CHAPTER 2

This chapter reviews and evaluates the existing techniques for the design of colour scanning filters. The problem of the design of colour scanning filters is formulated in a far broader sense than it is in the publications reviewed. The advantages of such a formulation are two-fold. Firstly, this formulation allows the incorporation of the physical constraints of fabricating a filter. Secondly, the formulation allows the application of the method to filter design for various applications, including colorimetry, colour correction and satellite imaging, unlike narrower formulations. This chapter is organized as follows. The first sections provide the background for the problem. Section 1 deals with the preliminaries of the representation of continuous colour signals in the discrete domain. The method of data correction which is commonly used to manipulate colour scanner measurements is presented in section 2. Section 3 presents a model of the additive colour display. The subtractive colour process is reviewed in section 4.

The last sections review previous work related to the design problem. Orthogonal matching functions and 'most orthogonal matching functions' as defined by MacAdam [20] are reviewed in section 5. Section 6 deals with means of constructing designed filters from an existing set of filters. The different techniques of constructing and designing colour scanning filters are reviewed in section 7, and applications of the vector space approach to colour filter design are discussed. Section 8 reviews methods of colour correction that advocate the use of more than three colour scanning filters.

Neugebauer's q-factor [22] is presented as a measure of goodness of a single colour scanning filter in section 9. Section 10 presents a summary.

2.1 Preliminaries

Most current research in colour systems assumes that the visual frequency spectrum can be represented by samples taken over the range 400-700 nm. The usual sampling interval is 10 nm although there are some cases where finer sampling is required. Integrals are approximated by summations, and the infinite-dimensional Hilbert space of visible spectra with the usual 2-norm is reduced to an N-dimensional Hilbert space, where N is the number of samples (for 10 nm sampling, N=31). A continuous function of wavelength is represented by an N-vector of its sampled values. Hence, visual spectra will be treated as vectors in an N-dimensional Hilbert space in this dissertation. The CIE colour matching functions defined in section 1.1.1 are represented by N-vectors, $\mathbf{a}_i$.

The notation in this dissertation follows that of Trussell [31] and Vora and Trussell [35]. Let $\mathbf{S} = [\mathbf{s_1}\ \mathbf{s_2}\ \mathbf{s_3}]$, where $\mathbf{s_1}, \mathbf{s_2}$ and $\mathbf{s_3}$ are N-vectors that represent the colour sensitivity functions of the three types of cones in the eye. Let $\mathbf{A} = [\mathbf{a_1}\ \mathbf{a_2}\ \mathbf{a_3}]$ and $\mathbf{P} = [\mathbf{p_1}\ \mathbf{p_2}\ \mathbf{p_3}]$,where $\mathbf{a}_i$ and $\mathbf{p}_i$ (i = 1, 2, 3) are N-vectors representing the CIE matching functions and the corresponding CIE primaries, respectively. The matrix $\mathbf{P}$ is defined by the equation $\mathbf{A}^T\mathbf{P} = \mathbf{I}$ [31], and is not unique. The set $\{\mathbf{a}_i\}_{i=1}^3$ denotes the set of matching functions; $\{\mathbf{m}_i\}_{i=1}^r$ denotes any set of r scanning filters, and $\mathbf{M}$ denotes the matrix of scanning filters, $\mathbf{M} = [\mathbf{m}_1\ \mathbf{m}_2 \ldots\ldots \mathbf{m}_r]$. The range space of a matrix $\mathbf{X}$ is the span (set of linear combinations) of its column vectors and is denoted by $R(\mathbf{X})$. Hence $R(\mathbf{M})$ denotes the set of linear combinations of the scanning filters.

An arbitrary scanning filter is represented by the vector $\mathbf{m}$.

The CIE matching functions are distributions of radiant power that characterize the human visual system. The colour of an object as seen by the human eye depends on the radiant power emitted by it. This, in turn, is a function of the radiant power incident on the object and the reflectance spectrum of the object. When the incident radiation is uniform as a function of wavelength, the radiant power incident on the eye characterizes the the reflectance spectrum of the object without any distortion. For a given reflectance spectrum $\mathbf{f}$ viewed under a uniform radiance source, the colour matching of section 1.1.1 is represented by

$$\mathbf{S}^T\mathbf{f} = \mathbf{S}^T\mathbf{P}\mathbf{t}$$

where $\mathbf{t} = \mathbf{A}^T\mathbf{f}$ [31] represent the CIE tristimulus values of $\mathbf{f}$, and are a valid approximation to the tristimulus values defined in equation (1.3). The spectra $\mathbf{f}$ and $\mathbf{Pt}$ provide identical colour stimuli and are known as metamers. Thus, the colour stimulus of $\mathbf{f}$ may be reproduced exactly if $\mathbf{t}$, the vector of tristimulus values, is known and the primaries are realizable. It can be shown that $\mathbf{t}$ is uniquely determined by the projection of $\mathbf{f}$ onto the subspace spanned by the set of linearly independent vectors, $\{\mathbf{a}_i\}_{i=1}^3$ [28] which is also the subspace spanned by the set of linearly independent vectors, $\{\mathbf{s}_i\}_{i=1}^3$. The spectrum $\mathbf{g}$ is a metamer of $\mathbf{f}$ under spectrally flat illumination iff $\mathbf{A}^T\mathbf{f} = \mathbf{A}^T\mathbf{g}$, i.e. the projection of $\mathbf{g}$ onto the space spanned by $\{\mathbf{a}_i\}_{i=1}^3$ is identical to the projection of $\mathbf{f}$.

