JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2006) **3**, 71–85
doi:10.1098/rsif.2005.0076
Published online 6 September 2005

Optimality of the basic colour categories for classification

Lewis D. Griffin[†]

Department of Computer Science, University College, London, UK

Categorization of colour has been widely studied as a window into human language and cognition, and quite separately has been used pragmatically in image-database retrieval systems. This suggests the hypothesis that the best category system for pragmatic purposes coincides with human categories (i.e. the basic colours). We have tested this hypothesis by assessing the performance of different category systems in a machine-vision task. The task was the identification of the odd-one-out from triples of images obtained using a web-based image-search service. In each triple, two of the images had been retrieved using the same search term, the other a different term. The terms were simple concrete nouns. The results were as follows: (i) the odd-one-out task can be performed better than chance using colour alone; (ii) basic colour categorization performs better than random systems of categories; (iii) a category system that performs better than the basic colours could not be found; and (iv) it is not just the general layout of the basic colours that is important, but also the detail. We conclude that (i) the results support the plausibility of an explanation for the basic colours as a result of a pressure-to-optimality and (ii) the basic colours are good categories for machine vision image-retrieval systems.

Keywords: colour histograms; image retrieval; colour quantization; basic colour terms

1. INTRODUCTION

The question of why the continuously variable quality of colour is linguistically segregated into the categories of the basic colours (red, blue, etc.) has been studied by anthropologists, linguistics and psychologists for 50+ years. Recently, researchers in machine vision have also considered colour categorization; not for communication but to facilitate retrieval from image databases. We have investigated whether these two fields of study have an overlap, by assessing whether optimal categories for machine vision agree with human linguistic categories. Our results are consistent with this agreement. Thus, the contribution that we hope to make to the debate on the origin and nature of human colour categories, is to increase the plausibility of a pressure-to-optimality explanation. In this introduction we review work on colour categorization by human- and machine-vision.

1.1. Nature of the basic colours

Although the quality space Gärdenfors (2000) of colours form a continuous manifold (Riemann 1854), superficially free of landmarks and subdivisions, human language contains colour words that reference particular colours or regions of colour. These colours' names vary in status; typically there are a special few that (i) are not defined primarily by reference to others, (unlike, for example, 'yellowish-green') and (ii) have unrestricted applicability (unlike, for example, 'blond'). Such names are said to be 'basic colour terms' (BCTs), and their referents are 'basic colours'. Since being proposed, the criteria for 'basic-ness' have been refined beyond the simple statements above and added to (Berlin & Kay 1969), but difficult cases still arise (Paramei 2005). However, interest in basic-ness does not primarily revolve around asking which or why some terms are basic, rather the focus is on whether and why the basic colours of different languages are the same.

An extreme position that can be taken on these questions is 'linguistic relativism' which maintains that the basic-ness or not of colours are facts that are grounded completely in language, and as such should be expected to vary between cultures dependent on local concerns and idiosyncratic history. At the other end of a spectrum of positions is 'semantic universalism' which holds that the basic colours are widely agreed upon across cultures, so simple and correct translation is possible. This debate, now in its fifth decade, is often characterized as a battle (Brown 1991; Kay 1999; Saunders 2000) with both sides sometimes suspecting the other of a more substantial agenda than settling a question in colour science.

Central to the debate is Berlin & Kay's landmark study (Berlin & Kay 1969) of colour words in 20 languages. On the basis of that study they claimed that the 'basic color terms of any given language are always drawn' from a universal inventory of 11: black, grey,

[†] l.griffin@cs.ucl.ac.uk

Received 18 April 2005
Accepted 3 August 2005

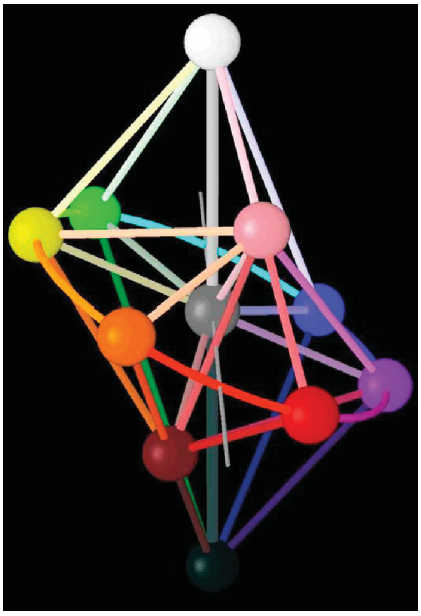

Figure 1. This refinement of a previously published diagram (Griffin 2001), shows the psychological structure of the 11 basic colours.

white, red, orange, yellow, green, blue, purple, pink and brown (see figure 1). These 11 colours are now often referred to as the basic colours.

Berlin and Kay's thesis has been challenged on several grounds, including the following.

(i) Criticism of the methodology used (Lucy 1997; Saunders & van Brakel 1997; Roberson *et al.* 2000).
(ii) Identification of counter-examples in non-industrialized societies. For example, the language of the Berinmo, a Melanesian people, has only five basic colour categories, but two of them—'nol' and 'wor'—have a shared boundary that lies firmly within the English 'green' region (Davidoff *et al.* 1999; Roberson *et al.* 2000).
(iii) Identification of counter-examples in industrialized societies. For example, the French term 'pourpre' does not include violet colours as the English term 'purple' does (Schirillo 2001), and Russian has arguably an additional BCTs for light blues (Paramei 2005).
(iv) Small but significant inter-cultural variation in the colour chosen as the most pure exemplar of a category, even between cultures with superficially the same system of categories (Webster *et al.* 2000; Lin *et al.* 2001).

However, despite these evidential challenges the findings of the recently completed World Colour Survey of 110 languages broadly support the thesis that there is a strong tendency towards BCTs with similar referents to the Berlin and Kay 11 BCTs Kay & Regier 2003; Kay *et al.* 2005). It remains to be seen whether the improved methodology of the World Colour Survey will answer all the criticisms levelled against the original Berlin and Kay study.

1.2. Explanations of the basic colours

Assuming that one accepts the evidence for inter-cultural consensus on which colours are basic, then one must accept that some part of the explanation of this must reside in a universally accessible locus. Shared neurophysiology, language, ecological optics and visual ecology have all been suggested and we review these below. The aim of this review is to convey the types of explanations that have been proposed, rather than to enumerate only those proposed explanations, which have not yet been disproved. For example, the hypothesis that the BCTs might be explicable in terms of cone-opponent channels, although an historically important hypothesis, has been undermined by studies (Webster & Mollon 1994; Webster *et al.* 2000; Kuehni 2004) that have shown that some unique hues (those judged to be pure yellow, green or blue) do not correspond to cone-opponent axes, though it is possible that they correspond to recently discovered non-opponent coding mechanisms (D'Zmura & Knoblauch 1998).

Neurophysiological explanations tie the basic colours to the joint measurement span of the cone spectral-sensitivity functions (Griffin 2001; Buchsbaum & Bloch 2002), to some later processing stage such as opponent channels (Hering 1920; Hurvich & Jameson 1957; Kay & Maffi 1999), or to dedicated neural mechanisms (Steels & Belpaeme in press). Neurophysiology may also limit what categories are possible, for example we may lack cognitive structures capable of representing disconnected or non-convex regions of colour space (Gärdenfors 2000).

Explanations based on shared language appeal to factors such as: limitations of some language-acquiring brain module (Dowman 2002), effects of the process of achieving consensus on semantics (Steels & Belpaeme in press), and advantages of agreement despite inter-individual variations in colour vision (Jameson in press).

Explanations from ecological optics (Gibson 1979) consider how common physical processes affect colour, and what invariants exist despite these processes. Examples are: categories being shaped so that they exploit that neither shadowing nor highlights alter the hue of reflected light; and categories being such that they achieve reasonable stability despite variations in illuminant (D. Bimler 2004, personal communication.

