A metric for predicting perceptual blackness

S Westland, TLV Cheung and OR Lozman*, Centre for Colour Design Technology, University of Leeds (UK), *FujiFilm Imaging Colorants Ltd, Manchester (UK)

Abstract

The performance of the black ink in ink-jet printing is of great importance to both ink-jet manufacturers and ink suppliers. Many black ink formulations contain several dyes or pigments and have a noticeable hue. The industry requires a suitable method for assessing the relative perceptual blackness of black inks. In this study, a set of black printed samples have been visually ranked in terms of perceptual blackness. Various candidate blackness indices have been evaluated and the best has been shown to be able to make pair-wise blackness comparisons with greater accuracy than the average observer.

Introduction

The performance of the black ink in ink-jet printing is of great importance to both ink-jet manufacturers and ink suppliers. Many black ink formulations contain several dyes or pigments and some have a noticeable hue. Current research is concerned with the design of new black ink formulations that produce optimal performance in a range of different properties. One of the key properties, however, is whether the ink formulation produces a satisfactory black colour. This can only be assessed visually or by measurements that can be correlated with visual assessments. Whereas the assessment of whiteness – in terms of both visual assessment and instrumental whiteness indices – has been extensively researched (because of the importance of good whites in certain industries such as textiles and paper) blackness has been less well studied. A collaborative project between Leeds University and FujiFilm Imaging Colorants Ltd aims to develop an instrumental method to assess perceptual blackness. A set of black printed samples produced using a variety of ink formulations were provided by FujiFilm Imaging Colorants and have been assessed visually in terms of perceptual blackness. The spectral reflectance factors for the samples were also measured to allow various candidate blackness indices to be developed and their performance compared with the visual assessments.

Experimental

A set of 100 black samples were prepared using an ink-jet printer and a variety of ink formulations. The samples were cropped to a size of 2.5cm × 2.5cm and mounted onto Munsell N5 grey card. The spectral reflectance factors were measured for each of the black samples at 10nm intervals in the visible spectrum to allow the calculation of CIE tristimulus values and various candidate blackness metrics. A ranking method was proposed for the visual assessment but it was considered inappropriate and impractical to request observers to rank 100 samples simultaneously. Therefore, the samples were randomly divided into 5 subsets each containing 20 samples. Figures 1 and 2 illustrate the samples in b^* vs a^* and L^* vs C^* diagrams respectively. Figures 3 shows the volumetric area of the samples in CIELAB space.

The samples of each subset in turn were viewed by observers in a viewing cabinet illuminated by a light source approximating the D65 illuminant. Twenty five observers were recruited to take part in psychophysical experiments and their colour vision was assessed using the Farnsworth-Munsell 100-hue test. Two of the observers were deemed to have either abnormal colour vision or poor colour discrimination and therefore twenty three observers were asked to rank the samples in order of their perceptual blackness. The rank orders were converted to interval-scale Z values using Torgerson's Categorical Scaling method[1,2]. One of the advantages of analyzing Z scores is that Z scores are interval data and their differences are linearly related to differences in the corresponding visual assessments.

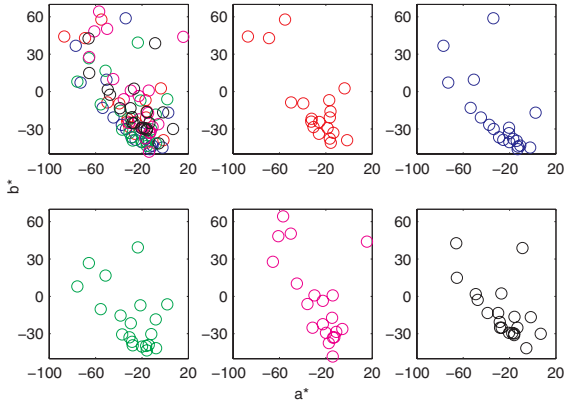

*Figure 1: Colour distributions of the 100 black samples (upper row left) and the five subsets in CIELAB a*b* space.*

