN will decay quickly with a time constant T_1. Likewise, the rate of decay of δN_t will initially be zero, but it will very quickly change over to a rate governed by T_3.

As photoconductivity is mainly governed by δN_h, which is determined by δN_t, photoconductivity will decrease at a rate determined by T_3. Thus, the effect of trapping is not only to increase the responsivity of the photoconductor, but also to increase the effective time constant.

In practice all these effects are easy to observe. When radiation is first incident on a material, free charge carriers are trapped and an appreciable time passes before reaching an equilibrium between the new free charge carrier density and the fraction of the trapping centres occupied. Likewise, when the radiation is reduced, the charged carriers are released relatively slowly, so that a long time can elapse for the new equilibrium value of photoconductivity to be obtained.

As the rate at which trapped electrons are removed is approximately proportional to their concentration, when the incident radiation is removed, the time constant by which the current decreases, can be quite large.

The physical significance of equations (4.14) and (4.15) is that the decay of photoconductivity can take place in two stages[6]. The initial decay stage is caused by recombination of free holes with electrons. The succeeding slower decay stage is due to the recombination of electrons, which have been thermally liberated from trapping centres. As, at low radiation intensities, the density of the trapped carriers may be much greater than that of the free carriers, the decay at low intensities is mainly of the second type, where the decay time is more dependent on the time required for a carrier to be liberated from a trap than by the free carrier lifetime. Thus, at low radiation intensities, the decay can be incredibly slow.

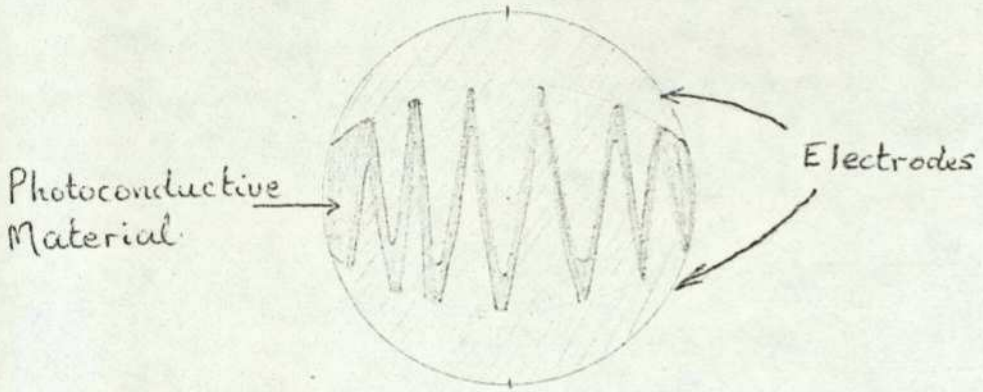
It can be seen from the above, that the behaviour of such photoconductors is to some extent dependent on the light history of the material. This behaviour excludes photoconductors as detectors for good quality light meters, unless certain conditions (such as exposing the detector to a constant light level previous to measurement, such an instrument has been designed by the author) are met.

The rise time constant can be defined as the time taken for the conductivity to rise to 63% of its final value, when the photoconductor is suddenly irradiated. Likewise, the fall time constant can be defined as the time required for the conductivity to fall by 63% of its original value, when the cell is suddenly placed in darkness.

Typical time constants at normal levels of light illumination are 100 mS for Cadmium Sulphide and 10 mS for Cadmium Selenide.

4.6 Construction Details of Photoconductive Cells

The most common structure of a photoconductive cell is shown below:



The doped semiconductor film is deposited on an insulating substrate and the electrodes are formed by evaporating a metal such as gold or indium on to the surface of the film. A mask is used during the deposition of the electrodes so that they are interleaved in a comblike pattern. This geometry enables a large sensitive area to be employed between closely spaced electrodes. A high sensitivity is thus obtained.

The effective "length" of the semiconductor is also quite small, as required by equations (4.4) and (4.5) for high photoconductive gain.

It was quite clear even with early studies, that both the resistance and optical absorption are very much dependent on the layer structure. Even though CdS layers have been prepared, where with careful choice of optimum pressure and rate of evaporation, properties approaching those of a single crystal have been achieved, it is generally accepted that the crystallinity of the layer is much less perfect than that of a single crystal.[7]

Effects associated with crystal imperfections (such as trapping) may determine the performance of a layer.

Single crystals, are therefore prepared in cases where systematic studies of these materials is to be undertaken.

In the early investigation stages of photoconductivity, many measurements were made on powder samples. Such photoconductors are not however reliable, because of the effect of particle-to-particle contacts. Large sensitivities can however be obtained by this method, due to the large area employed.

A variation of the powder layer which approaches the characteristics of a single crystal is the sintered layer. Sintered layers of Cadmium Sulphide or Cadmium Selenide are made by spraying a water mixture of the sulphide or selenide with chloride and copper onto a suitable surface. The surface is then fired so that the dried layer forms a polycrystalline sintered layer. This method is very useful in producing large-area photoconductors with crystal-like characteristics without involving the difficulties met in growing large-area single crystals or in producing large-area photoconductor layers by evaporation in vacuum.

The author has recently carried out a series of tests to compare the performances of the evaporated layer type of photoconductor to that of the sintered layer type. Three gallium arsenide light emitting diodes were used as sources, and these and the photocells were placed in a specially designed box whose temperature could be varied and accurately controlled. The sintered layer type was found to be much more sensitive for all three wavelengths used (5500 A, 5800 A and 6328A). It was also found to have a much higher speed of response and a much better temperature coefficient. Generally it also showed better behaviour as far as history or light memories were concerned. Both types were made of Cadmium Sulphide.

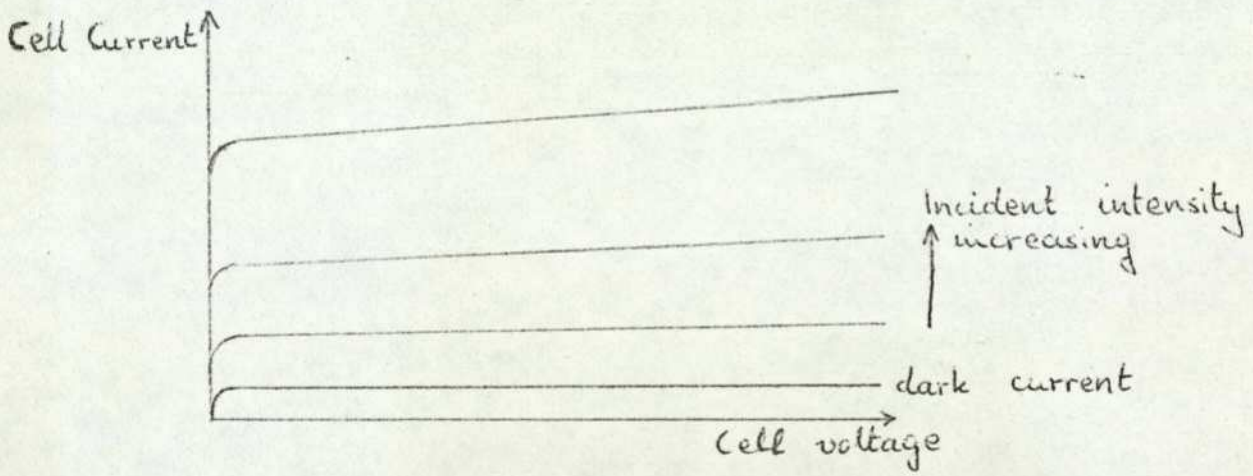
5. JUNCTION PHOTOELECTRIC DEVICES

5.1 Junction Photodiodes

Junction photodiodes are mostly used as photovoltaic cells for convertion of radiant energy to electrical; if however, they are reversed biased they may be used as photoconductive devices. In either way of operation they can be used as detectors of the intensity of incident radiation. Furthermore, as photovoltaic cells they can be operated in the open circuit mode (logarithmic response) or in the short circuit mode (linear response).