In *explanations from visual ecology*, the colour statistics of the environment are taken into account. For example, categories could correspond to clusters of naturally occurring colours (Yendrikhovskij 2001). Another possibility is that categories are particular effective for certain types of interaction with the world, such as search, identification, recognition, discrimination or classification (Roberson 2005). Effectiveness for classification ties in well to a recent suggestion about psychological categories in general: good systems of categories are those that effectively support induction (Ellison 2001). In the context of colour, the argument would go like this. A 'green' category is useful as it allows inferences like the following: the majority of 'green' things that I have seen have been plants, therefore, this 'green' thing is probably a plant. If instead of a 'green' category one had 'turquoise' and

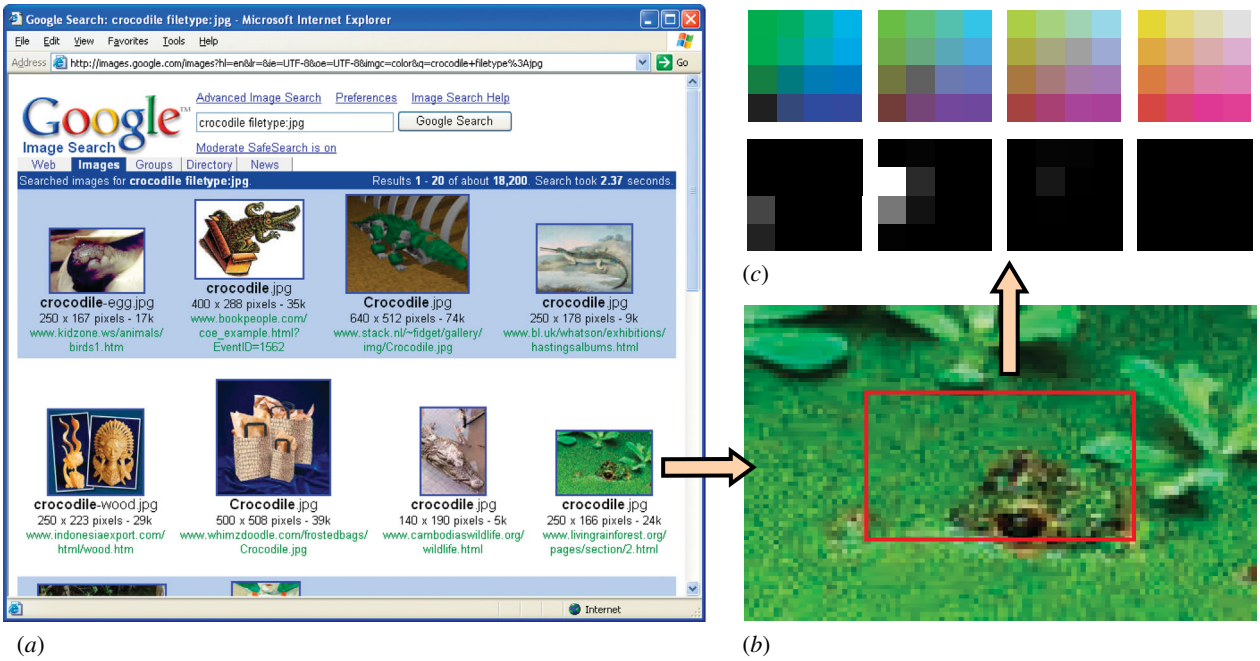

Figure 2. Shows the use of 'Google Image' to create the test data. (*a*) The search term 'crocodile' has been used. (*b*) The central quadrant of each image is then analysed to produce (*c*) a coarse-binned RGB histogram. The top row of (*c*) shows the 4^3 cells of the quantization of RGB that we use; the bottom row of (*c*) is laid out in the same manner, but shows the histogram, with intensity coding for frequency. (Image in (*a*) reproduced courtesy of Google Inc. Google™ is a trademark of Google Inc.)

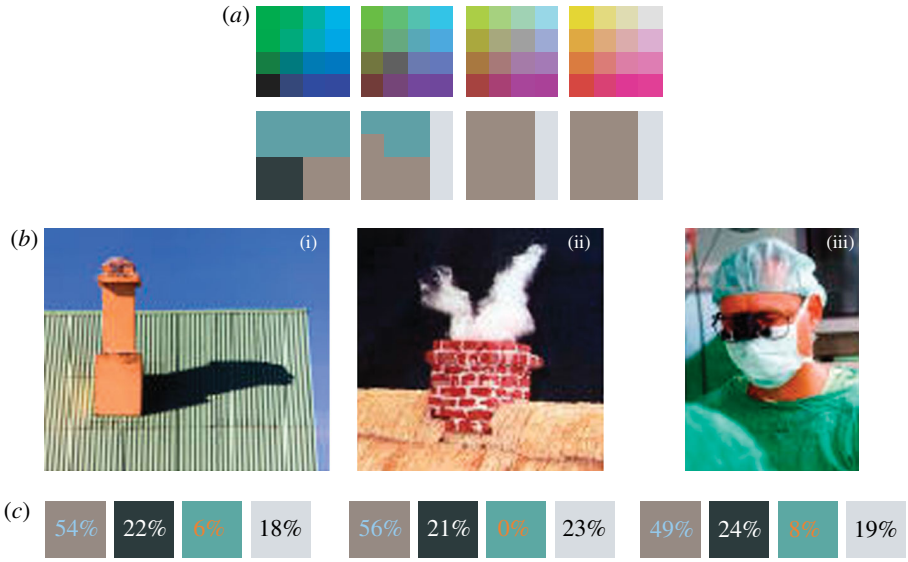

Figure 3. Illustration of the o–o–o task used in the study. (*a*) An example system of four categories. (*b*) Example instances from the classes 'chimney' (i, ii) and 'doctor' (iii). (*c*) Histograms of the central quadrants of the images using the categories from (*a*) as bins. In this particular case the distances between the histograms were computed as (i)–(ii) = 0.89, (i)–(iii) = 1.06 and (ii)–(iii) = 0.80. So, the o–o–o is incorrectly guessed to be (i); the correct answer being (iii).

'glaucous' categories then this useful inference would no longer be so easily made.

The various sources of explanation are not mutually exclusive. For example, categories could be tied to opponent channels and yet could also be particularly effective for classification. In such a case one could speculate that the pressure to classify has worked through evolution to shape the opponent channels, or effective classification could simply be an epiphenomenon to other evolutionary pressures. Disentangling these possibilities and assessing whether they are truly causative of the categories is likely to be difficult. Easier to assess is whether the categories are consistent or not with some possible explanation. For example, the hypothesis that basic colours are particularly stable under illuminant changes could be empirically assessed while remaining neutral on whether it explains them. We adopt this same neutral attitude when, in this paper, we investigate if the basic colour categories are particularly useful for classification: the most we can expect to do is change the plausibility of explanation-from-optimality.

1.3. Colour histograms in machine vision

The use of colour for content-indexing into a large database of images was introduced by Swain & Ballard (1991). They showed that using an image-similarity measure based on the similarity of the histograms of pixel colours, was robust, efficient and effective at retrieving from a database pictures that agreed in content with a query image. They concluded that colour histograms are object representations that are stable in the presence of occlusion and viewpoint changes, and that they can differentiate among a large number of objects.

Swain and Ballard used coarsely binned RGB histograms with 64 bins defined by the cross product of the two most-significant bits of each of the three colour channels. The motivations given for using coarse histograms are: less storage, faster indexing and helpful towards making the histogram-similarity measure accord well to human perception. This final point has been taken up by other authors who have constructed similarly coarse histograms, but defined on more perceptually uniform colour spaces such as HSV (Smith & Chang 1995), Luv (Sclaroff *et al.* 1997) and CIE-Lab (Sclaroff *et al.* 1997). The general idea of using these spaces is that, if histogram bins are such that all the individual colours within each bin are perceptually similar, then images whose colours differ in perceptually unimportant ways will automatically have similar histograms. The alternative is to use colour histograms with a fine binning structure and to incorporate information about colour similarity into the histogram comparison metric, but this is a far slower and more complex computation, and requires much more storage. More recently, several image-retrieval systems have taken this use of perceptually coherent colour histogram bins to the extreme of using just 11 bins based on the basic colours (Gong *et al.* 1996; Gagliardi & Schettini 1997; Ciocca & Schettini 1999; Ingemar 2000). Unfortunately, the use of colour histograms in these fulsome machine vision systems is only one of many factors that contribute to the computation of image similarity and so there is a lack of assessment of whether the use of such basic colour bins does actually improves retrieval performance compared to the other binning strategies previously used. The results reported in this paper can be regarded as filling this absence.