A change of primaries from $\mathbf{P}$ to $\mathbf{Q} = [\mathbf{q_1}\ \mathbf{q_2}\ \mathbf{q_3}]$, such that $\mathbf{A}^T\mathbf{Q}$ is non-singular, results in a change of matching functions from $\mathbf{A}$ to $\mathbf{B} = [\mathbf{b_1}\ \mathbf{b_2}\ \mathbf{b_3}]$, such that

$$\mathbf{B} = \mathbf{AT} \tag{2.1}$$

where $\mathbf{T}$ is an invertible linear transformation dependent on the primaries $\mathbf{P}$ and $\mathbf{Q}$ [31]. This implies that the space spanned by a set of matching functions is independent of the primaries used. This space is called the Human Visual Subspace (HVSS). Its orthogonal complement is called the black space and is the null space of the linear transformation represented by $\mathbf{A}^T$.

If the incident illumination is not uniform as a function of wavelength, the radiant power incident on the human eye depends not only on the characteristic reflectance spectrum of the object, but also on the illuminant spectrum. This implies that the characteristic spectrum of the object is distorted by the illuminant spectrum. Suppose the spectrum of the illuminant is represented by $\mathbf{l}$. The reflectance spectrum $\mathbf{f}$ is seen as $\mathbf{Lf}$, where $\mathbf{L}$ is a diagonal matrix, such that $\mathbf{L}_{ii} = \mathbf{l}(i)$. For a perfect colour match between the signals $\mathbf{f}$ and $\mathbf{g}$, as viewed under illuminant $\mathbf{l}$, $\mathbf{A}^T\mathbf{Lf} = \mathbf{A}^T\mathbf{Lg}$, or $(\mathbf{LA})^T\mathbf{f} = (\mathbf{LA})^T\mathbf{g}$. Let the matrix $\mathbf{A}_L$ denote the matrix product $\mathbf{LA}$. Then, $\mathbf{A}_L^T\mathbf{f} = \mathbf{A}_L^T\mathbf{g}$ and the spectrum $\mathbf{g}$ is known as a metamer of $\mathbf{f}$ under illuminant $\mathbf{l}$. The visual stimulus of a signal is now determined uniquely by its projection onto the subspace spanned by the set of vectors, $\{\mathbf{La}_i\}_{i=1}^3$. This subspace is defined as the Human Visual Illuminant Subspace(HVISS) for the illuminant $\mathbf{l}$ and can be denoted $R(\mathbf{LA})$. The projection of $\mathbf{f}$ onto the HVISS for $\mathbf{l}$ is

$$P_V\mathbf{f} = \mathbf{A}_L(\mathbf{A}_L^T\mathbf{A}_L)^{-1}\mathbf{A}_L^T\mathbf{f} \tag{2.2}$$

and is also called the fundamental of $\mathbf{f}$ with respect to the illuminant $\mathbf{l}$ [3]. It can be shown that the fundamental with respect to a particular illuminant represents completely the visual stimulus of the signal with respect to that illuminant [28, 31] and:

$$P_V\mathbf{f} = P_V\mathbf{g} <=> \mathbf{A}_L^T\mathbf{f} = \mathbf{A}_L^T\mathbf{g}. \tag{2.3}$$

Colour reproduction begins with correctly determining the fundamental or the projection of a given spectrum onto the HVISS. A perfect set of scanning filters is one whose measurements are within a linear transformation of the tristimulus values of the signal. Equivalently, a perfect set of scanning filters determines the fundamental of the signal and equivalently also, a perfect set of scanning filters is one whose span includes the HVISS. The set $\{\mathbf{La}_i\}_{i=1}^{3}$ is a perfect set of scanning filters and a basis for the HVISS, but it is not the only one.

2.2 Data Correction

The fabrication process limits the kinds of optical filters that can be manufactured. This implies that manufactured filters will most often not be perfect, even if designed filters are, and scanning filter measurements are often 'corrected' to give estimates of tristimulus values. Given a set of colour measurements for an ensemble with known tristimulus values, it is common to derive the 3x3 matrix that premultiplies the measurements to give a minimum mean error square estimate of the tristimulus values. This correction is dependent on the particular ensemble. It is commonly performed in colorimetry when the scanning filters are to be used on a well-characterized data set. The fundamental may be estimated from the corrected data, where the corrected data is the linear minimum mean square error approximation of the actual tristimulus values. Such a fundamental will represent a signal which has the corrected data as its tristimulus values. As demonstrated in [35], the correction reduces both mean square and $L^*a^*b^*$ errors considerably.

The correction of these measurements for scanning filter errors is done by a linear transformation $\mathbf{B}$ such that $\mathbf{h} = \mathbf{B}\mathbf{M}^T\mathbf{f}$, where $\mathbf{M}^T\mathbf{f}$ is the set of scanning filter

measurements and $\mathbf{h}$ is the set of corrected measurements. The matrix $\mathbf{B}$ is chosen such that $E[||\mathbf{A}_L^T\mathbf{f} - \mathbf{B}\mathbf{M}^T\mathbf{f}||^2]$ is a minimum over a given ensemble. The corrected scanning filter data is:

$$\mathbf{h} = (\mathbf{A}_L^T\mathbf{R}\mathbf{M}(\mathbf{M}^T\mathbf{R}\mathbf{M})^{-1})(\mathbf{M}^T\mathbf{f}) \tag{2.4}$$

where $\mathbf{R} = E[\mathbf{f}\mathbf{f}^T]$ is the sample correlation matrix of the ensemble. The estimated fundamental is:

$$\hat{P_V}\mathbf{f} = (\mathbf{A}_L(\mathbf{A}_L^T\mathbf{A}_L)^{-1})\mathbf{h}$$

which gives:

$$\hat{P_V}\mathbf{f} = (\mathbf{A}_L(\mathbf{A}_L^T\mathbf{A}_L)^{-1}\mathbf{A}_L^T\mathbf{R}\mathbf{M}(\mathbf{M}^T\mathbf{R}\mathbf{M})^{-1})(\mathbf{M}^T\mathbf{f}) \tag{2.5}$$

As the transformation from fundamental to tristimulus values is linear, the same expression is obtained by minimising the error between fundamentals.