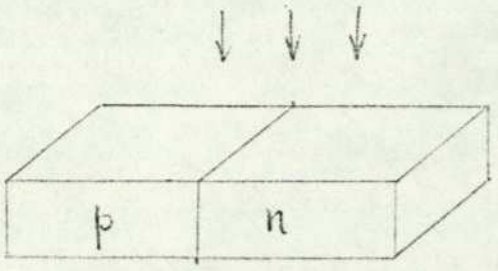
When radiation falls on to a pn junction, photons with energy greater than that of the band gap will produce charge carriers. These charge carriers, do predominately affect the concentration of minority carriers. As, in a reverse biased diode, the minority carriers pass through the junction, the current passing through the device is therefore increased.

The characteristics of a photodiode are shown in the diagram below:



In the absence of an external bias to the pn junction, the impinging photons will produce electron-hole pairs which will be separated by the natural junction potential and each charge carrier will move to the side of the junction where it is a majority carrier. This increased concentration of majority carriers on each side of the junction reduces the natural junction potential, resulting in p side becoming positive with respect to the n side of the material.

5.2 Properties of Parallel Illuminated Junctions



For low current densities through the junction the drift current is negligible as compared with the diffusion current. The entire applied voltage can be considered to be across the transition region of the p-n junction. The continuity equation for electrons in the p-type region is thus,

$$\frac{\partial \delta_n}{\partial t} = g + D_e \frac{\partial^2 \delta_n}{\partial x^2} - \frac{\delta_n}{T_e} \qquad (5.1)$$

Similarly for holes in the n-type region.

Solving equation (5.1) for i_e and adding the hole current i_h, we obtain the total current, 1,2

$$\left.\begin{aligned} I &= \left\{ q \left(\frac{D_p P_n}{L_p} + \frac{D_e N_p}{L_e} \right) \left[\exp\left(\frac{qV}{KT}\right) - 1 \right] - qg\,(L_p + L_e) \right\} A_j \\ &= I_s \left[\exp\left(\frac{qV}{KT}\right) - 1 \right] + I_{ph} \\ &= I_d + I_{ph} \\ &= I_d - S_p P \end{aligned}\right\} \qquad (5.2)$$

Equation (5.2) is the equation of a family of current voltage characteristics of the photodiode. It is interesting to note that I_{ph}, representing the current due to carriers generated by the radiation within a diffusion length of the potential barrier in the junction, is independent of the applied voltage. This is contrary to what is observed in a homogeneous photoconductor.

It can also be seen from equation (5.2), that the photoelectric current is directly proportional to the bulk carrier generation rate g and to the sum of the diffusion lengths of the carriers on each side of the junction.

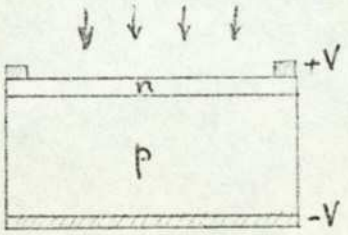
Further analysis of the complete solution of the continuity equation for a parallel-illuminated junction, reveals that the sensitivity of the photodiode is independent of the cross-sectional area of the junction.

It is furthermore found out that the only design parameter which affects the sensitivity of the photodiode is the length l of the p and n type regions, or strictly speaking, the ratio of their lengths to the diffusion lengths of minority carriers in these regions. It is also found that the dark current is also a function of this ratio.

The ratio of the photoelectric current to the dark current is shown to be inversely proportional to the wafer thickness.

5.3 Properties of Perpendicularly Illuminated Junctions

This method of illuminating is the most common. It functions as shown below:



According to R.Cummerow[8], the continuity equation in this case, has the form for holes in the n-type region.

$$D_p \frac{\partial^2 \delta_p}{\partial x^2} + g(x) - \frac{\delta_p}{\tau_p} = 0 \qquad 5.3$$

Similarly for electrons in the p-type region. Solving equation (5.3) for the hole component of the current in the n-type region, and also the similar equation for the electron component of the current in the p-type region, the total current can be written in the form:

$$
\left.
\begin{aligned}
I &= \left\{ q(g_p L^1_p + g_n L^1_n) \left[\exp \frac{(qV)}{(KT)} - 1 \right] - qg(0)(L_1 + L_2) \right\} A_j \\
&= I_s \left[\exp \left(\frac{qV}{KT} \right) - 1 \right] + I_{ph} \\
&= I_d + I_{ph} \\
&= I_d - S_p P
\end{aligned}
\right\} \qquad (5.4)
$$

where $g_n = \frac{N_p}{T_n}$ and $g_p = \frac{P_n}{T_p}$ are the rates of thermal generation of minority carriers in both regions.

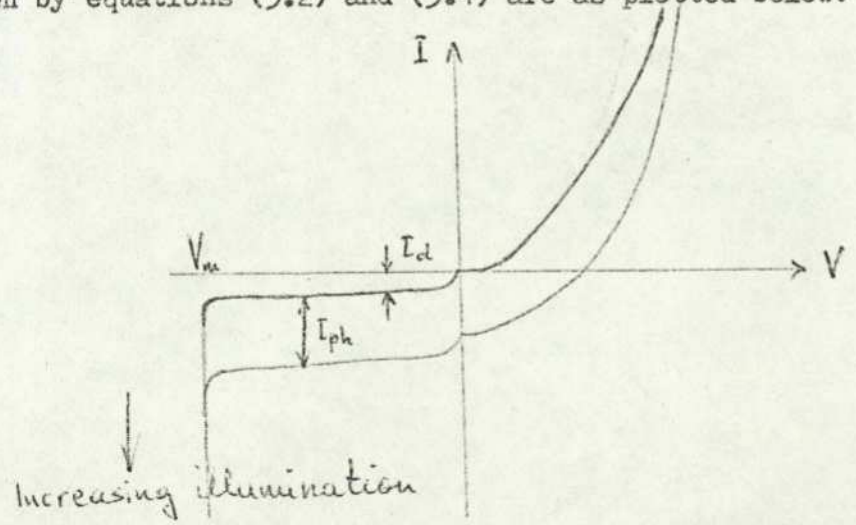
L^1_p and L^1_n are effective diffusion lengths.
L_1 and L_2 are coefficients depending on L_p and L_n.

Equation (5.4) represents the family of current-voltage characteristics of a photodiode. Cummerow[8] shows in his paper the conditions necessary for maximum sensitivity. To obtain a high photodiode sensitivity, the illuminated region of the junction must be sufficiently thin for the generated carriers to be able to reach the potential barrier, while the back region should be sufficiently thick to enable complete absorption behind the barrier, for the radiation which penetrates the face region. This of course is quite a logical conclusion.

It is furthermore shown that the ratio of the photoelectric to the dark current in a photodiode with a perpendicularly illuminated junction, is independent of the p-n junction area.

5.4 General Remarks on the Properties of Photodiodes

The equations of the family of current-voltage characteristics of photodiodes as given by equations (5.2) and (5.4) are as plotted below:



The normal operation of the photodiode is under a reverse bias condition ie. in the third quadrant of this diagram. The branches of the characteristics as shown in the first and fourth quadrants describe the condition where an external bias is not applied, ie. the photodiode is operating as a photoelectric cell.

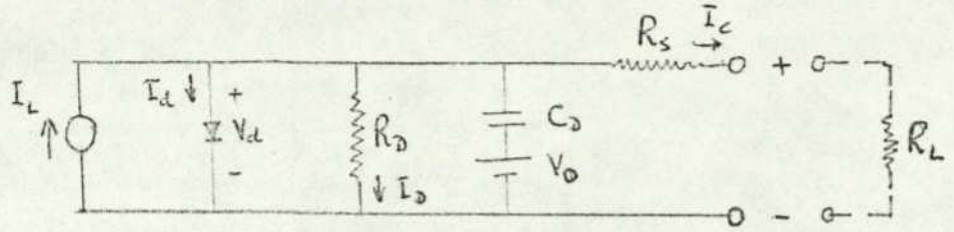
As in normal operation, the photodiode always passes current in the reverse direction, the current and voltage axes are usually reversed so that the operating region is in the first quadrant as shown in the diagram of section 5.1.