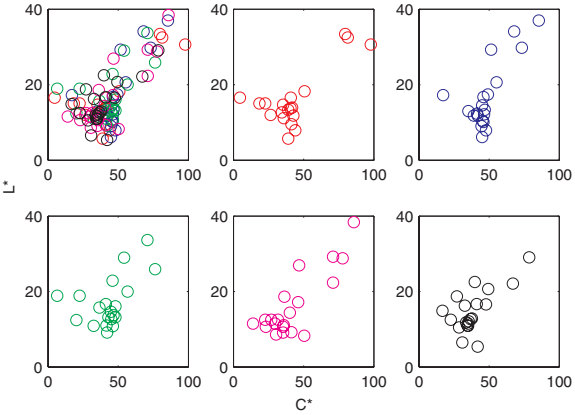

*Figure 2: Colour distributions of the 100 black samples (upper row left) and the five subsets in CIELAB L*C* space.*

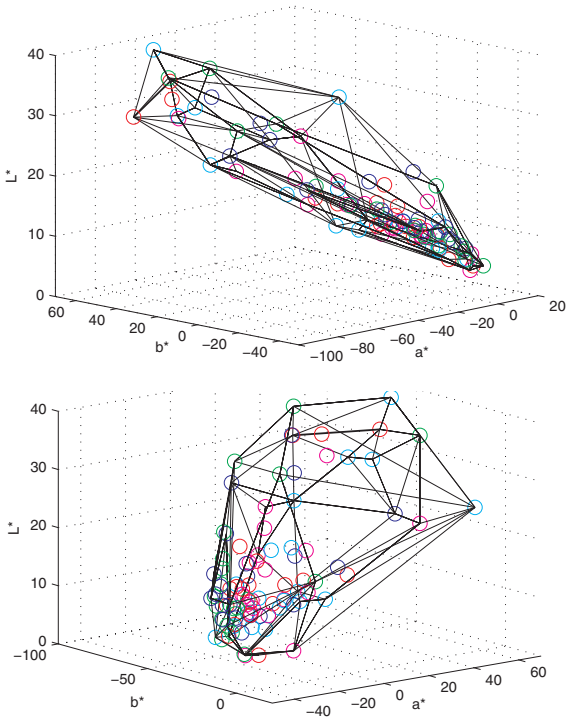

Figure 3: CIELAB convex hull (meshed facets) of 100 black samples.

Blackness metrics

Various metrics were considered for the prediction of perceptual blackness. One of the methods was inspired from studies of whiteness prediction[3,4]. The CIE whiteness metric is based on a linear combination of luminance and chromaticity, designed so that the components can be weighted to suit various hue preferences.

The first candidate metric (B1) is therefore given as equation 1,

$$B1 = a_1 + a_2Y + a_3(x - x_n) + a_4(y - y_n), \quad (1)$$

where Y,x,y are the luminance and chromaticities of the samples, x_n,y_n are the chromaticities of the white point (D65 illuminant) and the coefficients a_1-a_4 were variables whose values were determined by optimization. This metric will have a clear hue preference because of the terms in parentheses. An alternative metric (B2) was also considered, thus

$$B2 = a_1 + a_2Y + a_3(x - x_n)^2 + a_4(y - y_n)^2. \quad (2)$$

Whereas equation B1 would exhibit a strong hue preference the rationale behind equation B2 is that perceptual blackness may be correlated with the saturation of the sample (the smaller the saturation, the better the blackness) and that the coefficients a_2-a_4 would allow various hue directions to be weighted and also allow saturation and luminance to be appropriately weighted. Thus, in contrast with B1, B2 assumes that for a fixed luminance a perfectly neutral sample would be perceived as being blacker than any saturated sample. Results (which will be discussed later) showed that the performance of B2 was superior to that of B1. The notion of applying an analogue (B3) of B2 in the approximately uniform CIELAB space was also examined, thus

$$B3 = a_1 + a_2L^* + a_3a^{*2} + a_4b^{*2}, \quad (3)$$

where L^*, a^* and b^* are the CIELAB coordinates of the samples. Finally, an additional metric (B4) which employed the simple weighted sum of tristimulus values was considered.

$$B4 = a_1 + a_2X + a_3Y + a_4Z, \quad (4)$$

In all cases, the a_1 coefficient was required to provide sufficient freedom in the models to fit the data given the fitting criterion which is discussed in the next section.