The corrected data set always provides a lower mean square error than the uncorrected data because the corrected fundamental is *the* minimum mean-square error linear estimate of the true fundamental. Hence, the mean square error between the corrected fundamental estimate and the true fundamental will be lower than that between the true fundamental and any other linear estimate, including the uncorrected fundamental. It should be noted that data correction is a way of making the best use of the data obtained from a set of scanning filters if the correlation of the data set is known. Data correction does not change the data obtained in any fundamental manner, and the colour estimate can only be as good as the data obtained with the set of scanning filters.

2.3 Additive Colour Displays

Colour CRTs (Cathode-Ray Tubes) are a common example of additive colour displays. In a colour CRT, the image on the display is formed by the addition of the radiant power from (usually three) phosphors. These phosphors form the primaries for the display. A displayed colour signal may be expressed as the additive linear combination of the radiant power contributions of each phosphor. Negative coefficients in the linear combination are not physically possible because the coefficients represent the radiant power emanating from the phosphor.

A displayed signal on a colour monitor may be expressed as:

$$\mathbf{g} = \mathbf{P}\mathbf{p}$$

where $\mathbf{g}$ is an N-vector representing the displayed signal, $\mathbf{P}$ is an $N \times s$ matrix whose columns represent the responses of the s phosphors, and $\mathbf{p}$ is an s-vector whose elements represent the strength of the signal from the corresponding phosphor. If $\mathbf{g}$ is to be a metamer of $\mathbf{f}$, then

$$\mathbf{A}_L^T\mathbf{g} = \mathbf{A}_L^T\mathbf{f}$$

and

$$\mathbf{A}_L^T\mathbf{P}\mathbf{p} = \mathbf{A}_L^T\mathbf{f}$$

To reproduce the visual colour stimulus of $\mathbf{f}$ it is sufficient to determine $\mathbf{p}$, which is [33]

$$\mathbf{p} = (\mathbf{A}_L^T\mathbf{P})^{-1}\mathbf{A}_L^T\mathbf{f} \tag{2.6}$$

if $(\mathbf{A}_L^T\mathbf{P})^{-1}$ exists, or the primaries are visually independent, i.e. the visual stimulus of one cannot be matched by a linear combination of the other two. Notice that the values $\mathbf{p}$ are a linear combination of the CIE tristimulus values of $\mathbf{f}$. This is an

example of a case where it is not the CIE tristimulus values that are to be determined, but a linear combination of the CIE tristimulus values is to be determined for colour reproduction.

2.4 Subtractive Processes

A colour photograph is an example of a colour reproduction produced by the subtractive principle. At each point in the photograph, three dyes, cyan, magenta and yellow, are present in varying concentrations. The cyan dye absorbs radiation mostly in the red region of the visible spectrum. It passes the green and blue radiation, which gives cyan its definitive colour. Hence, if white light is incident on a white paper coated with cyan dye, the red part of the visual spectrum is absorbed, and the blue and green reflected. The cyan dye, in effect, serves to 'subtract' the red from the incident radiation. Similarly, the magenta dye serves to subtract the green part of the spectrum, and the yellow dye subtracts the blue part of the spectrum.

The amount of light absorbed and transmitted at each wavelength is a function of the density of the corresponding dyes. An increase in density implies a decrease in transmissivity over all wavelengths, and

$$T(\lambda) \;=\; r(\lambda)^{-d}$$

where T is the transmissivity, d the dye density, and r the transmissivity at wavelength λ of a dye with unit density. The dye transmissivity curve is not a 'block' curve, i.e. the curve is not uniformly high over a range of wavelengths and then uniformly low. In fact, there is a continuous change from high to low. The subtractive principle of hard-copy colour may be contrasted with the additive principle discussed in section 2.3. In contrast with the radiant power of the individual phosphors, which goes towards

increasing the contribution of the respective phosphor to the color reproduction, the dye density of a particular colour goes towards decreasing the contribution of the dye to the colour reproduction.

The fact that a colour photograph is made from the combined effect of only three dyes whose transmissivities may be determined as a function of wavelength, implies that the colour properties of each point in the photograph are completely determined by three parameters, the three dye densities. A densitometer is the instrument used to determine the optical densities. For colour measurement, it uses three narrow-band filters. The center of the band of a filter is determined by the wavelength of the maximum density of a corresponding dye. The readings from the three filters are used to obtain the three dye densities. The work of Spooner [29] indicates that the measured radiant power is a function of the distance of the aperture of the measuring instrument from the sample to be measured. It also indicates that the measurements obtained from a certain area are dependent on the spectra of the surrounding points. Hence, accurate measurements can be taken only if the radiation from the area surrounding the measurement aperture has an identical (or very similar) spectral composition. Calibration of measuring instruments like scanners and densitometers should take this into account.

2.5 Orthogonal Matching Functions

MacAdam [20] defines orthogonal matching functions. His motivation for defining them is to simplify representation of the HVSS and also to simplify error calculation in chromaticity space. A negative value for a filter transmissivity is not a physically realizable value. As orthogonal curves that are linear combinations of the CIE matching

functions have negative components, they are not physically realizable, and for quite some time the only role they played was in representation of the HVSS. MacAdam also defines 'most orthogonal' non-negative linear combinations of the CIE matching functions.

Pearson and Yule [24] mention that 'most orthogonal' curves result in higher colour saturation in photography and higher accuracy in colorimeters, though they do not justify this claim on mathematical grounds. They also do not attempt to define what they mean by 'most orthogonal'. As realizable scanning filters are non-negative, 'more orthogonal' filters imply 'more non-overlapping' filters. As realizable primaries are also non-negative, the identity, $\mathbf{M}^T\mathbf{Q} = \mathbf{I}$ [31] implies that realizable primaries can have non-zero values only where the other scanning filters have zero values. Hence, 'more orthogonal' (and hence 'more non-overlapping' or 'narrower') scanning filters are 'more likely' to correspond to realizable primaries. This is one more reason that orthogonality is seen as a criterion for optimality. The recent use of coloured fluorescent lights in lieu of scanning filters and primaries may relax the conditions on realizable primaries.