The spectral response of a pn junction photodiode is determined by the semiconductor material used, just as in the case of homogeneous photoconductors. Response at the long wavelength end of the spectrum is controlled by the energy band gap of the material used, since radiation at long wavelengths might not have sufficient energy to break a valence band in that material. At the short wavelength end of the spectrum, the response is controlled by surface phenomena. Short wavelength photons are absorbed near the surface of the cell and the electron-hole pairs created by p-n junction. Hence, they are not separated at the junction and do not contribute to the cell output current.

For silicon which is by far the most commonly used material for p-n junctions, the long wavelength range is approximately 1200 nm, the exact value being modified by the impurities present. The short wavelength end is around 350 nm, the exact value again depending on surface phenomena. Two methods that have only recently been used to improve the short wavelength response of silicon photodevices have been firstly the use of shallower junction depths and secondly the use of antieffective coatings on the surface[9].

5.5 Equivalent Circuit Model of a Photodiode

The most significant properties of a junction photodiode can be obtained from the equivalent circuit shown below:[9, 10]



The generated current I_L is proportional to the intensity of the incident radiation. The current through the diode is given by equation 1.9.11, viz

$$I_d = I_o \left[\exp \left(\frac{qVd}{KT} \right) - 1 \right] \qquad (5.5)$$

where I_o is the diode saturation current. In the circuit shown above, only the diode is a non-linear element. R_D is the depletion layer resistance. R_S is the series resistance in the material itself and the leads.

C_D is the depletion layer capacitance. R_L is the load impendance. V_o is the pn junction potential.

5.6 Linear Operation of the Photodiode - Short-Circuit Mode

Linear operation is achieved when only an insignificant proportion of the generated current passes through the diode which is a non-linear element. This implies a very low lead impedance - short circuit being the optimum method of operation.

Consider a constant level of illumination. Thus the effect of C_D can be neglected. Then by Kirchoff's law,

$$I_L = I_d + I_D + I_C \qquad (5.6)$$

Using equation (5.5), and the equivalent circuit,

$$I_L = I_o \left[\exp \left(\frac{qVd}{KT} \right) - 1 \right] + \frac{Vd}{R_D} + \frac{Vd}{R_L + R_S} \qquad (5.7)$$

$$\text{Hence} \quad IC = I_L \left[1 + \frac{R_L + R_S}{R_D} + \frac{I_o (R_L + R_S) q}{KT} \right]^{-1} = KI_L \qquad (5.8)$$

This is true, provided $I_C (R_L + R_S) \ll \frac{KT}{q}$, which is the condition for linear operation of the photodiode.

Inclusion of the capacitance C_D does not alter the result that for linear operation R_L and R_S have to be as small as possible.

5.7 Logarithmic Operation of the Photodiode - Open Circuit Mode

Logarithmic response can be achieved when an insignificant proportion of the light generated current passes through the load, ie. the load has to be as high as possible - open-circuit being the optimum method of operation. Under such circumstances,

$$I_L = I_d + \frac{Vd}{R_D} \qquad (5.9)$$

$$\text{ie. } I_L = I_o \left[\exp \left(\frac{qVd}{KT} \right) - 1 \right] + \frac{Vd}{R_D} \qquad (5.10)$$

$$\text{Thus } Vd = \frac{KT}{q} \log_e \left[\frac{I_L}{I_o} - \frac{Vd}{I_o R_D} + 1 \right] \qquad (5.11)$$

The conditions for achieving logarithmic response with respect to I_L are therefore,

(a) $\frac{I_L}{I_o} \gg 1$

and

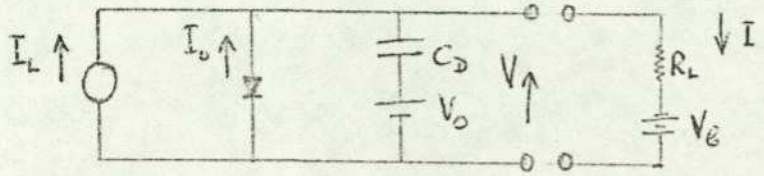
(b) $\frac{I_L}{I_o} \gg \frac{Vd}{I_o R_D}$

It can be seen from condition (a) that since I_L is proportional to the intensity of the incident radiation, the response deviates from a logarithmic response, at low intensities.

With this mode of operation Vd is very temperature dependent - for silicon about ten times worse than in the linear mode. From equation (5.5) it can be seen that I_d is dependent on I_o. But near room temperature, I_o doubles every about $10^{\circ}C$ temperature rise, and I_d varies at the same time. Thus V_d (the open circuit voltage of the photodiode) is very temperature dependent.

5.8 Photoconductive Operation of the Photodiode[10]

The equivalent circuit in this mode of operation is shown below:



In this mode of operation the photodiode junction is reverse biased by an external voltage Vb, as shown above. Thus, the current through the junction is in the same direction as the current from the light current generator.

The photodiode steady steady equation can in this case be given by,

$$V = - V_b + (I_L + I_o) R_L \qquad (5.12)$$

It can be seen that for this condition R_S is negligible and has been neglected. Thus the current through the junction reduces to the saturation current I_o. Therefore the output voltage across R_L is due to the effect of sum of I_L and I_o.

5.9 The Phototransistor

This is a much more sensitive device than the p-n photodiode. In construction it is identical to an ordinary transistor, except that the base-collector junction is uncovered so that light can fall on it.

In an npn phototransitor, the current I consists mainly of electrons moving from the n type emitter region through the slightly forward biased emitter-base junction, into the p type base and then through the reverse biased base-collector junction, into the n type collector.

As usual, the main obstacle to this current flow is the base-collector barrier. When this barrier is irradicated, an incident photon can create

a hole-electron pair, in the p-type base. Since these charge carriers are produced quite near to the barrier, the electrons will be attracted into the collector by the field in the depletion region. The holes are however trapped in the p type base by the potential energy hills at the two junctions on either side of the base. These trapped holes constitute a net positive charge causing an additional bias to appear at the emitter base junction. Additional electrons will therefore now cross the junction into the base, some of them neutralizing the holes by direct recombination, whilst the majority find their way through the base and into the collector.

Thus the primary holes created by the incident photons act as a trigger, causing a number of electrons to leave the emitter and travel through the base to the collector to finally constitute the output current.

The output current of a phototransistor exceeds that of a photodiode by a factor equal to the current gain of the transistor, this factor normally being between 50 and 100.

The spectral response of a phototransistor is identical to that of a photodiode, it is dependent on the semiconductor material but modified slightly by impurities.

It must be noticed that even though a higher light current can be obtained from a phototransistor than from a photodiode, the dark current (ie, the collector leakage current, which is very temperature sensitive) is amplified by the same amount. The signal to noise ratio can be improved by incorporating a resistor between the base and the emitter.

Quite often, silicon planar phototransistors can be used without any connection to the base, as their leakage current can be negligible, viz 10^{-9}A.

The phototransistor may not provide usable output signal levels at low irradiance levels. Due to a drop in gain at low currents, a phototransistor may not produce a higher signal output than a photodiode, at low intensities of incident radiation.

As the generation of the charge carriers occurs at varying distances from the base-collector junction, it is obvious that individual electrons will take different times to travel to the collector. This, very much limits the frequency response of the phototransisitor, which is much worse than that of the photodiode - a few kHz, as compared to a few MHz for a photodiode.

Another disadvantage of the phototransistor is that it can only be produced with very small sensitive areas. Modern photodiodes can be produced with very large sensitive areas. Therefore, the sensitivity of the phototransistor is very much limited in this respect.

The main application of a phototransistor is in light switches.

5.10 Photofets

Photofets is the short name for light sensitive field effect transistors. In this device the incident radiation produces electron-hole pairs in the gate junction region. These charge carriers form the gate current, which causes a change in the gate potential when flowing through the gate resistor. This change of the gate potential multiplied by the mutual conductance of the FET, is equal to the change in the output current flowing through the drain resistor. Thus the sensitivity of the Photofet is determined by the value of the gate resistor.

The sensitivity may vary from 20 nA to mA drain current when the gate resistor is changed from zero to 500 M ohms, for an illumination of around 10 lux.

The input time constant increases and therefore the response time increases, when large values of gate resistors are used to increase the sensitivity.