1.4. Hypothesis tested in this paper

We hypothesize that the basic colours are optimal colour categories for classification. We can make this more precise as follows.

OH: the colours of common things of the same general class are typically more *similar* than the colours of common things of different class. This effect is maximized when colours are *described* using the basic colours.

OH contains vague terms, in particular 'similar' and 'described'. These terms can either be made more precise by relating them to a particular algorithm or by appealing to some ideal algorithm that optimally extracts useful information from colour signals. In this paper, we follow the former course by testing ***OH*** for a particular algorithm that we implement in a machine vision setting. We claim that the particular algorithm used makes good use of colour data, but we make no claim to its optimality, nor to its similarity with the comparable algorithm used in human vision.

2. METHODS

To test ***OH*** we used the following task: given descriptions of the colours of three things, two of which are instances of the same class (for example two fishes) while the third is something else (for example a tree), decide which two of the descriptions are most similar, and so identify the remainder as the odd-one-out (o–o–o). ***OH*** predicts that the success rate at the o–o–o task will be maximized for colour descriptions based on the basic colours.

To test ***OH*** we need to decide on:

(i) a set of classes of thing,
(ii) a procedure for obtaining instances of each class,
(iii) a method of forming (relative to a system of colour categories) colour-based descriptions of instances and
(iv) a dissimilarity-measure on colour-based descriptions.

We describe our approach to these in the following sections.

2.1. Preparation of data

The ***OH*** specifies that the hypothesized optimality of the basic colours is particularly for the commonly encountered contents of the world. So, in an attempt to focus on common classes of thing, we used the 758 nouns (see appendix A) from a children's vocabulary book (Amery 1997; e.g. 'acrobats', 'baby', 'cabbage', etc.) as our classes.

As instances of our classes, we collected images using the web-based search engine 'Google Image' in response to a query using one of the class nouns (see figure 2). The searches used modifiers that specified that only colour jpeg-format images should be returned; this mostly eliminates results, which are web-page graphics (which are typically in .gif format) rather than images. After eliminating duplicates (approximately 1%) we used the first 80 images returned for each search. For convenience we used the thumbnail images that Google displays in its results page as our instances, rather than the indexed images themselves. These thumbnails varied in size between 60^2 and 140^2. Figure 2*a* shows typical results from a search.

For each instance, we computed and stored the RGB histogram of the central quadrant of the image. The logic of using only the central quadrant of each image was that, on average, the fraction of pixels closely related to the search term would be greater there than in the entire image. The RGB histograms were computed relative to a coarse 4^3 quantization of RGB; figure 2*b*,*c* illustrates the process.

For the purposes of statistical analysis, the class-defining nouns were randomly divided into two sets—A and B—each of 379 nouns; and each set of 80 instances was randomly divided into two sets—1 and 2—each of 40 images. Thus, we had four datasets: A_1, A_2, B_1 and B_2. How we make use of four smaller datasets rather than a single large one is described in detail later, but in brief they allow calculation of statistical measures that are not biased by over-fitting, they allow computation of confidence limits for our results, and they allow us to assess whether the number of instances we have per class is sufficient.

2.2. Systems of categories considered

The systems of categories that we considered were the partitions, compatible with a 4^3 quantization, of the RGB cube into two or more convex (Kim & Rosenfeld 1982) categories. The definition of convexity that we use is as follows: a *path* is any sequence of cells, with consecutive cells adjacent in a 26 neighbour sense. A *line* is a shortest *path* between its endpoints. A set of cells is *convex* if and only if every pair of cells in the set can be connected by a *line* entirely in the set.

2.3. Testing a system of categories

The first step in testing a system of categories was to compute category-based colour descriptions for all instances. These colour descriptions were simply the histograms using categories (i.e. collections of one or more cells) as bins rather than individual cells. Such histograms are easily computed from the stored 64-bin histograms by totalling the weights of cells belonging to the same category.

The next step was to subject all weights, of all category-based histograms, to a square-root transformation. This commonly used procedure in histogram processing makes the distribution (across histograms) of the weights associated with any particular bin less skewed, more normally distributed and so better behaved (Aherne *et al.* 1997). The final step was to whiten the square-rooted histograms so that the variance of the weights for different categories all equal unity. To achieve whitening, for each category we subtracted the mean and divided by the standard deviation of the distribution (across all histograms) of the square-rooted category-bin weights. We then treated each whitened square-rooted histogram as a point in n-dimensional space (where n is the number of categories) and used Euclidean distance as a measure of dissimilarity.

We used a Monte Carlo method (Manly 1997) to compute an o–o–o score for a category system. In each trial, two instances (i and ii) from one class and a third (iii) from a different class were randomly chosen. The three pair-wise dissimilarities (i–ii, i–iii, ii–iii) between the corresponding category-based histograms were calculated as above. The guessed o–o–o was the instance not involved in the smallest dissimilarity. When the guessed o–o–o was the instance which alone in the trial was from its class, the trial was rated a success, otherwise a failure (figure 3). Each o–o–o score was based on 10^6 trials, which resulted in a precision of $\pm 0.05\%$.

2.4. Defining a basic colour category system

To test ***OH*** we needed to identify a category system for the 4^3 quantized RGB-cube that closely approximated the partitioning of the colour solid into the basic colours. This was a multi-step process that we now describe.

We made use of an assignment of colour names to 267 Munsell-coordinate-specified chips (Kelly & Judd 1976). We will illustrate the following stages using one of these chips (5.3R, 5.9/3.5 labelled a 'light greyish red') as an example. To assign the 267 chips to the basic colours, we considered only the primary designator of the associated colour name ('red' in the example). Eleven 'violet' chips were classified as 'purples', and six 'olives' as 'greens'. The Munsell coordinates of the chips were transformed into XYZ under illuminant ($C_{XYZ} = \langle 88.2, 90.0, 107.3 \rangle$) using the Munsell Company's conversion software (v6.22); this gave $\text{example}_{XYZ} = \langle 32.0, 28.9, 29.0 \rangle$. Then, a von Kries transform (von Kries 1902) was used to transform to illuminant D65 ($\text{D65}_{XYZ} = \langle 95.0, 100.0, 108.9 \rangle$); this gave $\text{example}_{XYZ} = \langle 34.5, 32.1, 29.5 \rangle$. Then, CIE-Lab coordinates were computed; this gave $\text{example}_{\text{Lab}} = \langle 63.4, 14.4, 7.6 \rangle$. We then computed the convex hull (Chazelle 1993) of each of the 11 subsets of points to identify the CIE-Lab extents of the basic colours (figure 4*a*).

We then turned the colour extents in CIE-Lab space into data expressed over a uniform 32^3 sampling of RGB. This was done by transforming each of the 32^3 RGB triples into CIE-Lab space by assuming a monitor gamma of 2.4 and standard phosphor chromaticities (ITU 1990); and for each transformed triple detecting which, if any, of the basic colour extents it lay within. If it did lie within one, then the RGB triple got the corresponding basic-colour label, if not it was unlabelled. We then settled the unlabelled RGB triples by finding the closest basic-colour extent to the CIE-Lab image of the triple. Distance for this purpose was computed using the CIE94 colour metric (Griffin & Sepehri 2002). The result was a fully labelled cube as shown in figure 4*b* (right),*c*. Although, as we have explained, we mapped RGB grid points in CIE-Lab, not vice versa, for completeness we note that the example colour that we were following through the sequence of transformations ends up at $\text{example}_{\text{RGB}} = \langle 184, 144, 141 \rangle$.