Pearson and Yule [23] use orthogonality as an optimality criterion in designing scanning filters while converting a densitometer into a colorimeter. They claim [24] that MacAdam determined linear combinations of the matching functions that had no negative components and were as 'non-overlapping' as possible. There does not seem to be a major flaw in this claim. The equivalence of 'non-overlapping' with orthogonal is justified because the required functions are non-negative. The method used works primarily because of the nature of the matching functions which allows non-overlapping to be made equivalent with 'narrowest'.

The most important point in an investigation such as the one performed by Pear-

son and Yule should have been a mathematical definition of 'most non-overlapping', in the form of a single function to be optimised. The researchers do not define any such function, but try to reduce the pair-wise inner products, which give three separate functions. As they do not define a function to be optimised, the only way they can justify MacAdam's choice is by trial and error.

A measure of the orthogonality of two vectors $\mathbf{u}$ and $\mathbf{v}$, is defined by Pearson and Yule as [24]:

$$\frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}.$$

A definition more consistent with other mathematical and statistical applications is:

$$\frac{\mathbf{u}^T\mathbf{v}}{||\mathbf{u}||||\mathbf{v}||}.$$

In their attempt to define families of curves from which the 'most orthogonal' are chosen, the decision to normalise the curves so that the 1-norm is unity is very impractical. Reasons favouring the 1-norm are the fact that the vectors are non-negative, and that the 1-norm represents the energy of the light-signal. This is justification from the physical point of view, but makes mathematical problem formulation far more difficult. The 2-norm is far easier to manipulate mathematically than any other norm, and MacAdam himself, many years before [23] was published, used filters normalised so that the 2-norm is unity. This would seem a better choice when inner products are being optimised, because the 2-norm is consistent with inner products. Further, instead of defining the families of curves to be linear combinations of the CIE colour matching functions, it would have been better to define them as linear combinations of orthogonal curves that span the HVSS. Such curves can easily be obtained by the Gram-Schmidt process, and have been obtained by MacAdam himself [20]. The trial

and error approach limits the accuracy of the results, and no effort is made to limit the number of further trials at each step, as is normally the case with sophisticated trial and error techniques.

2.6 Construction of Designed Filters

An important criterion in the design of scanning filters is their realizability as optical filters. For this reason, all scanning filters belong to the set:

$$C_n = \{\mathbf{m} | 0 \leq \mathbf{m}(i) \leq 1\} \tag{2.7}$$

Transmission space is defined as the space of filters defined by their spectral transmissivities, hence C_n is a subset of transmission space.

One method of realising a designed scanning filter is by combining already existing optical filters. This is a popular approach, and is a relatively inexpensive proposition as far as fabrication is concerned. Suppose a scanning filter, $\mathbf{m}$, is to be obtained from a set of n existing filters. It is sufficient to construct a filter that is a scalar multiple of $\mathbf{m}$. Let $\{\mathbf{r}_i\}_{i=1}^{n}$ denote the set of the n given filters where $\mathbf{r}_i$ (i=1, 2, ...n) is a vector in N-dimensional transmission space. If it is possible to express $\mathbf{m}$ as a product of these filters, i.e.

$$\mathbf{m} = \alpha \prod_{i=1}^{n} \mathbf{r}_i^{u_i}, \ \textit{for some} \ \ \alpha, \ u_i \geq 0, \tag{2.8}$$

then $\mathbf{m}$ belongs to the set C_T in transmission space [31]:

$$C_T = \{\mathbf{f} | \mathbf{f} = \alpha \prod_{i=1}^{n} \mathbf{r}_i^{u_i}, \ \alpha, \ u_i \geq 0\} \tag{2.9}$$

where u_i is a scalar representing the density of the optical filter $\mathbf{r}_i$ in the construction of $\mathbf{m}$ (see equation(1.8)), and α is a scalar multiplying factor. As a realizable filter also belongs to C_n, the vector

$$log\ \mathbf{m} = \begin{bmatrix} log\ \mathbf{m}_1 \\ log\ \mathbf{m}_1 \\ \vdots \\ log\ \mathbf{m}_N \end{bmatrix}$$

exists. The set of all filters defined through the logarithms of their spectral transmissivities is referred to as density space. The vector $log\ \mathbf{m}$ belongs to the cone:

$$C_d = \{\mathbf{m_d} | \mathbf{m_d} = \sum_{i=1}^{N} u_i log \mathbf{r}_i - a; u_i \geq 0; a\ is\ real\}, \tag{2.10}$$

where a represents the multiplying factor α. A realization of $\mathbf{m}$ as a product of filters, as in equation (2.8), is known as a cascaded realization of $\mathbf{m}$, or simply as a product realization. This is the most common realization, though attempts have also been made to express $\mathbf{m}$ as a sum of filters, or a product of sums, or even a sum of products.

2.7 Exact vs. Approximate

A number of researchers have attempted to design and realize colour scanning filters. These attempts are reviewed and evaluated. Each technique designs a colour scanning filter independent of physical realizability and then attempts to find the closest realizable approximation.

2.7.1 Construction of MacAdam's 'Most Orthogonal' Scanning Functions

Pearson and Yule [23] make an attempt to construct MacAdam's 'narrowest' or 'most orthogonal' scanning functions discussed in section 2.5. Allowing for the viewing

illuminant, the recording illuminant and other necessary light sources and filters, they obtain the desired responses with a cascaded (see equation 2.8) combination of the Kodak Wratten (gelatin) filters. The procedure of filter selection is extremely non-rigorous trial and error, in fact, the authors do not even define an error measure to be minimised. They admit that the error in the constructed filters is too large for colorimetry.

The authors claim that more sophisticated mathematical techniques were not used because the filter thicknesses were not continuously variable. Wratten filters are produced commercially in various densities but the densities are not continuosly variable. Glass filters, whose thicknesses can be continuously varied, are either too thick or too brittle, and could not be used due to space limitations. The authors do not discuss why no attempts were made to find any other 'most orthogonal' sets of filters that could perhaps have been more accurately constructed.