Also the response time can be increased, when a large resistor is used in the drain circuit, which effectively increases the input capacitance by the Miller effect. The highest speed response can be obtained by connecting the device as a source follower.

Photofets have got a greater sensitivity than phototransistors, as well as an increased gain-bandwidth product.

Photofets, like phototransistors cannot be manufactured with large sensitive areas - thus limiting their sensitivity. They also suffer from the fact the charge carriers are not generated at exactly the same spot - thus limiting their speed response.

Photofets are primarily used in switching circuits.

5.11 Avalanche Photodiodes[11,28,29]

As the shot noise in an ordinary photodiode is very small, Johnson noise in the load resistor can be the limiting factor for its sensitivity at low intensities. To overcome the problem of the load resistor, photodiodes have been produced where amplification of the charge carriers can be obtained by the Avalanche effect. A high electric field can transfer enough energy to the generated charge carriers to move electrons from the valence band into the conduction band. The increased number of electrons can repeat the process and since the process is cumulative, amplification can be obtained. Gains of the order of 10^4 have been reported.[28] Improved signal to noise ratio can be obtained by operating at liquid air temperatures.

This photodiode represents a step towards the replacement of the photomultiplier by solid state devices.

5.12 The PIN Photodiode[12]

Devices using the internal photoelectric effect in semiconductors such as photoconductive detectors, pn junction detectors or phototransistors are too slow for some applications, either because of the physical effect employed or because the structures are not suitable for radio-frequency operation.

The PIN photodiode is a detector which can operate at speeds in excess of those obtained from a photomultiplier. It has also a spectral response which makes it useful for work in the visible and the near infra red regions.

The PIN photodiode consists of a wafer of nearly intrinsic semiconductor (I) bounded on one face by a fairly thin layer of very heavily doped n type semiconductor (N), and on the other face by a fairly thick layer of very heavily doped p type semiconductor (P).

When a reverse bias is applied across the PIN photodiode, the depletion layer of the junction opens out unitl it occupies the entire volume of the intrinsic material.

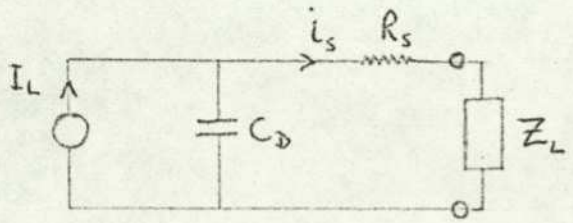
A high and uniform field is thereby established in the intrinsic material and any charge carriers generated in this region by absorption of photons from the incident radiation, will be quickly swept either into the p type or the n type material. The intrinsic material has a very high resistivity and acts as the dielectric of a capacitor.

The time required by the carriers to be collected at the heavily doped regions is given by,

$$t_o = \frac{W}{U} \qquad (5.13)$$

where W is the thickness of the intrinsic layer and U is the limiting velocity of the carriers.

The equivalent circuit of the PIN device is as shown below:



Useful output from the device will be obtained if the impedance of C_D is greater than the load impedance Z_L, ie if

$$\frac{1}{WC_D} > Z_L \qquad (5.14)$$

The transition capacitance of the junction is given by,

$$C_D = \frac{A\varepsilon}{W} \qquad (5.15)$$

Where A is the junction area and ε the dielectric constant of the semiconductor.

Using the last three equations, the upper frequency limit of photodetector response is,

$$W < \left\{ \frac{2\pi U}{Z_L A \varepsilon} \right\}^{\frac{1}{2}} \qquad (5.16)$$

For example the frequency limit for a germanium device[6], for a device 6 mm in diameter, and $Z_L = 10\Omega$ works out to 1.84×10^{10} Hz.

The noise performance of the PIN photodiode may be assessed as follows:

Let the mean signal current generated by the incident radiation be i_s.
Let the load resistance be R_L, so that the output signal power is $i_s^2 R_L$.

The shot noise fluctuations in i_s and in the reverse junction saturation current i_r over a bandwidth δF, represent a noise power of $2q\,(i_s + i_r)\,R_L \delta F$.

Therefore the signal-to-noise power ratio is

$$\left(\frac{S}{N}\right)_{diode} = \frac{i_s^2}{2q(i_s + i_r)\,\delta F} \qquad (5.17)$$

Suppose that the signal from the photodiode is amplified by an amplifier of bandwidth δF and noise temperature T. Then the new signal-to-noise ratio is,

$$\left(\frac{S}{N}\right)_{diode + amplifier} = \frac{i_s^2 R_L}{\left[2q\,(i_s + i_r)\,R_L + KT\right]\delta F} \qquad (5.18)$$

Except for very large currents (a few milliamps) the addition of such an amplifier reduces the signal-to-noise ratio quite considerably, ie. the diode-amplifier combination do not constitute an efficient low noise photodetector.[25]

One way of overcoming this difficulty is to operate the PIN photodiode as a heterodyne detector[12]. With this arrangement, an additional fairly strong light signal i_a whose frequency differs from that of the signal by the intermediate frequency is incident on the diode. This could be achieved by the use of a Gallium Arsenide diode, positioned next to the photodiode, and operated by a local oscillator. At the intermediate frequency, the signal-to-noise ratio is,

$$\left(\frac{S}{N}\right)_{i.f.} = \frac{i_a^2 R_L}{2q\,(i_a + i_{dc} + i_r)\,R_L \delta F} \qquad (5.19)$$

where i_{dc} is the direct current component of the combined signals. Now, for a strong signal i_a,

$$i_a^2 \approx i_s\, i_{d.c}$$

and also $i_{d.c} \gg i_a,\ i_r$

Under these conditions, equation (5.19) becomes,

$$\left(\frac{S}{N}\right)_{i.f} = \frac{i_s}{2q\delta F} \qquad (5.20)$$

For a photomultiplier, the signal-to-noise ratio is,

$$\left(\frac{S}{N}\right)_{p.m.} = \frac{i_s^2}{2qi_s\delta F} = \frac{i_s}{2q\delta F}$$

i.e. the PIN photodiode when used as a heterodyne detector, then its signal-to-noise ratio is similar to that of a photomultiplier.

5.13 Selenium Barrier Layer Photocells

These are photovoltaic cells which are made by pouring molten selenium on to a metal base plate. The selenium is then converted to a crystalline form by annealing in an oven. During the annealing an insulating barrier layer forms on the top surface, on which a transparent thin layer of gold is evaporated. When the top is irradiated, the gold layer becomes negative with respect to the metal plate.

The physical basis of this photocell is as follows: As soon as the gold film is deposited, electrons diffuse from the gold film into the selenium. Thus the gold film becomes positive with respect to the selenium. Incident radiation creates electron-hole pairs in the selenium. The electrons will therefore be collected by the gold film thus making the film negative with respect to the metal base.

Selenium cells show fatigue and history effects. Their frequency response is limited to only about 1 KHz. They are normally used for inexpensive light meters.

6. ON THE RELATIVE COMPARISON AND THE ULTIMATE SENSITIVITY OF PHOTOELECTRONIC DEVICES

6.1 Introduction

The ability of a photoelectronic device to detect or sense or measure the intensity of radiation incident on the device is quite often described by an awkwardly large collection of numbers. The reason for this situation, is that there are no standards laid down relating to the conditions of testing such devices. This situation, is not to be quite unexpected, as different devices have different applications, and therefore testing conditions should vary depending on the application.

Thus one would not expect a system of reference by which, to directly compare a device which is used, say, to detect the presence of a beam of light (such as in burglar alarms), with a device which is used for accurate individual photon counting.

In practice, one finds terms such as sensitivity, responsivity, noise equivalent input (NEI), signal to noise ration (SNR), detectivity, responsive quantum efficiency, detective quantum efficiency, quantum efficiency etc. used to describe particular devices. Needless to say that manufacturers more often than not, use the term that makes their own device look superior to others. Confusion faces the user who is not very familiar with those terms and would like to choose a device for a particular application.