Data Analysis

The values of the weighting parameters in equations 1-4 were optimized to minimize the root-mean-squared (rms) error between the blackness index and the Z values derived from the psychophysical experiment. Optimization was performed using MATLAB's *fminsearch* function which performs a multidimensional unconstrained nonlinear minimization (Nelder-Mead). The Nelder-Mead method is a simplex method for finding a local minimum of a function of several variables, in this case a_i. The starting values for all cases were that $a_i = 1$, $i \in \{1,2,3,4\}$. The performance of the optimized equations was assessed by the rms error score. The performance of the candidate metrics was additionally assessed using the wrong-decision criterion[5] (WDC). In the WDC method the average or consensus rank order for all observers is deemed to be correct and can be used to define the relative ranking between any pair of samples. Thus if a sample A is higher in the average ranking than another sample B then it is deemed that sample A is blacker than sample B. An individual observer may agree with this 'decision' if they rank sample A higher than sample B or may disagree with the decision if they rank sample A lower than sample B. In the latter case, the observer is said to have made a wrong decision. Adjacent pairs in the consensus rank order were analyzed and the number of (or per cent) wrong decisions made by each observer was calculated. The per cent wrong decisions made by each of the metrics were also calculated on a similar basis by comparing each metric decision with that of the visual consensus.

Finally, the same group of observers assessed all five sample subsets and the candidate blackness indices were optimized for each set of visual data.

Results and Discussion

The mean per cent wrong decisions made by the observers was 35% (individual errors ranged from 22% - 55%). (We note that if every pair had been compared with every other the observer performance would have been better since by analyzing only those pairs formed from samples adjacent in the consensus order we are considering only the most difficult comparisons.) Tables 1 and 2 summarize the training performance obtained from the models for each subset of samples. So, for example, the third row of data in each table shows the results of the five B3 models, each of which was optimized to one of the subsets.

Table 1: Summary of training performance (rms)

model	B1	B2	B3	B4
subset 1	1.1526	0.9489	0.7633	1.0085
subset 2	0.7777	0.6859	0.6693	0.7704
subset 3	1.2715	1.5284	1.2258	1.4371
subset 4	1.9451	1.7898	1.9797	2.1831
subset 5	0.9334	0.7405	0.6924	0.9364
average	1.2161	1.1387	1.0661	1.2671

Table 2: Summary of training performance (%WD)

model	B1	B2	B3	B4
subset 1	38.9	38.9	27.8	33.3
subset 2	31.6	31.6	36.8	31.6
subset 3	31.6	42.1	31.6	36.8
subset 4	47.4	47.4	52.6	36.8
subset 5	36.8	42.1	47.4	47.4
average	37.3	40.4	39.2	37.2

To illustrate the degree to which the fits indicated by Table 1 are acceptable, Figure 4 illustrates an example of the performance of model B3 when it was optimized to fit subset 1.

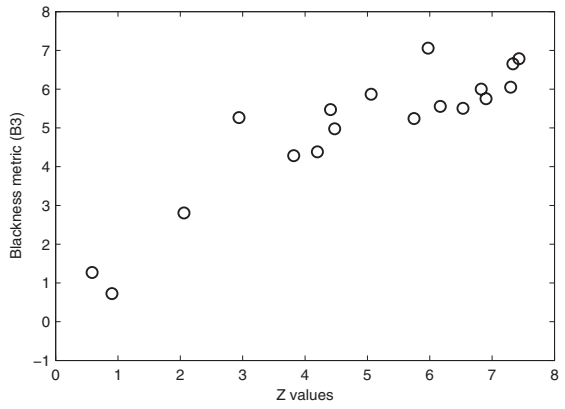

Figure 4: Performance of model B3 for subset 1 (rms = 0.7633).