2.7.2 A Minimum (Square) Error, Product-Approximation in Transmission Space

Wright, Saunders and Gignac [41] use 'curve-fitting' and claim to have obtained a minimum square error approximation to a required spectral response for a photometer. The approximation, $\mathbf{m}_{approx}$, is constructed by cascading a set of filters. The required response, $\mathbf{m}$, is normalised by fixing a value at a particular wavelength.

The authors do not describe the program used to solve the system of non-linear equations that leads to the filter combination, nor do they indicate how positive densities of the filters are ensured. They claim to have limited the maximum density possible of the filters used in the construction so that the light transmitted through the constructed filter is sufficient to stimulate the photo-detectors that are often used

with scanning filters. They do not discuss how they obtained such a maximum value nor how it ensures that the transmission of light through the constructed filter is sufficient.

If the authors found the minimum-square-error approximation to the required spectrum, this is a useful contribution not just to filter design, but to optimisation theory. The minimum-square approximation represents the orthogonal projection, with respect to the euclidean norm, of the required spectrum onto the set, C_T, see equation (2.9). To this researcher's knowledge, such a projection is not readily found. As the authors do not describe the method used, it is difficult to comment on whether they have achieved what they claim.

2.7.3 A Signal-Specific Product Approximation

Wright, Saunders and Gignac [41] describe another error criterion, called the photometric error. It is defined as

$$e_k = \frac{\sum_{i=1}^{N} \mathbf{E}_k(i)\mathbf{m}(i) - \sum_{i=1}^{N} \mathbf{E}_k(i)\mathbf{m}_{approx}(i)}{\sum_{i=1}^{N} \mathbf{E}_k(i)\mathbf{m}(i)}$$

which can be seen to be

$$e_k = \frac{\sum_{i=1}^{N} \mathbf{E}_k(i)\mathbf{e}(i)}{\sum_{i=1}^{N} \mathbf{E}_k(i)\mathbf{m}(i)}$$

where $\mathbf{E}_k$ represents one of seven source spectra and $\mathbf{e}$ represents the error in constructing the filter. The above normalisation represents a fractional error. If the photometric error is interpreted as the component of the error (between the required and constructed filter responses) along a particular source spectrum vector $\mathbf{E}_k$, it should be normalised by the 2-norm of the source spectrum, $\mathbf{E}_k$.

An error measure, $\sum_{k=1}^{7} e_k^2$ is minimised. This minimisation is justified for photometers that are used for a particular set of source spectra, and their linear combinations, and is very source(signal)-specific. Once more, the authors do not indicate what kind of program was used to minimise this error.

The two methods described are used to design filters for nine source-detector combinations, which are then tested on the seven light sources, $\{\mathbf{E_k}\}_{k=1}^{7}$. For the case described in section 2.7.2, which is not signal-specific, the maximum photometric error obtained is about 7%. For the signal-specific method, a lower maximum photometric error of 0.6% is obtained, as expected. The results are good for the particular application, but are not related to other error measures.

2.7.4 Minimum (Weighted Square) Error, Product-Approximation in Density Space

Davies and Wyzecki [5] try to construct an approximation of the matching functions in cascaded (see equation (2.8)) form. For this, they define an error measure in the density domain. For a required filter $\mathbf{m}$, they attempt to match $log\ \mathbf{m}$ to $log\ \mathbf{m}_{approx}$ where

$$\mathbf{m}_{approx} = \alpha \prod_{i=1}^{n} \mathbf{r}_i^{u_i} \tag{2.11}$$

by minimising the error measure

$$\mathbf{e} = \sum_{i=1}^{N} (log\ \mathbf{m}(i) - log\ \mathbf{m}_{approx}(i))^2 \mathbf{m}(i) \tag{2.12}$$

Without the weighting factor $\mathbf{m}$, this error measure would give the orthogonal projection of $\mathbf{m}$ onto the set C_d in euclidean density space, defined in equation (2.10).

Working in the density domain has the advantage of reducing the unwieldy product expression (2.11) to a sum, making analytical minimisation results far easier to obtain. The inherent problem in the density domain of transforming small transmission values to large negative density values is reduced by the weighting factor $\mathbf{m}$ in (2.12). The decision to use $\mathbf{m}$ as a weighting function is made on the basis of trial and error. The other weighting functions used were powers of $\mathbf{m}$, and it was reported that the error in the tail ends of the spectra increased with an increase in this power. This is to be expected, because higher powers increase the weighting of the error at the peaks.

Solving the minimisation problem is not easy, as the error function is neither linear nor quadratic. If the error function was quadratic, as it would be without the weighting, quadratic programming could be used to solve the minimisation problem. The authors obtain the solution to the minimisation problem by trial and error. Filters that seem to look like the desired curve are used to get a rough fit, and other filters are used to trim the fit. This leads to underutilisation of the set of filters, and inaccuracy in the minimisation process. Plots of the error versus wavelength are presented. The maximum error is about 4 percent of the peak transmissivity. This error is not related to perceptual or tristimulus error.

If the vector $\mathbf{m}$ is assumed to have no zero values and, instead of just a factor $\mathbf{m}$(i) in equation (2.12), $\mathbf{m}(i)^2$ is used instead, the resulting error function defines an inner-product-norm. With this norm, the underlying density space is hilbertian. The solution to the error minimisation problem is then an orthogonal projection with respect to this norm, and as the set C_d may be shown to be convex this projection exists and is unique. The projection may be obtained by the method of quadratic programming.

An attempt is also made to fit the spectrum, log $\mathbf{m}$, exactly on a set of particular

wavelength values, $\{\lambda_k\}_{k=1}^{N_0}$ where $N_0 << N$. This method should not be expected to give results as good as those obtained by projecting the filter onto the span of the filters used, in the density domain. This is because the total error in a fit is not necessarily minimised by performing an exact fit at particular points.