6.2 Terminology

(a) Responsivity: This is the ratio of the detector output to the detector input. Thus, the Responsivity of a detector may be expressed in volts (or amperes) per watt (or lumen). The responsivity of a detector usually depends on the detector sensitive area, its speed of response, the wavelength of the incident radiation, the operating temperature and the modulation (chopping) frequency of the radiation. The dependence of responsivity upon frequency (f) may be of the form[13],

$$R = \frac{R_{max}}{\left[1 + (2\pi f T)\right]^{\frac{1}{2}}} \qquad (6.1)$$

where T is the time constant.

(b) Sensitivity: This is a more ambiguous term than Responsivity, as for a given radiation signal, the detector with the highest sensitivity may be the one with the greater output voltage, or it may be the one with a greater signal-to-noise ratio. Because of its ambiquity Sensitivity is not used very often.

(c) Noise Equivalent Input (NEI): This is the rms fluctuation in the output expressed in terms of the detector input. When expressed in watts, it is called the Noise Equivalent Power (NEP). According to Rose[15], for a high degree of reliability, the minimum detectable input is usually about five times the rms Noise Equivalent Input. This result was arrived at by psychophysical measurements.

(d) Minimum Detectable Input: This is a useful concept, but it does however have the limitation that the reliability of the detector must also be known. This limitation does not, for example, apply to the Noise Equivalent Input.

(e) Quantum Efficiency: For solid state devices, this is the ratio of the number of photons that generate electron-hole pairs, to the total number of incident photons. For an emissive surface, Quantum efficiency is the number of electrons emitted per incident photon.

There are two main classifications of Quantum Efficiency: Responsive and Detective. These are discussed in more detail further on.

(f) Detectivity: This term was introduced by R.C. Jones in 1952[16]. It is defined as the reciprocal of the Noise equivalent input. Detectivity, D, can also be described under a particular reference condition (such A or C as described by Jones); e.g. in the reference condition C, detectivity is given by,

$$D = \frac{A^{\frac{1}{2}}}{P_n} \times \left(\frac{\delta f}{f}\right)^{\frac{1}{2}} \qquad (6.2)$$

where A is the sensitive area in square centimetres, f is the modulation (chopping) frequency, f is the noise equivalent bandwidth and P_n is the noise equivalent power.

(g) Performance: This is a very vague term, which could mean any one of many properties of photoelectronic devices, or their average behaviour. Rose in 1942 used this term to define the 'number of photons required to produce a just detectable signal in a single resolution element'. As this term implies knowledge of the reliability of the detector, Rose abandoned it in favour of 'detective quantum efficiency'.

(h) Gain: This term is normally applied to photoconductivity where it can be defined as the number of charge carriers produced at the output per absorbed photon. It is also used for photomultipliers, where it is normally defined as the number of electrons reaching the anode per electron leaving the cathode. It so happens that on average these two gains are of the same order, viz. 10^6.

(i) Noise: This is the root mean square fluctuation in the output of a detector expressed in terms of the total detector output. Noise may be expressed in rms volts or rms amperes.

(j) Noise Figure: The detectivity of a radiation detector has an upper limit due to the photon noise of the background radiation field in which the detector operates. The noise figure of the detector is defined as the factor by which the upper limit exceeds the actual detectivity.

(k) Signal-to-Noise Ratio: The use of this term to describe the relative performance of detectors involves a logical confusion. This is because this concept is a combined property of a detector and a signal, and not only of the detector.

6.3 Detecting a Radiation Signal in the Presence of Steady Ambient Radiation

Ambient radiation normally consists of,

(a) The blackbody radiation field produced by the detector and its environment, and

(b) Other steady radiation such as daylight or steady manmade illumination.

The radiation signal to be detected is usually varying with time.

When considering the detecting ability of a radiation detector, there are two entirely contrasting cases to be taken into account[13]: (a) The photon noise of the ambient radiation is negligible, and (b) the photon noise of the ambient radiation is the dominant noise.

Case (a) implies that the noise level inherent in the detector itself, is quite high. This category of detector includes photo-conductive cells, thermocouples and bolometers. These detectors are normally operated so that the only incident radiation present is the detector blackbody radiation and the signal radiation to be detected.

R. Clark Jones has considered case (a) very thoroughly,[13] and concludes that under these conditions, 'Detectivity' is the particular characteristic of the device, that should be used to evaluate the relative merits of photoelectronic devices.

In case (b), the statistical fluctuations in the intensity of the ambient radiation (photon noise), cause the prevailing noise in the detector output — the photon noise of the 'steady' ambient radiation, limits the detectivity. This category includes human vision and the photomultiplier.

According to R. Clark Jones[14] the concept of Detective Quantum Efficiency, is the appropriate means to characterize the detecting ability of a detector in this kind of situation.

6.4 Responsive Quantum Efficiency

The Responsive Quantum Efficiency of a detector may be defined in various ways, depending on what one uses as the number of input events. Thus different values will be obtained if the number of absorbed photons is used, instead of the number incident. Due to historic reasons, the standard practice used today, is to use the incident number of photons for photoemissive tubes and the absorbed number of photons for photoconductive cells.

The Respnsive Quantum Efficiency may be greater than unity, e.g. for a photomultiplier it may be 100,000.

The more universal definition of Responsive Quantum Efficiency would be the ratio of effective to incident photons.

6.5 Detective Quantum Efficiency[14]

Due to the difficulties arising in some cases (such as the human or photographic negatives) in ascertaining the number of effective photons or even in defining effectiveness, the concept of Detective Quantum Efficiency was introduced by Rose in 1946.

The Detective Quantum Efficiency of an actual detector is defined as the square of the ratio of the measured detectivity of the detector to the maximum possible detectivity on the given signal in the presence of the given ambient radiation.

Rose defined DQE in terms of an ideal detector, which is one that makes full effective use of every photon incident upon the detector sensitive area.

6.5.1 Detecting ability of an ideal detector

For an ambient radiation consiting of an average number M_a of photons per unit time reaching the detector sensitive area, the rms deviation (assuming a Poisson distribution) from Ma is given by,

$$\delta M_a = M_a^{\frac{1}{2}} \qquad (6.3)$$

Thus the rms noise N, measured in photon numbers is,

$$N = M_a^{\frac{1}{2}} \qquad (6.4)$$

Let a signal, S, whose photon number per unit time is M_S (where M_S M_a), be incident on the detector. The signal-to-noise is then given by,

$$\frac{S}{N} = \frac{M_S}{M_a \cdot t} \qquad (6.5)$$

and the noise equivalent number of signal photons is given by,

$$S_N = M_a^{\frac{1}{2}} \qquad (6.6)$$

This corresponds to the detecting ability of an ideal detector.

Equation (6.5) can be written in the form,

$$\frac{S/N}{M_s/M_a^{1/2}} = 1 \qquad (6.6)$$

and applies only to an ideal detector. For an actual detector the fraction will be less than one, as in this case $\left(\frac{S}{N}\right)_m$ will be smaller than one. The square of the left hand side of equation (6.6) is defined as the detective quantum efficiency, Q_D, ie.

$$Q_D = \frac{(S/N)^2_m}{M_s^2/M_a} \qquad (6.7)$$

Thus, the DQE is the square of the ratio of the measured signal-to-noise ratio $(S/N)_m$, to the signal-to-noise ratio of the ideal detector on the given signal in the presence of the given ambient radiation.

Contrary to RQE, DQE can never be greater than unity.

6.5.2 DQE for the quasi-ideal detector

The quasi-ideal detector is defined as the combination of the ideal detector covered by a filter transmitting a fraction F of the incident photons, ie, it makes maximal use of the fraction F of the incident photons.

For the quasi-ideal detector, the effective number S of signal photons is FM_s. Similarly, the effective mean number of ambient photons is FM_a. Thus the rms fluctuation N in the effective number of ambient photons is $(FM_a)^{\frac{1}{2}}$.

The square of the signal to noise ratio is therefore given by,

$$\left(\frac{S}{N}\right)^2 = \left[\frac{FM_s}{(FM_a)^{\frac{1}{2}}}\right]^2 = \frac{FM_s^2}{M_a} \qquad (6.8)$$

Comparing equations (6.7) and (6.8), it can be seen that,

$$Q_D = F \qquad (6.9)$$

This is the basic justification for calling Q_D, a 'quantum efficiency'.