Table 1 shows that the best predictions in rms terms were made by the B3 model. However, in terms of the WDC (Table 2) it is not so clear which of the models gives the best performance. Moreover, to evaluate the models robustly it is necessary to know the generalization ability of the models. Table 3 summarizes the testing performance. For each class (B1-B4) of model five distinct sets of coefficients are generated depending upon which data subset the coefficients were optimized on. Each of these sets of coefficients is then tested on the other subsets and performance averaged to yield a generalization or testing measure.

Table 3 shows the best performing model is B3 (Eqn 3) which makes 37.73% wrong decisions compared with the average visual performance of 35% (the best observer makes 22% wrong decisions). Final models were generated by averaging the five sets of coefficients for each model class and Table 4 shows the WDC results when these average models are tested on the various data subsets. We note that even for the best model the performance is limited by the accuracy to which the visual assessments were obtained. It is interesting, but not totally unexpected, that the best model is one based on an approximately perceptually uniform colour space. Other spaces (including other colour-appearance spaces) will be explored in further work. However, overall, the performance of the B3 blackness index in this study is encouraging.

Table 3: Summary of testing performance (%WD)

model	subset	1	2	3	4	5	avg	overall
B1	1	---	36.8	47.4	47.4	31.6	40.8	41.8
	2	38.9	---	47.4	52.6	36.8	43.9	
	3	44.4	31.6	---	42.1	42.1	40.1	
	4	50.0	42.1	47.4	---	42.1	45.4	
	5	33.3	31.6	42.1	47.4	---	38.6	
B2	1	---	42.1	47.4	42.1	36.8	42.1	42.8
	2	33.3	---	47.4	52.6	47.4	45.2	
	3	38.9	47.4	---	42.1	36.8	41.3	
	4	38.9	47.4	42.1	---	47.4	43.9	
	5	33.3	42.1	42.1	47.4	---	41.2	
B3	1	---	36.8	36.8	36.8	42.1	38.2	37.7
	2	33.3	---	42.1	31.6	42.1	37.3	
	3	38.9	31.6	---	31.6	36.8	34.7	
	4	38.9	47.4	47.4	---	42.1	43.9	
	5	27.8	36.8	42.1	31.6	---	34.6	
B4	1	---	31.6	31.6	42.1	42.1	36.8	38.4
	2	36.8	---	47.4	36.8	47.4	42.1	
	3	33.3	36.8	---	42.1	36.8	37.3	
	4	38.9	42.1	42.1	---	42.1	41.3	
	5	33.3	26.3	42.1	36.8	---	34.6	

Table4: Summary of testing performance (%WD) using averaged coefficients over 5 subsets

model	subset 1	subset 2	subset 3	subset 4	subset 5	avg
B1	38.9	36.8	42.1	47.4	42.1	41.3
B2	33.3	36.8	47.4	47.4	42.1	41.2
B3	33.3	36.8	42.1	31.6	47.4	36.0
B4	38.9	36.8	42.1	31.6	36.8	37.4

$$B3 = 8.6542 - 0.2583L^* - 0.0052a^{*2} + 0.0045b^{*2}, \quad (5)$$

The best-performing model based upon Table 4 is of the B3 form and is shown in Equation 5. When observers were asked to rank the samples in order of perceptual blackness they were also asked to signify the position in the ranking below which they didn't consider the samples to be black at all. Results are shown for one subset in Table 5. Those samples above the dashed line in Table 5 are deemed to be black because more than 50% of the observers agreed that they were black.

A summary of all the data is shown in Figure 5 where it can be seen (upper-left pane) that there were limitations in the sample availability that have limited the impact of this work. For example, note that samples close to the point $a^* = b^* = 0$ are not classed as being black. This is because these samples were too light (see Figure 6). However, intuition informs us that an achromatic sample of very low Lightness should certainly be seen as black. Also, the limit of how chromatic samples may be before they cease to appear to be black was clearly not reached in the sets of

samples, particularly in the blue direction. The reason for this limitation is the particular ink range and substrate that was used. It is suggested that further work of this nature is needed with a larger and more comprehensive set of samples in order to validate the blackness metric that is proposed (Equation 5).