The final step of our procedure was to consider the 512 basic-colour-labelled RGB-triples that corresponded to each cell of the 4^3 quantization. We could have simply seen which of the basic colours had the largest volume fraction for any given cell, but this would have produced an answer that corresponded poorly with expectation. Consider, for example, the cell R, G, B $\in$ [0,63], which is the one that we expected to receive the label 'black'. Although this cell contained 77 of the 78 RGB-triples that received the 'black' label, they accounted for only 15% of the cell's volume; whereas the largest volume fraction was 34% for 'green' RGB-triples. Hence, with the volume-fraction approach, no cell would have been labelled 'black'. So, instead of using raw volume-fractions, we computed the fraction of pixels of each label type that fell within the cell. Where the pixels considered were from the totality of all 60 640 images in

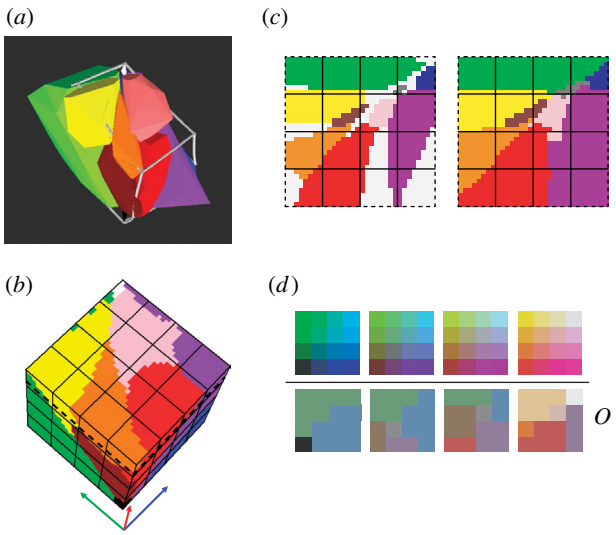

Figure 4. The stages in our mapping of the basic colours into RGB: (*a*) CIE-Lab space with regions of definite colour label (11 coloured polyhedra) and the edges of the monitor-typical RGB cube (grey). The orientation is similar to figure 1. (*b*) Both panels show the same slice through a 32^3 quantization of RGB with the basic colour extents from (*a*) mapped into it; in the right panel, the labels have been extended to all sites. (*c*) A perspective view of the completely labelled RGB cube in (*b*). The orientation is similar to (*a*), and the dotted line shows the position of the slice in (*b*). (*d*) Below the line is ***O***, the basic-colour category system we have defined, above the line is shown the same schematic of RGB as used for orientation in figures 2*c* and 3*a*. In the diagram of ***O***, and in diagrams of other category systems in other figures, we colour the categories with the average RGB value of pixels (from the full set of instances) that fall within them.

our study. So in the case of the cell discussed, 59% of image pixels that have RGB values within this cell were labelled 'black', 2% were 'red', 13% 'green', 6% 'blue', 8% 'purple' and 12% 'brown'. We labelled cells with the most common label that occurred within them, so the cell discussed was labelled 'black' as per expectation. The results of this procedure are shown in figure 4*d*. We will refer to this basic-colour category system as ***O***. Each of the categories of ***O*** is convex, though this was not enforced in its construction.

2.5. Finding optimal category systems

We wished to find the category systems that maximize the o–o–o score. The Monte Carlo method of estimating o–o–o scores is, however, ill-suited to finding highest-scoring category systems because its imprecision can be larger than the score difference between partitions that are being pair-wise compared. Instead we used a faster, but approximate, method of scoring based on measures of within- and between-class dispersion. The approximate scores were given by:

$$\text{score}(nr) = r\int_{w\in\mathbb{R}^+}\left(\chi^2_n(r\cdot w)\left(\int_{b\in[w,\infty]}\chi^2_n(b)\right)^2\right),$$

where n was the number of categories, χ^2_n is the probability density function of a chi-squared distribution of n degrees of freedom, r was the ratio between the within- and between-class root-mean-squared dissimilarities, and w and b are integration variables that range across different distances between histograms. We will refer to scores so calculated as equational, to distinguish them from Monte Carlo scores.

The equation was derived on the assumption that the histogram n-tuples of each class and of the entire set were each distributed like isotropic n-dimensional normal distributions. The square rooting and whitening that was done on raw histograms gave this assumption some grounds. A comparison study of the equational and Monte Carlo scores found a strong linear relationship (correlation coefficient, $r^2=0.76$) between the two, for category systems of a given size. However, it was found that the regression parameters of the relationship varied with the number of categories, so use of equational scores was restricted to comparisons between systems of the same number of categories. We used equational scoring on dataset A_1 to find the highest-scoring category system of a given size. In each case we started with a random category-system of the correct size and iteratively made changes to it, a single cell at a time. The change at each step was randomly chosen from the changes that (i) kept the number of categories constant, (ii) preserved category-convexity and (iii) improved the equational score. The process terminated when no valid change was possible. For each category size we repeated the optimization several times using a different random starting category system so that we could assess whether we were achieving convergence. From the systems that resulted from the multiple optimizations, the one that had the highest equational score was chosen. This system then had its Monte Carlo score assessed on datasets A_2, B_1 and B_2. Our use of different datasets for optimization and final evaluation was important as it prevented spuriously high scores due to over-fitting.

2.6. Computations carried out

The main computations that we carried out were to calculate o–o–o scores for random category systems, to find and score optimal category systems and to score the basic-colour system ***O***.

3. RESULTS

Figure 5 shows our results for random, optimal and the basic-colour systems of categories. We discuss these following the figure.

3.1. Random system scores

In figure 5, the grey line surrounded by the pale green band shows the mean and one sd of scatter of the o–o–o scores of random category systems. The data for each size of category system were based on 10 random systems each evaluated on the instance datasets B_1, A_2 and B_2. The results in figure 5 show that the mean o–o–o score for random systems rose from the baseline pure chance score of 33.3% for one category to a maximum of 36.6% for 20 categories, it then declined with further increases in the number of categories, dropping to

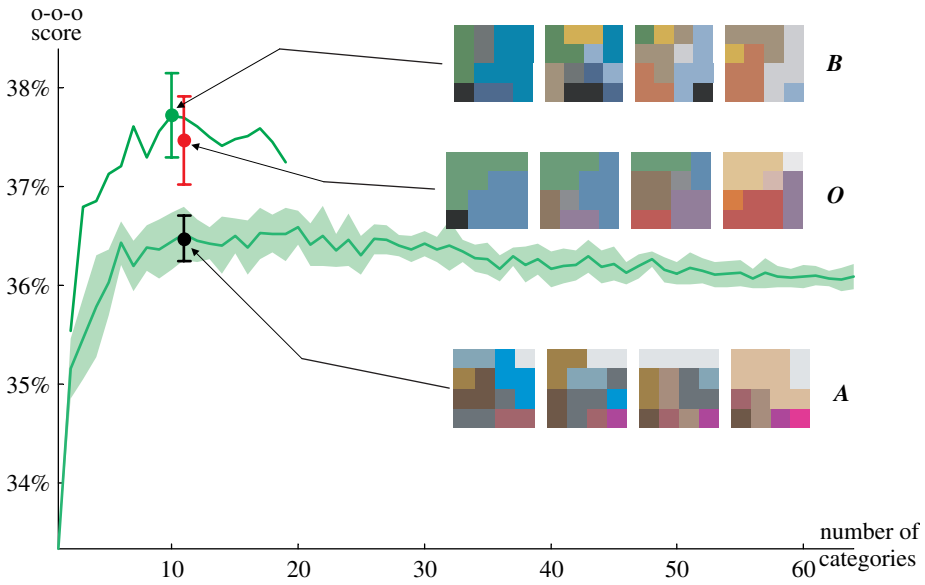

Figure 5. The main results of the study. The grey curve and surrounding pale green band show the mean score and one s.d. of variation for random category systems of different sizes. The black error bar shows the 95% confidence interval for the o–o–o score of a typical scoring 11-category partition (A) which is shown at the lower–right of the figure using the same format as figures 2*a* and 3*d*. The bright green curve shows the maximum likelihood estimates of the o–o–o scores of the optimal category-systems of size from 2 to 19. The bright green error bar shows the 95% confidence interval for the best optimal category system found (B) which is shown at the top-right of the figure. The red error bar is the 95% confidence interval for the basic-colour system ***O***, shown at the right of the figure.