2.7.5 Minimum (Square) Error Sum-Approximation

Davies and Wyczecki [5] also report attempts to approximate the matching functions with sums of filters from a given set. This is an extremely simple problem, it suffices to express $\mathbf{m}$ as a linear combination of the set of filters. Negative coefficients do not present a physical problem, and if $\mathbf{m}$ does not lie in the span of the given filters, a least-squares projection onto the span may easily be obtained. The results obtained by the authors are not physically feasible, as too many filters are required to produce an acceptable approximation. It should be noted that the number of filters required to approximate a given filter depends on the given filter set. Hence the fact that too many filters are required is not an inherent flaw of the sum-approximation. It is more difficult, however, to implement a solution with a large number of filters in the sum-approximation method than it is in the product-approximation method.

An interesting combination of filters was the product of a sum of filters with the best approximation obtained using cascaded filters. The main problem is still the fact that too many filters are needed. Errors obtained with this procedure seemed to be similar to those obtained by the procedure described in section 2.7.4.

2.7.6 Minimum (Infinity-norm) Error, Product-Approximation in Density Space

Falletti, Premoli and Rastello [9] have developed a method of finding an approximation to a given filter in the density domain. This approximation has a minimum infinity-norm error. Recall that the infinity norm of a vector $\mathbf{f}$ is defined as:

$$||\mathbf{f}||_\infty = max|\mathbf{f}(i)|$$

The approximation is constructed in the form of cascaded (see equation (2.8) filters. The filters used for the construction are glass filters, and filters that are too thin are rejected by the algorithm used, as are filter combinations with unacceptably low transmissivity. Analysis of the algorithm shows that the authors have done exactly what they claim to have, and that their technique is very efficient. It appears unlikely that there would be a more efficient algorithm to find a minimum-infinity-norm-error-approximation using cascaded filters.

The infinity-norms of the errors of the approximation are reported. Unfortunately, the authors do not go into a discussion of how the error translates into transmission space, although their approximations to the CIE Photopic Observer Response are very good, judging by the diagrams of the actual and the constructed responses. The maximum relative transmissivity obtained is low, at 40 percent, for the good approximations. They do not discuss why they did not attempt to minimise the square error, which would have seemed to be a natural choice. A disadvantage of the chosen minimisation is the fact that the maximum error can occur at any (and all) points in the signal. It is not known how the errors in density space relate to errors in tristimulus space or to errors in uniform colour spaces, like CIE $L^*a^*b^*$ space defined in section 1.1.2.

2.7.7 Simulation of CIE Illuminant D65

Liu, Berns and Shu [18] have designed coloured glass filters to simulate CIE illuminant D65. They design both a sum realization and a sum-of-products realization. The method used is almost entirely trial and error. The authors pick six critical parameters that represent the quality of the simulator, and attempt to construct one that matches desired values of these parameters within a certain range. Such a simulator is found by trial and error, by trying various combinations from the given set of glass filters. All combinations that satisfy these conditions are obtained, and then an 'optimal' one from within these is found, once more by trial and error. 'Optimal' is not defined mathematically.

Set theoretic techniques could probably have been used more profitably than the trial and error technique for obtaining the simulant. The six parameter values along with the respective ranges define sets in the space of visual spectra. The required simulant lies in the intersection of these sets. Instead of using trial and error once an acceptable solution has been obtained, variational techniques could be used to obtain an 'optimal' solution, though, before that can be done, 'optimal' must be more clearly defined.

The authors claim that the six parameters used to define the simulant cannot be related mathematically, and also that they cannot be optimised simultaneously. At least one of these statements is likely to be in error. It is probable that the parameters are related though a relationship may not be immediately obvious.

The trial and error method is inherently inaccurate and inefficient, practicality demands that many filter combinations cannot be tried. This limits possibilities, and makes the method inaccurate. Efficient trial and error techniques choose the next

trial based on the degree of feasibility of the previous one, thus moving towards a feasible solution. The authors attempt no such improvements in their method, and the result is a highly inefficient algorithm.

The results of the optimisation process are reported in the form of the values of the six parameters, and curves of both the D65 illuminant and the designed simulator are shown. It is probable that a more accurate and efficient optimisation algorithm could find a better fit.

2.7.8 The Vector Space Approach

The vector-space approach to the design and analysis of colour systems has been used by a number or researchers in recent years [3, 13, 31, 35, 38, 39, 40]. This section reviews some of the important terminology and notation of the vector space approach. This terminology makes it easier to frame the problem of the design of colour scanning filters in a manner that eliminates some of the disadvantages of the previously-reviewed methods.

Let $\mathbf{f}$ be a reflectance spectrum to be scanned. If $\mathbf{f}$ is illuminated by a scanning illuminant $\mathbf{l}$ then the resulting spectrum is one whose i^{th} value is $\mathbf{l}(i)\mathbf{f}(i)$. Let $\mathbf{o}$ be a vector representing the optical path of the illuminated spectrum. The signal that reaches the scanning filters has i^{th} value $\mathbf{o}(i)\mathbf{l}(i)\mathbf{f}(i)$. The output of a scanning filter $\mathbf{m}$ for such a signal has i^{th} value $\mathbf{m}(i)\mathbf{o}(i)\mathbf{l}(i)\mathbf{f}(i)$. If $\mathbf{d}$ represents the detector response, then the output of the detector is $\sum_{i=1}^{N} \mathbf{m}(i)\mathbf{d}(i)\mathbf{o}(i)\mathbf{l}(i)\mathbf{f}(i)$. This can be represented as $\mathbf{m}^T\mathbf{H}\mathbf{f}$, where $\mathbf{H}$ is a diagonal matrix such that $\mathbf{H}_{ii} = \mathbf{d}(i)\mathbf{o}(i)\mathbf{l}(i)$. Hence, if $\mathbf{M}$ represents a set of scanning filters and $\mathbf{H}$ the combined effect of the scanning illuminant, the optical path and the detector response, $\mathbf{M}_H = \mathbf{H}\mathbf{M}$ denotes the scanning system. The vector $\mathbf{h}$, such that $\mathbf{h}(i) = \mathbf{H}_{ii}$, is defined as the scanner