6.6 Detectivity [16, 17]

The ability of a radiation detector to detect weak signals, under specified conditions, could be defined as the amount of input radiation required to produce a steady electrical output equal to the rms noise voltage. Unfortunately the noise equivalent input suffers from a crippling psychological defect: it is upside down - the better detector having a lower noise equivalent input. Thus a simple concept is needed to describe that property of a detector which is greater for the detector that achieves a greater signal-to-noise ratio for a given radiation signal. R. Clark-Jones [14] introduced Detectivity to describe this concept. It is defined as the reciprocal of the noise equivalent input.

6.6.1 Conditions of testing

In specifying the performance of a detector, one has to realize that the properties of the detector are dependent on the test conditions. Thus a complete description of a detector and the tests conditions employed, might involve the shape and size of the detector and the type of surface. It might also involve the method of activation, method of cooling, window material, window thickness, shape of sensitive area, time constant, electrical resistance, biasing current, temperature, noise voltage, signal voltage etc. One should also describe the radiation source and chopper, ie. chopping waveform, chopping frequency, spectral distribution of the source, incident power per unit area etc. One should also give details of the amplifier, ie. gain versus frequency, input impedance at the chopping frequency, noise equivalent bandwidth, etc.

Thus, a direct comparison of detectors would be impossible unless the above conditions were standardized. It is soon however realized that some of the conditions (such as method of activation, cell resistance, biasing current etc) are not necessary.

According to R. Clark Jones the necessary items required for direct comparison of detectors are:

(a) Sensitive area, A	}	
(b) Noise voltage, E_n	}	Relevant to detector
(c) Signal voltage, E_s	}	
(d) Chopping frequency, F	}	Relevant to source
(e) Incident power per unit area, J	}	
(f) Noise equivalent bandwidth, F	-	Relevant to amplifier

At the same time one ought to specify the temperature of the cell during the test and also the spectral distribution of the source ie. blackbody radiation at a particular temperature, monochromatic source etc.

The next step taken by R Clark Jones was to define detectivity, D, by the relationship,

$$D = \frac{E_s}{JAE_N} \qquad (6.10)$$

Thus, four of the items in the above list, can be replaced by a new single item - property.

Therefore, one could now make direct comparison of two detectors by specifying:

(i) The detectivity D
(ii) Chopping frequency F
(iii) Noise equivalent bandwidth δF

The final step taken by R Clark Jones was to introduce a new term: "The detectivity D_c in the reference condition C".

In introducing this new concept, he implied that the power spectrum of the noise in the output varies inversely as the frequency F (see para 3.5) and that the detectivity of cells of different sensitive area A, varies inversely as the square root of the sensitive area. Therefore D_c may be defined as,

$$D_c = DA^{\frac{1}{2}} \left(\frac{\delta F}{F}\right)^{\frac{1}{2}} \qquad (6.11)$$

This concept has the advantage over any other single test result, that a knowledge of it alone permits a direct and significant comparison of two detectors, even though the cells may have different areas, and even though the tests conditions are different. The only real limitation is the dependance of D_c on the temperature of the photocell and the spectral distribution of the source. Thus these two items have to be specified when denoting the value of D_c.

6.6.2 Applying the concept of detectivity D_c

As explained in section 6.2, the concept of detectivity D_c should be used to compare the performance of radiation detectors only when the photon noise of the ambient radiation is negligible compared with the other noises in the system. Detectivity D_c in the reference condition C should not be used in the case where the photon noise of the ambient radiation is the dominant noise. Under such conditions, the detective quantum efficiency should give a clearer indication of the relative performance of two or more radiation detectors.

6.7 The Ultimate Sensitivity of Radiation Detectors [18]

6.7.1 Summary

R. Clark Jones[19] extended the result obtained for the ultimate sensitivity of a thermal receiver, to apply to any radiation detector whether or not its mechanism is thermal. P.B. Fellgett[18], examined Jones's proposals and concluded that they are not valid for certain types of detector. Fellgett then extended the treatment to cover all cases, and obtained a uniform theory that can be applied to all detectors where ultimate sensitivity is determined by fluctuations in thermal radiation from their surroundings.

6.7.2 The ultimate sensitivity of a thermal detector

The energy fluctuation, E, in a thermal capacity, C, is

$$(\delta E^2)_{av} = KT^2 \frac{\partial E}{\partial T} = KT^2 C \qquad (6.12)$$

The corresponding temperature fluctuation is,

$$(\delta T_n^2)_{av} = \frac{KT^2}{C} \qquad (6.13)$$

When the thermal capacity is shunted by a heat resistance r, in which the bandwidth of the generated fluctuations is $\delta F = \frac{1}{4\ cr}$, then the fluctuation across r may be given in the form,

$$(\delta T_n^2)_{av} = \frac{KT^2}{C} = 4KT^2\ \frac{r}{4cr} = 4KT^2\ r.\delta f \qquad (6.14)$$

When radiant energy, W, is incident on the thermal detector, the temperature rise is $T_s = E_s r W$ where E_s is the fraction of the signal W absorbed. This rise in temperature will equal the rms noise when,

$$W = \frac{1}{E_s}\left(\frac{4KT^2\delta F}{r}\right)^{\frac{1}{2}} \qquad (6.15)$$

6.7.3 Application of the theory to a radiation detector

R Clark Jones assumed that for a radiation detector of area A, effective emissivity ε for thermal radiation, and interacting with its surroundings only by radiation, then,

$$\frac{1}{r} = \frac{d}{dT}\ (\sigma A \varepsilon T^4) = 4\sigma A\ \varepsilon_1 T^3 \qquad (6.16)$$

where $\varepsilon_1 = \varepsilon + \frac{T}{4}\ \frac{d\varepsilon}{dT}$ and σ is Stefan's constant. Then (6.15) becomes,

$$W = 4\left[KT^5\ \delta F\ \sigma A\ \frac{\varepsilon_1}{\varepsilon_s^2}\right]^{1/2} \qquad (6.17)$$

It can be assumed that the fluctuations occur in the radiation which the receiver exchanges with its surroundings. It can therefore be concluded that equation (6.17) gives the ultimate sensitivity of any radiation detector whether its mechanism is thermal or not. The only assumption necessary is that ε_s and ε_1 are interpreted as quantum efficiencies.

6.7.4 Fellgett's generalised theory

Fellgett notes that Clark Jones's proposition (Eqn 6.17) is not applicable to all radiation detectors, because, the response to fluctuations in the temperature radiation does not remain in the same ratio to the response to incident signals whatever the mechanism of the detector. The nature of the detector affects this ratio in two ways:-

(a) In the case of the thermal detector, the total noise power is made up of two equal components: Fluctuations due to the incident photons and fluctuations due to the reradiated photons. In the case of, say, photo-emissive cells, however, there is no noise component due to the reradiated photons, and their ultimate sensitivity would therefore be better by a factor of $\sqrt{2}$. A similar effect, but of variable magnitude is noticeable where a thermal radiator is used at low temperatures so that reradiation is suppressed.

(b) Variations also exist between detectors in which the output is dependent on the number of photons incident (photo-emissive) and those in which the output is dependent on the total absorbed energy (thermocouple). It can be seen from equation 6.17 that the variation with frequency of the fraction of the signal absorbed e_s must be known. This can be deduced from the spectral response, as long as it is known to which class the detector belongs. Also, it can be seen from this equation that the signal to noise ratio does not only depend on e_s/e_i but also on the absolute value of e_s. Thus for a given spectral response, the ultimate sensitivity is obtained by making e_s equal to unity at the peak of response of the detector. The wavelength of this peak varies depending on whether the detector is taken to an energy or a photon number response. This could lead to large changes in the value ascribed to e_s and thus to the ultimate sensitivity.