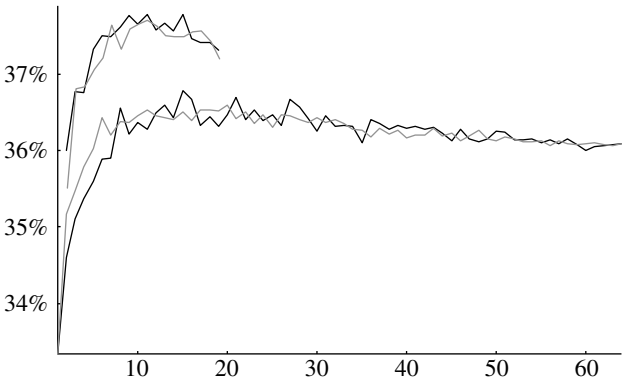

Figure 6. A comparison between system of categories that are convex (black) and unrestricted (grey). The lower curves are for random systems, the upper optimal.

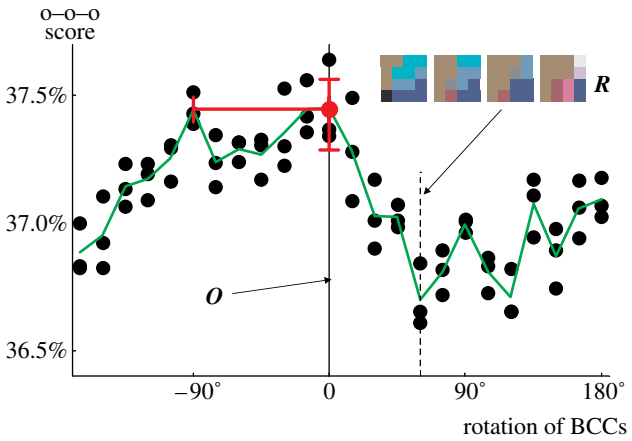

Figure 7. o–o–o scores for category systems derived from rotating the basic colours. Zero rotation is just the system ***O***. For each angle of rotation, the three black dots show the scores using the three test datasets A_2, B_1 and B_2. The green line connects the mean scores for each angle. The red symbol shows the median and 95% CI for the position and o–o–o score of the maximum of the mean score curve. The lowest scoring 'rotated basic-colour' system, which occurs at a rotation of 60°, is shown and labelled *R*.

36.1% for the unique partition of 64 categories. For all sizes of system greater than one, the mean o–o–o score is significantly greater than chance ($p<0.0005$).

In figure 5 we also include a typical random system of 11 categories which we refer to as ***A***. The estimated score for ***A*** is 36.5%, close to the average of 36.5% for random 11-category systems.

3.2. Optimal system scores

To assess whether our computation of optimal category-systems was achieving convergence we compared the maximum equational-score achieved based on either five or (a distinct) 10 optimizations. Finding no significant difference between these we concluded that five optimizations was sufficient; but even so the results in figure 5 are based on the highest equational scoring system of all 15 optimizations that we performed. The optimizations used A_1 but we then computed Monte Carlo scores based on B_1, A_2 and B_2; and it is the means of these scores that are plotted in the figure as a bright green line.

For systems of only two categories the optimal score was 35.5%, only slightly better than the mean random score for two-category systems (35.2%). The optimal score rises with the number of categories, peaking at systems of 10 categories, and then falls; eventually going down to the performance of the unique 64-category system. The optimal ten-category system is

denoted ***B*** and is shown in figure 5. The individual scores for ***B*** were 37.6, 37.9 and 37.6%. Thus, its maximum likelihood estimate is 37.7% with a 95% confidence interval [37.3%,38.2%]. The score of ***B*** and its confidence interval are shown in green in figure 5. The confidence intervals for other optimal systems are not shown but are of a similar size.

3.3. The basic-colour category system (*O*)

The o–o–o scores for the basic-colour system (***O***) were 37.3, 37.7 and 37.4% for the three test datasets. This gives a maximum likelihood estimate of 37.5% and a 95% confidence interval [37.0%,37.9%]. The score for ***O*** is significantly ($p<0.005$) larger than the scores for random systems of categories. A test of whether the estimated scores for ***O*** and ***B*** are the same does not reject the null hypothesis that they are ($p\approx0.08$).

3.4. Effect of dataset used

All our results are based on computing three o–o–o scores (one for each of A_2, B_1 and B_2) for each category-system considered. By using three datasets we get an estimate of the o–o–o scores averaged across a wider population of datasets than we have, and we get to bracket these estimates with confidence limits. We have compared the three o–o–o scores for a large number of category systems and find no significant difference in the variation of o–o–o scores that results from using different instances within the same classes (B_1 versus B_2) compared to using different classes (A_2 versus B_1, and A_2 versus B_2). This suggests to us that the number of classes we used, and the number of instances within those classes were well matched.

4. CONTROL COMPUTATIONS

We describe in the following six subsections the methods and results of control computations that we performed.

4.1. Importance of classes

Our first control computation was a check whether it was actually the grouping of instances into classes that underlay the greater than chance (33.3%) o–o–o scores we obtained, or simply some error in the design or execution of the computation. This we did by randomly shuffling the data so that instances were still in groups of 80 but now unrelated to the search terms used to index the classes. Using this shuffled data, the o–o–o scores of the three partitions shown in figure 5 were: ***B*** 33.3%, ***O*** 33.4%, ***A*** 33.3%. This confirmed the importance of the grouping of instances into classes.

4.2. Optimization starting at *O*

The optimization method used starts with a random category system. It is natural to wonder what will be found if ***O*** is taken as the starting point instead. We examined this by performing 15 optimizations starting at ***O***, and looking more closely at the best of these (***P***). The estimated o–o–o score for ***P***, when evaluated on our test datasets, was 37.7% with a 95% confidence interval [37.2%, 38.3%]. This is not significantly different from either ***B*** ($p\approx0.43$) or ***O*** ($p\approx0.06$), and thus nothing is lost by starting optimizations at random systems. It is tempting to look in detail at the category differences between ***O*** and ***P***, but since ***P*** failed to significantly improve on the score of ***O***, any differences can be attributed to over-enthusiastic fitting to the dataset A_1 used for optimization rather than a genuine worthwhile tweaking of the structure of ***O***.

4.3. Relaxing the category-convexity constraint

All of the results so far presented are for systems of convex categories. More relaxed conditions on categories could be imposed, for instance that they be connected but not necessarily convex. The most extreme relaxation is that the categories are unconstrained, and we have repeated our computation using such. The results in figure 6 show that it makes no important difference.

4.4. Rotated basic colours

We wished to assess whether the fine detail of the arrangement of the basic colours is important for their performance on the o–o–o task, or just their general layout. To this end we produced category systems that were like the basic colours in layout but different in detail. We did this by rotating the named Munsell colours that were the basis of our determination of ***O*** and then following the same steps as was done for ***O***. Rotation was about the achromatic axis and was performed in the CIE-Lab space as its approximate perceptual uniformity helps preserve the basic-colour structure during rotation. We computed 'rotated basic-colour' systems for angles of rotation from—165° to 180° in 15° degree steps. The o–o–o scores of these systems were then evaluated.

Figure 7 shows the individual scores of the 'rotated basic-colour' systems as black points. The green curve is the maximum likelihood estimate (i.e. the mean of the three scores). By eye measure the curve peaks at close to 0° rotation and more weakly at 180°, and has poorly defined minima around 60° and −165°. To make this more precise we used a bootstrap re-sampling of the data to estimate that, with 95% confidence, the curve peaks in the interval [−90°,0°], as shown in red. The figure also shows the '60°-rotated basic-colour' system, designated ***R***, which had the lowest score.

4.5. Natural versus manufactured classes

We next addressed the concern that our methods were biased towards a high score for ***O*** because many of our images were of manufactured items, and the choice of colours of these may be influenced by the usage of the basic colours by producers and consumers. In particular, there could be a tendency towards 'focal colours'.