characteristic. Let the set $\{\mathbf{Hm}_i\}_{i=1}^r$ be defined as the set of 'effective' scanning filters. The set of linear combinations of the effective scanning filters may be denoted $R(\mathbf{M}_H)$. The output of the scanning system represented by $\mathbf{M}_H$ is:

$$\mathbf{c} = \mathbf{M}_H^T\mathbf{f} \tag{2.13}$$

A set of effective scanning filters which is exactly the set $\{\mathbf{La}_i\}_{i=1}^3$ will give output values $\mathbf{t}$. Given the set $\{\mathbf{La}_i\}_{i=1}^3$ it may not be possible to construct exactly a set of filters $\{\mathbf{m}_i\}_{i=1}^r$ such that $\mathbf{Hm}_i = \mathbf{La}_i$. A method that seeks to construct exactly the vectors $\{\mathbf{La}_i\}_{i=1}^3$ does not allow flexibility in the incorporation of limitations of the fabrication process or the fact that it is the space $R(\mathbf{A}_L)$ that is important.

Notice, however, that any invertible transformation of the tristimulus values can be manipulated to give the tristimulus values. Hence any such transformation of the tristimulus values will also serve to characterize the visual stimulus of the signal. If $\mathbf{X}$ represents an invertible transformation, it is enough to obtain a set of readings $\mathbf{M}_H^T\mathbf{f}$ such that

$$\mathbf{M}_H^T\mathbf{f} = \mathbf{X}\mathbf{A}_L^T\mathbf{f}$$

This is possible when

$$\mathbf{M}_H = \mathbf{A}_L\mathbf{X}^T$$

and

$$R(\mathbf{M}_H) = R(\mathbf{A}_L)$$

This is not a necessary condition, however. For example, it is possible that 3 filters whose span is equal to the HVISS cannot be constructed, but four filters whose span includes the HVISS can be. See [35] for an example. In such a case, there exists a linear transformation $\mathbf{Y}$, such that

$$\mathbf{Y}\mathbf{M}_H^T\mathbf{f} = \mathbf{A}_L^T\mathbf{f}$$

where $\mathbf{Y}$ is not necessarily square nor invertible. This implies that

$$\mathbf{M}_H \mathbf{Y}^T = \mathbf{A}_L$$

and

$$R(\mathbf{M}_H) \supseteq R(\mathbf{A}_L) \tag{2.14}$$

This implies that a perfect set of effective scanning filters is one whose span includes the HVISS.

A number of problems arise in the construction of a scanning system to obtain the required projection of $\mathbf{f}$. In particular, it is difficult to fabricate a designed scanning filter exactly and errors in filter construction will, in general, change the space spanned by the filters, resulting in an error in the measurement of the required projection. This error will lead to an error in the reproduction. None of the methods reviewed in sections 2.7.1-2.7.7 resulted in a perfectly constructed filter.

In the colour scanning problem, it is important not to construct particular filters exactly, but to construct filters that span a given space. Construction errors result in reproduction errors because the space spanned by the scanning filters is not the space that needs to be spanned. This implies that the optimization criteria used in the methods reviewed are lacking, simply because these criteria evaluate a single scanning filter based on how close it is to the corresponding CIE function for a given illuminant. The methods discussed do not incorporate the requirement of equation (2.14) into the design procedure. With this in mind, the next chapter proposes a measure which may be used to evaluate filter sets with respect to requirement (2.14).

2.8 More than Three Filters in Colour Scanning

Vrhel and Trussell [38, 39] address the problem of colour correction using the requirement of equation (2.14). Colour correction methods attempt to estimate the tristimulus values of a colour signal under more than one viewing illuminant. This implies that perfect colour correction requires the projection of the spectrum $\mathbf{f}$ onto the HVISS of each of the viewing illuminants. If the number of viewing illuminants is two, the space onto which the projection of $\mathbf{f}$ needs to be determined could be of dimension as large as six. The minimum number of scanning filters required to ensure perfect colour correction in such a case would be, as indicated by requirement (2.14), six. As the number of possible viewing illuminants increases, the dimension of the space to be spanned also increases, and hence the minimum number of scanning filters for perfect scanning also increases.

In [39] Vrhel and Trussell discuss ways and means to define the parameters required for 'best' colour correction given a set of P scanning filters and K illuminants. They discuss approaches using prior knowledge of the spectra of the viewing illuminants and the reflectance spectra of the data set. The error measures used to judge a reproduction are mean square tristimulus error and mean ΔE_{Lab} error over the data set. The simulation results presented indicate a significant improvement in colour correction, as should be expected, with the addition of a fourth filter to the set of three scanning filters. In other words, the knowledge of the projection onto a four-dimensional space improves colour correction results considerably, over the conventional methods that use the projection onto a three-dimensional space.

The methods of Vrhel and Trussell result in the definition of a P-dimensional space to be spanned. This P-dimensional space implies the optimal use of a set of

P scanning filters. Once such a space is defined, a set of P scanning filters needs to be physically realized such that the set of P effective scanning filters spans the space defined as closely as possible. To define and realize such a set, an optimization criterion that is related to how closely the space is spanned, is required. Such a criterion is developed in Chapter 3. The next section deals with existing measures of the effectiveness of scanning filters.

2.9 The Quality Factor Measure of Neugebauer

In an attempt to measure the goodness of a colour filter with respect to the error that occurs due to mismatch of spaces spanned, the quality factor or the q-factor of a colour filter was defined by Neugebauer [22]. If $\mathbf{m}$ represents a colour filter and $P_V(\mathbf{m})$ its orthogonal projection onto the HVISS, the q-factor of $\mathbf{m}$ is defined as:

$$q(\mathbf{m}) = \frac{||P_V(\mathbf{m})||^2}{||\mathbf{m}||^2} \tag{2.15}$$

where $||.||$ is the 2-norm in N-dimensional vector space. Further, it is clear from [22] that

$$q(\mathbf{m}) = \frac{\sum_{i=1}^{3} \mathbf{t}_i^2}{||\mathbf{m}||^2} \tag{2.16}$$

where the $\mathbf{t}_i$ are the tristimulus values of $\mathbf{m}$ with respect to some orthonormal basis for the HVISS (and not with respect to the CIE matching functions).