Fellgett then proceeds to obtain the following equations for the fluctuation in energy E and in the number of photons δM per second, affecting a detector.

$$(\delta E^2)_{av} = \frac{2\pi A}{C^2} \int_o^\infty \frac{\varepsilon(\nu)\, e^{h\nu/KT} h^2 \nu^4}{(e^{h\nu/KT}-1)^2}\, d\mu \qquad (6.18)$$

$$(\delta M^2)_{av} = \frac{2\pi A}{C^2} \int_o^\infty \frac{\varepsilon(\nu)\, e^{h\nu/KT} \nu^2}{(e^{h\nu/KT}-1)^2}\, d\nu \qquad (6.19)$$

The equations obtained for the total number of photons M and the total energy E, affecting the detector per second are:

$$M = \frac{2\pi\varepsilon A}{C^2} \int_0^\infty \frac{\nu^2 \, d\nu}{(e^{h\nu/KT} - 1)} \qquad (6.20)$$

$$E = \frac{2\pi\varepsilon A}{C^2} \int_0^\infty \frac{h^2\nu^4 d\nu}{(e^{h\nu/KT} - 1)} \qquad (6.21)$$

Equations 6.18 to 6.21 can be evaluated by numerical integration. The values obtained are as follows:

$$(\delta E^2)_{av} = \frac{8\pi^5 \varepsilon A (KT)^5}{15c^2h^3} \qquad (6.22)$$

$$(\delta M^2)_{av} = \frac{8\pi^3 \varepsilon A (KT)^3}{6c^2h^3} \qquad (6.23)$$

$$M = \frac{0.48\pi^3 \varepsilon A (KT)^3}{c^2h^3} \qquad (6.24)$$

$$E = \frac{2\pi^5 \varepsilon A (KT)^4}{15c^2h^3} \qquad (6.25)$$

7. THE PRACTICAL PERFORMANCE OF RADIATION DETECTORS

7.1 Introduction

In the first six chapters an exposition has been given of the mechanism of the most important solid state photoelectronic devices, of the causes of their limitations as radiation detectors, suggested methods of direct comparison of such detectors, together with suggested means of working the ultimate sensitivity of these devices.

In this chapter, a summary is given of the work done by several investigators in the relative performance of photoelectronic devices as well as in their ultimate sensitivities. It has been found helpful in some instances to directly compare the ability of solid states devices with that of a photo-multiplier, thermocouple, film negatives and to the most common detector, the human eye.

7.2 Phtoconductors as Radiation Detectors

The ultimate test of how sensitive a photoconductor is to incident radiation, is the signal-to-noise ratio. When it is used as a radiation detector, the significant parameters are (a) the intensity of the background noise currents, and (b) the signal-to-noise ratio for that particular incident radiation.

It is found in practice, that a photoconductor is a very noisy converter of incident radiation to current. According to Butler[20, 22], who investigated noise in CdS photoconductors quite extensively, the signal-to-noise ratio of the photocurrents is one thousand times lower than the signal-to-noise ratio in the incident photon stream.

The main source of noise in photoconductors is the 1/f noise. These are attributed to poor contacts either at the electrode - Semiconductor interface or at internal "surfaces".

The total noise in a photoconductor can be expressed in the same form as shot noise, viz.

$$I_n = \left[2(qG)\, I\delta f\right]^{\frac{1}{2}} \qquad (7.1)$$

where I_n is the total rms noise current in the bandwidth δf due to the average current I, and qG the total charge.

The signal-to-noise ratio is,

$$\frac{I}{I_n} = \left(\frac{I}{2qG\delta F}\right)^{\frac{1}{2}} \qquad (7.2)$$

The bandwidth, δF, is determined by the lifetime T, by

$$T = \frac{1}{2\delta F} \qquad (7.3)$$

$$\text{Therefore, } \frac{I}{I_n} = \left(\frac{IT}{qG}\right)^{\frac{1}{2}} \qquad (7.4)$$

The number of carriers, $F = I/qG$, and using equation (4.1),

$$\frac{I}{I_n} = (FT)^{\frac{1}{2}} = (N)^{\frac{1}{2}} \qquad (7.5)$$

$N^{\frac{1}{2}}$ represents the smallest signal that can be detected.

For CdS photoconductive cells it is found that the smallest detectable signal is of the order 10^{-14} watts per sq cm. Values of the order of 10^{-15} have also been reported. This corresponds to about 10,000 photons per sq cm per sec. at around 550 nm. This shows that a photoconductive detector is a very sensitive device indeed, quite comparable with the human eye sensitivity (125 photons/sec).[23]

The Responsive Quantum Efficiency of all photoconductors approaches 100% very closely.[14]

The Detective Quantum Efficiency of CdS photoconductors varies between 60% to close 100% (depending on surface reflections), for modulation frequencies around 100 cps. The DQE drops to around 10% at 1000 cps, and decreases even more at higher modulation frequencies. Lead Sulfide cells and most other photoconductive cells have got a much lower DQE.

The Detectivity of photoconductive cells seems to vary with a maximum of 5×10^{13} $(cps)^{\frac{1}{2}}$/watt and an average of 1.7×10^{12} $(cps)^{\frac{1}{2}}$/watt.[14]

7.3 P-n Junction Devices

These devices are substantially less sensitive than the homogeneous type of photoconductor. The reason of course is the enormous gain provided by the trapping process in photoconductors.

Both the RQE and the DQE are of course close to 100%, but the average detectivity is only around 10^{10} $(cps)^{\frac{1}{2}}$/watt.

A very great improvement is obtained with the PIN photodetector where a detectivity of around 10^{13},[21] has been obtained. These devices can detect energies down to 10^{-14} watts per sq cm. They can also operate at frequencies exceeding 10 GHz. This makes them very comparable with a photomultiplier, and they are also much more stable.[24, 25, 26]

7.4 Photoemissive Tubes

The highest RQE for a photocathode is only around 18% whilst the average is around 5%.[14]

The average DQE of a gas phototube or a photomultiplier is around 70%.[14]

The average detectivity for a gas phototube is around 10" whilst that for a photomultiplier around 10^{14}.[14]

8. USING PHOTOELECTRONIC DEVICES

8.1 Uses of Photoconductors[30]

The most obvious use of photoconductor is a radiation detector over the spectral region in which it has a high response. It can also be used for x-ray detection and nuclear particles detection. Another application is the use in light sensitive switches where the high power handling ability of the photoconductor can operate a relay directly without the use of an amplifier. In the Infra-Red region the photoconductive type of PbS practically monopolizes.

New uses of photoconductors are continuously found. This is because of their small size, power handling capabilities, ruggedness, and their extremely low cost. Because of their high sensitivity, they have replaced photomultipliers where speed of response is not an important factor.

Photoconductors have also been used in devices such as the Vidicon; photoconducting pickup tube, television camera, x-ray imaging tube. They are also used in light amplifiers.

The main drawback of the photoconductor is its very low speed response. Even at high light intensities the CdS type can only be used at modulation frequencies of a few hundred cycles per second.

Another disadvantage is the "history" or "memory" effect, which makes them unsuitable for light meters, even though some cameras do use them.

8.2 P-n Junction Phtodiodes

Contrary to photoconductors, they do not suffer from any memory effects and they are also much faster in response. They are however much less sensitive and more expensive. With the use of suitable amplifiers they can have extensive applicability ranging from switches to lightmeters. Because of their small size they are used for tape and card readers, GaAs laser detectors etc.

Their speed of response can exceed 1 GHz.

The PIN photodiode has already replaced the photomultiplier in quite a few applications ranging from spectrophotometers to star-tracking.[26] This must be the best overall photoelectronic device at the moment. As a fairly new device, its price is rather high (ranging from £5 upwards).

8.3.1 A three-colour indicator

This instrument was designed to indicate the three primary colour values (Red, Yellow and Blue) of oil samples. The idea was to build an automatic Lovibond. One of the main difficulties experienced with this measurement was the dirt on the sides of the cell containing the oil, dirt in the oil itself, convection currents in hot samples and air bubbles with on-line measurements. For these reasons a beam of light in the near I.R. region was used as a Reference Beam. This beam was affected by dirt etc. but not by the colour of the sample under test.