To assess this, we ordered the 758 class terms by the degree to which they were natural or manufactured,

and bisected the list into manufactured and natural halves. The ordering (see appendix A) was based on 6600 binary judgements of relative naturalness between random pairs of terms made by a group of 33 naïve subjects.

An impression of the order is conveyed by listing every 20th term, starting at the natural end: shell, sky, ostrich, milk, squirrel, snail, carrot, crocodile, winter, tomato, sheepdog, reindeer, chef, wall, walking stick, chips, chicken, bridge, signpost, gloves, axe, road, dinner, hole, cricket, tape measure, knives, bedroom, switch, railings, loft, forks, ghost train, car wash, bicycle, umbrella, train set and tablecloth.

We would expect that if the o–o–o performance of ***O*** was enhanced by a manufactured-item focal-colour bias, then its o–o–o score using just manufactured classes would be higher than using just natural classes. This hypothesis was not supported. The Monte Carlo score for ***O***, using just the 50% of classes at the natural end of the spectrum, was 37.5%, and using just the manufactured classes was 37.2%.

4.6. Photographs versus pictures

The database of images used included many that would be better characterized as pictures than as photographs. These pictures, having been produced by human artists, may bear a trace of the cognitive colour categories of those artists e.g. categorically indistinct colours such as yellow–green might subconsciously be avoided. The presence of these cognitive traces could bias our results towards a high score for ***O***. To assess this we identified the image in each class which most contributed to the score of ***O*** being higher than for random category systems. If the picture-bias hypothesis is correct then we would expect that these instances would contain an excessive number of pictures. The fraction of pictures was assessed by two subjects, unaware of the purpose of the experiment. They rated that, of the full database of images, the fraction of pictures was 19.4% and 31.2%, the difference in the fractions reflecting different attitudes to difficult cases such as photographs of painted shop signs. Of the images most positively contributing to the performance of ***O*** the fractions of pictures were 17.7% and 31.9%, respectively. The fractions were not significantly different for either subject, which argues against the presence of pictures being an explanation of the performance of ***O***.

5. DISCUSSION

The results of our study are consistent with the hypothesis ***OH***. First, the o–o–o scores for random category-systems were significantly better than chance, demonstrating that colour is a useable cue for classification. This is consistent with studies showing the benefit of colour for object recognition (Ostergaard & Davidoff 1985; Wurm *et al.* 1993; Tanaka *et al.* 2001). Second, a basic-colour system of categories performed significantly better than random systems, demonstrating that (i) not all systems of categories are equally effective for classification and (ii) that a basic-colour system is one of the better ones. Third, we showed that we could not find any system of categories significantly better than the basic-colour system, which is consistent with our hypothesis that the basic colours are optimal. In the following sections, we tackle several aspects of the results and methods that threaten to undermine our conclusion.

5.1. Significant but small effects

Prior to our experiment there were several factors that made it doubtful whether better-than-chance performance at the o–o–o task would be achieved at all. In particular the highly variable nature of the images returned by 'Google Image', the pollution of the data with a substantial fraction of irrelevant images (for example, the crocodile-wood image in figure 2*a*), the absence of a foreground/background segmentation step, and the complete discard of spatial information all made the task difficult. However, the results of figure 5 show that performance significantly better than chance is achievable, though it must be stressed that although significant the margin is very small. In fact, the margin is so small that it is essential to rule out various possible explanations for it, other than the intended.

The class-shuffling control computation of §4.1, eliminated the possibility that the performance of ***O*** was due to an undiagnosed design flaw or bug in the computation and so completely artefactual.

The second possibility was that performance was due to a small subset of 'easy' classes. To assess this we computed class-specific o–o–o scores for the system ***O***. These scores, shown in figure 8, indicate that although there is a rightwards skew, it is mild; so the o–o–o score of ***O*** is not in the main due to a small number of easy classes.

5.2. What underlies differences in category system performance?

To further understand how o–o–o scores are achieved, we have looked at some particular classes in detail. In figure 9 we show a selection of instances from the 'lettuce' class. This is a noteworthy class as (i) it is the second best class for system ***O*** and (ii) it is the class whose score increases the most between systems ***A*** and ***O***. Figure 9 also shows diagrams illustrating the dispersion of lettuce instance-histograms for systems ***O*** and ***A***. These diagrams make it clear that the higher score of ***O*** compared to ***A*** is because the histograms for ***O*** significantly depart from the population mean along a dimension corresponding to the 'green' category. The prominence of pixels of various shades of green is apparent in the instances shown in the figure, and of course makes sense because lettuces are green. In contrast, ***A*** has several categories covering the green category of ***O***, so in ***A*** the consistency of the lettuce images is not captured. This is similar to the motivating example in §1.2 concerning the usefulness for induction of a 'green' category.

In figure 10, we compare ***B*** and ***O*** to understand how they achieve similarly high o–o–o scores despite being

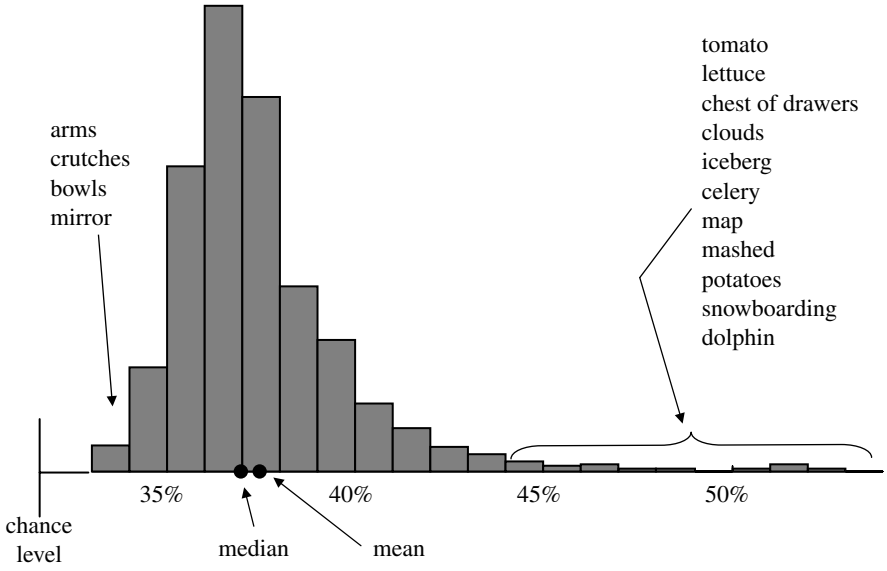

Figure 8. Shows a histogram of class-specific o–o–o scores for the basic-colour system ***O***. The 10 highest-scoring classes are shown at right ('tomato' is the best), the four lowest scoring at left ('mirror' is the worst).

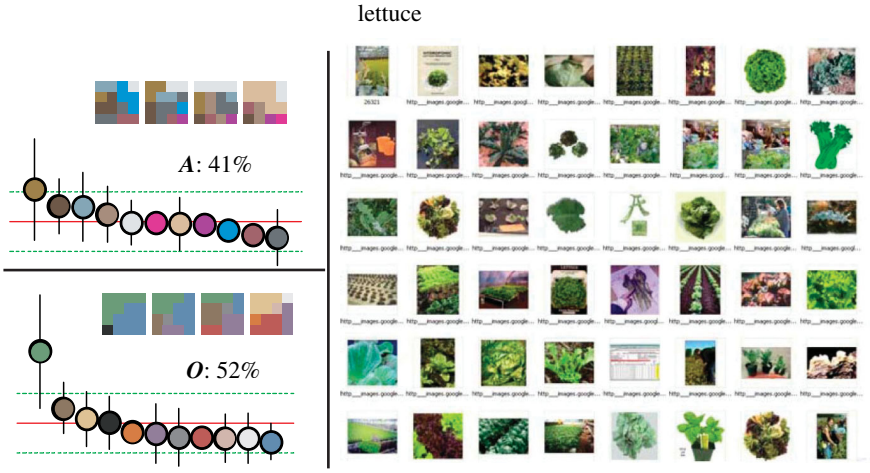

Figure 9. At right are typical instances of the 'lettuce' class. At left are shown the mean (coloured dots) and 1 s.d. of scatter (vertical line) of instance histograms for the class 'lettuce' when represented using the systems ***A*** and ***O***. Note that these are not raw histograms, but square-rooted whitened histograms. The background horizontal lines show the means (red) and ± 1 s.d. of variation (green) of the full population of instances. The colours of the dots are the same as in the category-system diagrams. The percentages are the class-specific scores for 'lettuce' with ***A*** and ***O***.

so different. The figure shows data concerning two classes: 'swimming pool', for which ***B*** most outperforms ***O***; and 'raspberry', for which ***O*** most outperforms ***B***. The better performance of ***B*** for 'swimming pool' is explained by its bluish and light-bluish categories, both of which are different from the population mean for this class, whereas ***O*** only has a single bluish category which is comparably deviant. For the 'raspberry' class, the better performance of ***O*** is because ***B*** lacks a reddish category and so fails to capture the similarity of these instances as well as ***O***. Together the 'swimming pool' and 'raspberry' examples illustrate that it is not a simple matter of there being a system of categories that is the best for all classes: best on average seems to be the most that can be expected.