Table 5: Analysis of one subset showing the per cent observers classing each sample as black

observers' agreement (%)	L*	a*	b*
100	10.11	-12.77	-44.13
100	11.76	-13.77	-39.65
96	7.86	-12.93	-46.20
100	13.29	-28.95	-36.78
100	12.09	-25.42	-38.76
100	12.29	-16.85	-37.93
96	8.99	-10.83	-43.35
96	10.41	-20.23	-40.52
78	16.70	-37.49	-26.77
70	6.14	-2.25	-44.82
74	14.51	-33.81	-29.95
74	11.82	-20.58	-33.44
26	12.99	-20.26	-28.87
17	17.41	-44.67	-20.69
22	20.67	-53.80	-12.97
0	17.18	2.56	-16.87
0	29.84	-72.90	7.09
4	29.34	-50.68	9.63
0	36.97	-77.14	36.77
0	34.13	-33.97	58.84

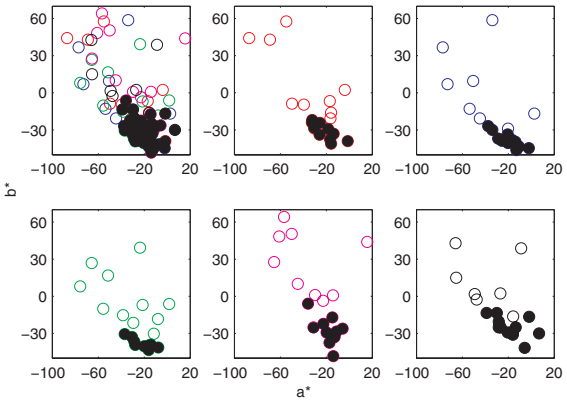

*Figure 5: Colour distributions of the 100 black samples (upper row left) and the five subsets in CIELAB a*b* space (closed symbols represent those samples that more than 50% of observers class as being black).*

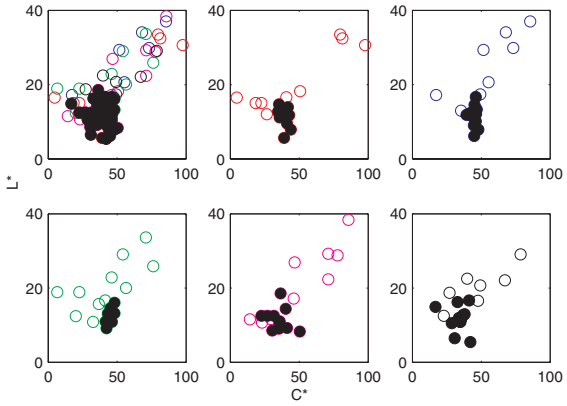

*Figure 6: Colour distributions of the 100 black samples (upper row left) and the five subsets in CIELAB L*C* space (closed symbols represent those samples that more than 50% of observers class as being black).*

References

[1] C.J. Bartleson, Measuring Differences, in Bartleson CJ and Grum F, editor: *Optical Radiation Measurements: Volume 5*, Academic Press (Orlando), (1984).

[2] W.S. Torgerson, *Theory and methods of scaling*, John Wiley and Sons (New York), (1962).

[3] E. Ganz, Whiteness: Photometric Specification and Colorimetric Evaluation, *Applied Optics*, **15** (9), 2039-2058, (1976).

[4] E. Ganz, Whiteness formulas: a selection, *Applied Optics*, **18** (7), 1073-1078, (1979).

[5] R. McDonald, *Colour Physics for Industry*, Society of Dyers and Colourists (UK), (1987).

Author Biography

Stephen Westland is Professor of Colour Science and Technology at the University of Leeds (UK). He has published about 100 technical papers in the area of colour science and his current research interests are human contrast sensitivity, blackness, and the application of colorimetry to dentistry and design.