Thus a disc with four filters is used on the present version: An I.R. filter used as a reference, a red filter to measure the blue component, a green filter to measure the red (magenta) component and a blue filter to measure the yellow component. A PIN photodiode is used to measure the transmission for each filter, as the filters periodically cut the optical beam which crosses the sample cell. A system of fibre optics and four phototransistors sychronize the measuring beams with the gates feeding the four amplifiers for the four individual measurements.

The PIN photodiode is used in the logarithmic mode. This allows the colour measurement beams to be sequentially subtracted from the reference I.R. beam, thus avoiding the introdction of an expensive devider. The three referred values thus obtained are suitably amplified and then displayed on three meters. The instrument is at the moment calibrated in Lovibond units.

The PIN photodiode was preferred for four reasons:

(a) Its spectral response - covering the whole of the visible region plus the essential part of the near I.R. region necessary for the referencing.

(b) Its large area - enabling a beam 3/8ths of an inch in diameter to be used, thus covering a sufficient volume of sample to be representative of the total.

(c) Its high speed - other considerations necessitated measurements to be made with a motor at 300 rpm. This corresponds to 1200 interruptions per minute. But the actual measuring period was only a fraction of a millisecond.

(d) Its logarithmic response - This mode of operation is very temperature sensitive, and thus the photodiode had to be housed in a constant temperature housing.

It should be noticed that the phototransistors are only used as switching devices.

8.3.2 Pressurised can leak detector

In the manufacture of products contained in cans at a high pressure, a check must be made for leaks. The pressurised cans are passed through a tank of hot water, so that a stream of bubbles is formed in the water as a leaky can enters the tank. The cans enter the tank attached to magnetic conveyor belts in rows of eight.

A beam of light is incident crosswise from one end of the tank to the other where a photoconductive cell of the sintered type detects it. As the beam is intercepted by the bubbles, a pulse from the photocell is fed into a counting circuit. An alarm is given when sufficient number of bubbles in a predetermined time cross the beam, (normally 3 bubbles per 300 milliseconds).

The photoconductive cell was chosen because of its low price, ruggedness and ease of mounting. The price was important as several of these detectors will probably be made.

8.3.3 Detection of cloud point

In the production of some liquid detergents, it is essential to know the point at which they start to cloud as the temperature changes. This

cloudiness is caused by the formation of minute particles in the liquid. The effect on a beam of light of these particles is twofold: Absorption and scattering.

A photoconductive cell is used to detect the intensity of the transmitted light and another one for intensity of the scattered light at right angles. As the cloudiness increases the resistance of the first photocell increases whilst that of the second decreases. The two photocells are used as the feedback and the input resistor of an operational amplifier. Thus, the output of the amplifier is proportional to the square of the cloudiness effect. The advantage of the two cell system, apart from getting an increased output is that by proper choice of photocells, drifts such as due to temperature changes or history effects, are eliminated.

8.3.4 Measurement of the thickness of a layer of wax on paper

To prevent moisture coming in contact with some hydroscopic products, the products are wrapped in paper which is covered with special wax. As this wax is rather expensive, its thickness has to be accurately monitored. This has been achieved as follows:

Ultra violet radiation which is strongly absorbed by this wax, is incident on the wax-covered paper, and the reflected radiation is detected by a Selemium Barrier type photovoltonic cell, which is situated next to the flourescent tube producing the radiation but optically isolated from it. As absorption of the incident radiation may be caused due to other effects such as dirt on the surface, a reference signal in the green region is used. The white paper which forms the base for the wax reflects uniformly through the visible and the near ultra violet regions. The reference signal is detected by another barrier selenide photocell on the other side of the flourescent tube.

The voltage outputs of the photocells, which are operated in the logarithmic mode are fed into a differential amplifier. The output of the amplifier is proportional to the thickness of the wax. This is because the differential amplifier acts as a divider for the two logarithmic signals.

The choice of the photocells was mainly governed by the large area necessary to cover. Also, because of the response in both the ultra violet region and the visible.

8.3.5 Dust monitor

The concentration of dust in a factory atmosphere has to be controlled within certain limits. This was achieved by the use of "narrow angle scattering" from a beam of a Helium-Neon laser.

As particles cross the beam, normally at relatively high speeds, they produce a signal at a p-n junction silicon photocell, situated slightly off the main laser beam. The photocell is A.C. coupled to the preamplifier. Thus a signal is produced at the output of the amplifier only when a particle crosses the beam. The amplifier output is integrated and the d.c. level is used as an indication of the particle concentration in the atmosphere.

The silicon p-n junction photocell was chosen because of its high speed, small area (to minimise the error introduced by a number of particles generating pulses on a large area photocell at the same time), high sensitivity, stability and price.

8.3.6 Detection of trace glycerine in water

An instrument was designed to detect minute traces of glycerine in water (one part per million). This was made possible, by using the Brillouin scattering from the glycerine molecules. It is well known, that the amplitude of the side bands of the Rayleigh scattering, is proportional to the concentration of the scattering centres, whilst their distance from the main band is characteristic of the volume of the scattering centres.

The main beam was provided by a Helium-Neon laser. The scanning was provided by a Fabry-Perrot Etalon, placed in a vacuum chamber, such that the optical path could be changed by a change in the air pressure. The detection was originally done by a photomultiplier. It was however soon realized that a PIN photodiode combined with a highly sophisticated amplifier, could replace the photomultiplier. The PIN photodiode reduced the complexity, the size, the cost and the adoptobility of the instrument.

REFERENCES

1. Ambroziolk A., Semiconductor Photoelectric Devices, Iliffe (1968).
2. Gorlich, P, Advances in Electronics, Vol. 4, (1952)
3. Bennett, Problems of Noise, Iliffe, (1965)
4. Cleraix Manual, Photoconductive Cell Applications (1971)
5. Smith, Jones and Chasmar, I.R. Radiation, Clarenden, 1968
6. R.H. Bube, J. Phys Chem. Solids, Vol. 1, (1957) p. 234
7. R.H. Bube, Proc. I.R.E., (Dec 1955) p. 1836
8. Cumerow R., Phys. Rev. Vol. 95 (1954) p. 16
9. Witherell and Faulhaler, Appl. Opt. Vol. 9 (1970) p. 73
10. Prince & Welf J. Brit. I.R.E. (1958) p. 583
11. Haitz, Goetzberger, Scarlet, Schockley, Z. Appl. Phy. 34 (1963) p. 1581
12. Riesz, Rev. of Sc. Instr., Vol. 33, (1962) p. 994
13. R.C. Jones, Adv. in Electronics, Vol. 5 (1953) p. 1-96
14. R.C. Jones, Adv. in Electr. Vol. 11, (1959) p. 87-183
15. A. Rose, J. Opt. Soc. Am Vol. 38 (1948) 196-208
16. R.C. Jones, Nature Vol. 170, p. 937-8 (1952)
17. R.C. Jones, The Rev. of Sc. Instr. Vol. 24 (1953) p. 1035
18. P.B. Fellgett, J. Opt. Soc. of Am, Vol. 39 (1947) p. 970
19. R.C. Jones, J. Opt. Soc. of Am, Vol. 37 (1947) p. 879
20. Rose, A. Proc. I.R.E. (1955), p. 1850-1866
21. U.D.T. Silicon Photodetector Design Manual (1972)
22. A. Van der Ziel, Proc. I.E.E.E. Vol. 58, p. 1178, (1970)
23. Imaging Devices, Biberman & Nudelman, Plenum, (1971)
24. Hewlett-Pachard Manual, HP5082-4200 (1970)
25. R.H. Hamstra P. Wendland, Appl. Optics. Vol. 11, No. 7 (1972)
26. R. Fisher, Appl. Optics, Vol. 7, No. 6, p. 1079 (1968)
27. R.D. Baertch, I.E.E.E. Trans., p. 987 (1966)
28. R.D. Bartch, I.E.E.E. Trans., p. 384 (1966)
29. R.Z. McIntyre, I.E.E.E. Trans., Vol. Ed-13, No. 1. p. 164 (1966)
30. Mullard Manual, Photoconductive Cells, (1969)
31. J.B. Dance, Photoelectronic Devices, Iliffe, (1969)

CONCLUSION

It is hoped that enough material has been presented in this report to guide a potential user to the best selection of a solid state photoelectric device.