Figure 11 shows the 'tomato' class which is the best class for ***O*** and the class for which the performance of ***O*** most outstrips that of ***R***. The ***O*** system successfully captures the fact that 'tomato' instances, although various in colour, are all varieties of red. In the ***R*** system, the red category of ***O*** has been dissected into purplish-pink, purplish-brown and brown categories and so the colour clustering of tomatoes is far less apparent.

5.3. Issues concerning the use of web imagery

Our use of images rather than objects is unusual in studies of human colour vision, but not unprecedented (Hurlbert 1998; Griffin 1999). However, the use of a database of imagery constructed by a search engine, rather than manually constructed and labelled is unprecedented. This raises the possibility that some non-standard aspects of the images we have used account for our results. We believe that the results in §5.2 already defuse this objection by showing that

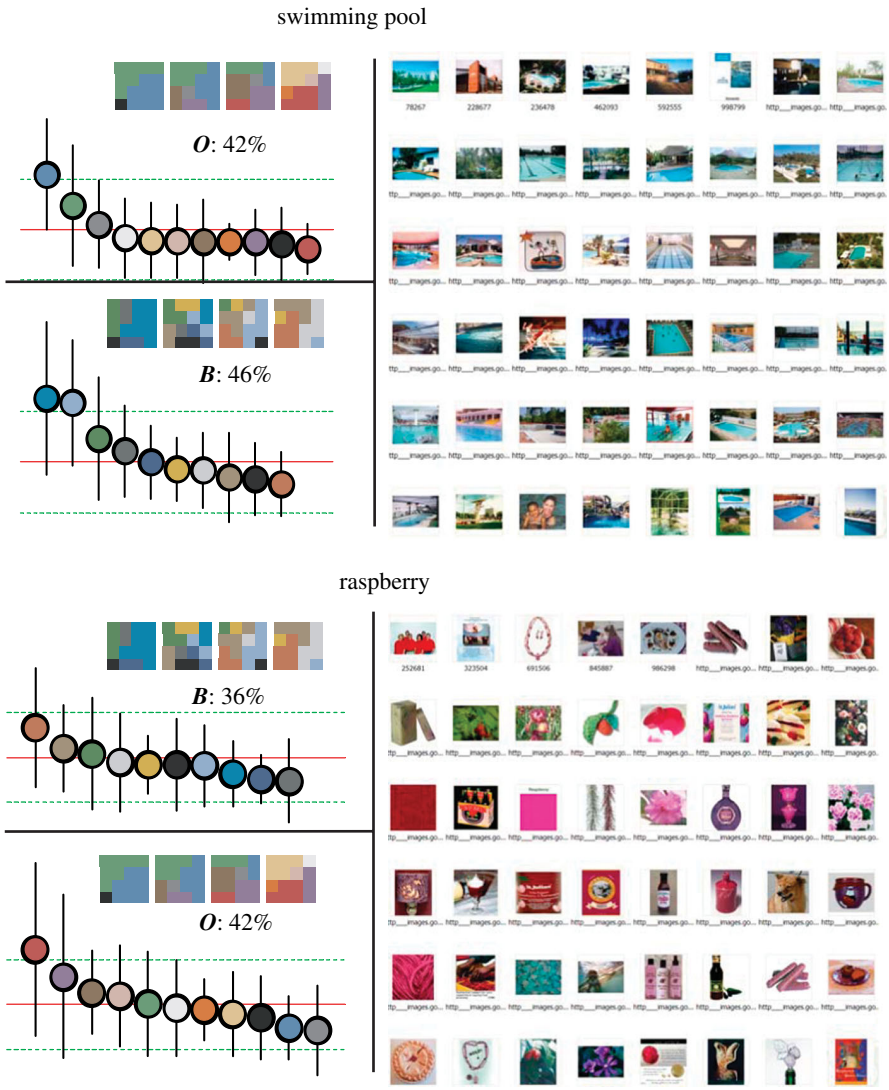

Figure 10. Laid out the same as figure 9.

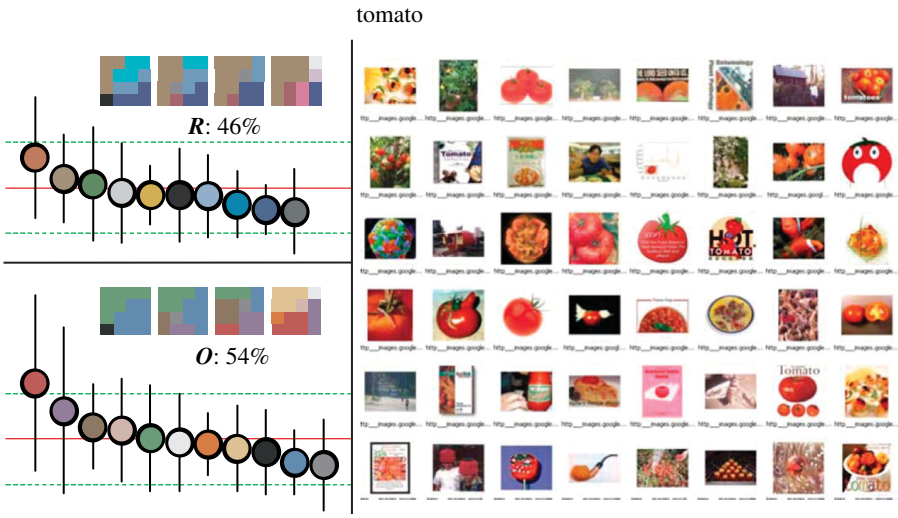

Figure 11. Laid out the same as figure 9.

categorization is based on colour in the intended manner, but in this section we deal with a few issues that merit direct address. These can be grouped into issues surrounding image content, the relation between the image and its content, and the image data itself.

The first image-content issue is that a substantial fraction (we estimate 10%) of the images we use, have contents that are not connected to the class term in a direct way. For example, in figure 2 we see an image of 'crocodile-wood' returned for the 'crocodile' query and