Notice that $0 \leq q(\mathbf{m}) \leq 1$, and the closer the value of q($\mathbf{m}$) to unity, the 'better' the colour scanning filter $\mathbf{m}$. If the value of q($\mathbf{m}$) is small compared to unity, the filter measurement does not give much information about the fundamental of the measured signal, and hence the filter is not appropriate for colour scanning. The q-factor seems

a reasonable quality measure for filters not in the HVISS because $||\mathbf{m}||^2(1 - q(\mathbf{m}))$ is the square of the euclidean distance of $\mathbf{m}$ from the HVISS. If any one of a set of three scanning filters, $\{\mathbf{m}_i\}_{i=1}^3$ is not in the three-dimensional HVISS (i.e. $q(\mathbf{m}_i) \neq 1$ for some i), then $R(\mathbf{LA}) \neq R(\mathbf{M})$, and $\{\mathbf{m}_i\}_{i=1}^3$ is inaccurate for colour sensing. A major disadvantage of the q-factor is that it is designed to be used with only a single filter. A measure that extends the idea of the q-factor to judge the effectiveness of a set of colour scanning filters would be very useful.

In most existing scanning systems, only three scanning filters are used. Three linearly independent scanning filters span the three-dimensional HVISS iff all three have unit q-factors. Hence the q-factor indicates a perfect set of filters if $q(\mathbf{m}_i) = 1$ *for every* i and the $\mathbf{m}_i$ are linearly independent vectors. It does not indicate whether the three filters are linearly independent or not, nor can it assist in differentiating among imperfect sets of filters. Hence, the q-factor cannot be used by itself to indicate a 'better' imperfect set of filters. Another disadvantage of the q-factor is that it may be used to judge at most a set of three filters.

There are at least two reasons why more than three filters may be used to improve the quality of the colour reproduction. First, in many cases, three parameters are not enough to define sufficiently the visual stimulus of an N-dimensional signal for colour correction. Typically, such a situation arises when the colour reproduction is to be viewed under two different illuminants. In such a case, as many as six parameters (representing the projections of the signal onto the two different three-dimensional Human Visual Subspaces defined by the two different illuminants) may be required to accurately represent the signal [31, 38, 39]. Second, the constraint of constructability on the filters might imply that no set of three constructable filters can span the HVISS though a set of four filters could be constructed so that the required projection is

obtained. When more than three parameters (four scanning filters, for example) are necessary, the q-factor is not an effective measure of the goodness of even each single filter as part of the set of more than three filters. For example, suppose $\{\mathbf{Hm}_i\}_{i=1}^4$ is a set of effective (see section 2.7.8 for a definition) scanning filters. It is possible that the HVISS is contained in the span of the set of four effective filters, i.e.:

$$R(\mathbf{M}_H) \supseteq R(\mathbf{A}_L)$$

but, $q(\mathbf{Hm}_i) < 1$ for $i = 1, 2, 3, 4$. Such a set could provide perfect colour scanning though the individual q-factors would not indicate this. An example of such a set is presented in Chapter 3.

The performance of a set of filters can be judged by the reproduction quality of a set of signals. Usually this set of signals is chosen so as to represent the ensemble of signals that the set of scanning filters is to be used on. The average error in the reproduction is often used as an indicator of the goodness of the set of filters. The q-factor of a colour scanning filter has the disadvantage of not being a good indicator of the perceptual error in colour reproduction.

As just discussed, the q-factor has three major disadvantages. First, it measures a single filter independently of other filters in the scanning set. A measure that extends the idea of the q-factor to judge the effectiveness of a set of colour scanning filters would be very useful. Second, it can be used to judge the merit of a single filter as part of a set of three filters in a limited sense because it indicates a perfect set of three filters only when they are linearly independent. The q-factor itself does not indicate linear independence for a perfect set of filters, nor does it differentiate among imperfect filter sets. Third, it is not useful in judging the merit of a single filter as part of a set of more than three filters. It is possible to develop a measure that

overcomes these disadvantages.

Engeldrum [7] suggests the average quality factor of a set of three scanning filters as a measure of the quality of the set. This measure is shown to be inadequate in Chapter 3. Firstly, it may not be used to evaluate a set of r filters ($r > 3$) which is used to span an r-dimensional space. Hence, for example, it is not useful in colour correction. Secondly, it may not be used to evaluate a set of r filters ($r > 3$) which is used to span a space of dimension smaller than r. Hence, for example, it may not be used to ensure spanning of the three-dimensional HVISS with a set of four scanning filters. Thirdly, the average quality factor of a set of three filters is indicative of the colorimetric performance of the set of filters only if the filters are non-overlapping. This is demonstrated in Chapter 3 through analysis and simulations.

2.10 Summary

A number of techniques that are used for the design of colour scanning filters have been reviewed and analyzed. It is concluded that the techniques are inadequate in part because they do not allow for a linear combination of the tristimulus values to be the goal of the scanning process. Instead of attempting to match each filter exactly, it would be more profitable to attempt to match the entire set of designed filters. The incorporation of the linear transformation of scanner measurements into the design procedure should allow for considerable flexibility in the design procedure. Other inadequacies of the reviewed methods include the lack of a rigourous optimization criterion and method, the failure to evaluate an imperfect filter set, and the failure to handle a set of more than three filters.

Current measures of goodness of a single filter and a set of colour scanning filters

are reviewed. A need for a measure that evaluates a set based on the mismatch between the space spanned by the scanning filters and the space to be spanned has been identified, and such a measure is developed in Chapter 3.