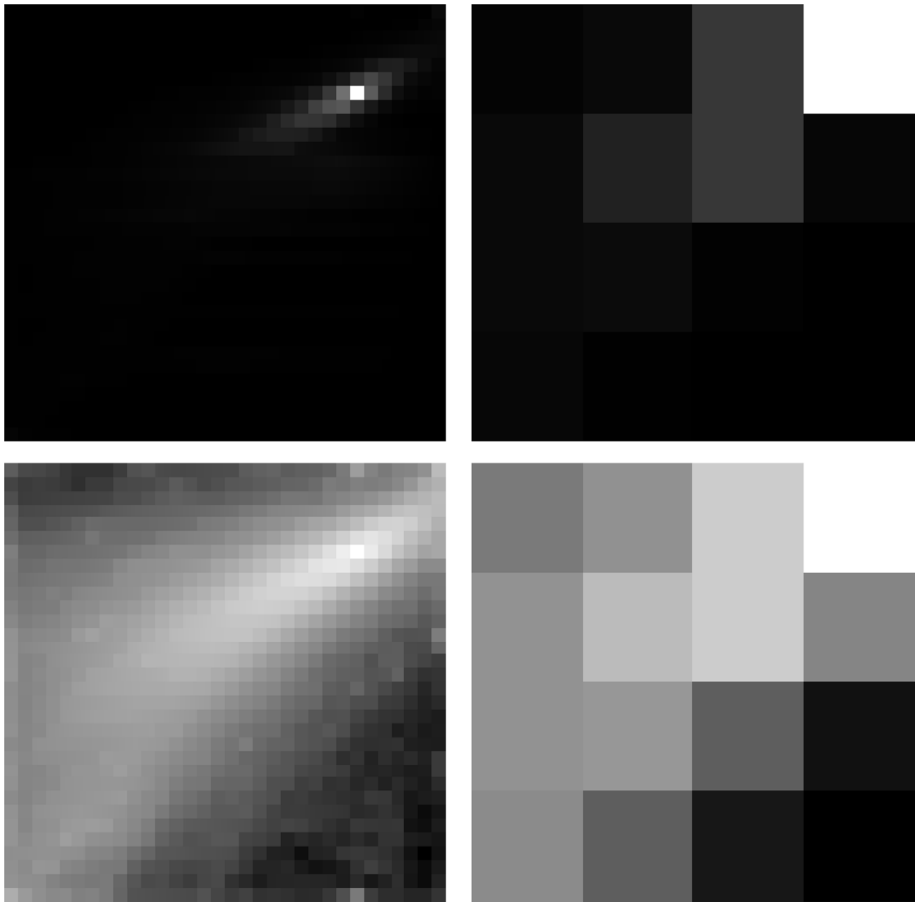

Figure 12. Shows slices through the RGB histogram for the full dataset. The left panels are for the histogram quantized to 32^3 bins (the same slice as in figure 4*b*). The right panels are for the top layer of the 4^3 quantization. In the top row, intensity codes directly for histogram weight; in the bottom row, intensity codes for log weight.

in figure 11 there is an image of a 'tomato hornworm moth' for the 'tomato' query. Since, these erroneous images are so unpredictable in their contents it is difficult to see how they could advantage ***O***. Similar comments apply to the issue of images that do contain directly class-related content, but not within the central quadrant. The second image-content issue is the possibility of a manufactured-good focal-colour bias. In §4.5, we showed evidence that this is not an issue. Informal scrutiny of the data suggests that any tendency to manufacture in focal colours is offset by a tendency to make products in diverse colours. The difference we found between natural and manufactured although small was consistent with the finding that colour is more useful for the recognition of natural than manufactured objects (Humphrey 1994).

Concerning the relation between the images and their content, one issue is that the images used are almost certainly non-veridical. The majority of them will have been acquired using uncalibrated cameras, with spectral sensitivities different from human, under a range of illuminants. They may well then have been manipulated to display well on a display device with gamma and chromaticities different from what we assume (§2.3) before being placed on the web. That this is a potentially serious problem can be seen by considering a population of images that were dramatically non-veridical. Imagine, for instance, that the RGB channels were permuted. This would undermine our claim that our results support ***OH*** and instead we would only be able to claim to have supported an hypothesis about the colours of web images of common things, rather than the colours of common things. However, acting against the possibility of significant non-veridicality is the fact that creators of web pages will favour images that allow viewers to correctly recognize their contents. A second issue on the relationship between images and their contents is the presence of pictures, as opposed to photographs, in the database. The issue with pictures is better described as their fictionality rather than their non-veridicality. As manufactured fictions they undoubtedly bear traces of the cognition of their creators. For example, painted seas might be blue more often than real seas. The control computation reported in §4.6 assessed whether the presence of pictures could account for the superior performance of the basic colour system ***O***, but found no evidence to support this.

Concerning the image-data itself, is the issue that widespread usage of compression algorithms may bias images towards containing certain RGB values. Examining the RGB histograms at full 256^3 quantization we do see evidence of this. However, as figure 12 shows, at the coarser quantization levels that we used these effects are lost and without influence.

The final issue with our use of web imagery concerns what is unknown about the operation of 'Google Image'. Google's description of its operation is 'Google

analyses the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content. Google also uses sophisticated algorithms to remove duplicates and ensure that the highest quality images are presented first in your results.' This is the only non-proprietary information available, so it is a possibility that some subtle and unknown bias has been introduced. However, this would also be the case with an image dataset constructed manually.

5.4. Issues around the use of RGB

In our method we had to choose the fineness of RGB quantizations used at two points. We choose 32^3 for representing the basic colour category extents from the named colour chip data, and 4^3 for representing category systems whose classification performance we assessed. It is the coarser of these two that is more plausibly the source of some problems. The possibility is that, although 4^3 allows a surprisingly good representation of the Basic colour categories (see figure 4), we cannot rule out that there is a category system that performs better than the basic colours but that can only be faithfully expressed at a quantization of 5^3 or higher.

Ignoring the slight non-veridicality of RGB (mentioned in §5.3), and so treating it simply as an alternative coordinate system for cone response space, there is an issue as to whether the direction and lengths of the RGB axes is influential on our results. This is a real possibility, since the category systems assessed were based on a regular subdivision based on the RGB system, and so the surprisingly good representation of the BCTs mentioned above could be due to the use of RGB, and the RGB system could be somehow tainted by the BCTs. However, against this concern it should be noted that the RGB system is not contrived around easy nameability of its axes (just as CMYK is also not). Rather the constraints on its design are that (i) it is a system for linear additive mixing of colour and (ii) its gamut should coincide with the gamut of surface colours as closely as possible. RGB succeeds in this second aim very well, as the gamut of surface colours is not far from being parallelepiped in form (Griffin 2001), which is inevitable given the non-convexity of the spectral locus (Koenderink & van Doorn 2003). So, in fact RGB is quite a natural coordinate system, and this may be why seven of the RGB cube's vertices coincide well with the foci of seven of the basic colours (black, white, red, green, blue, yellow and purple).

6. CONCLUSION

The starting point for this study was the observation that categorization of colour has both been *studied* as a window into human cognition, and has been *used* purely pragmatically by engineers building image-retrieval systems. This suggested the hypothesis that the best system of categories for pragmatic purposes coincides with human categories. We tested this hypothesis using a classification task and obtained results consistent with it. The strongest interpretation of our results is that they support an explanation from visual ecology for why humans use the colour categories that they do. However, because (i) our experimental method assessed the usefulness of colour categories when classifying image contents into noun-categories and (ii) noun-categories are human cultural-cognitive constructs, any ecological explanation that is supported would not be of a simple colour-clusters-present-in-the-environment type, rather it would have to be a far more complex beast dealing with the phenomenon of categorization in general. The first step in this grander project of explanation may already have been taken with the startlingly simple suggestion that 'good categories are those that support induction' (Ellison 2001).

However, we reject this strongest interpretation of our results for three reasons. First, our method tested the optimality hypothesis only for a particular machine vision algorithm, not for the unknown algorithm of human vision. Second, our method was only able to reveal a very weak advantage for the basic colours. Third, our experiment has absolutely nothing to say about any causal linkage between this advantage and humans having the basic colour categories; and causal linkage is required for explanation. Rather we feel that the most that can be concluded from this work is (i) the plausibility of an explanation of the basic colours as the result of a pressure-to-optimally classify is increased and (ii) the basic colours are good categories to use for classification in machine vision.

APPENDIX A

Here, we list the 758 nouns used as search terms. The terms are listed in the natural-to-manufactured order determined by the 6600 binary judgements of natural versus manufactured made between pairs of nouns. The precise ordering was not used in the reported research, only the segregation into a natural and a manufactured set.

A.1. The natural set

Shell, path, plant, cat, goat, kitten, lake, rocks, cliff, dew, bird's nest, cabbage, fog, island, mushroom, seeds, eagle, flowers, mist, night, sky, spider, star, tree, caterpillar, fish, lion, man, mountain, river, tiger, toad, worm, apricot, bone, bull, diamond, giraffe, leaves, lizard, ostrich, peas, polar bear, pond, rainbow, salt, sea, seaweed, summer, whale, autumn, baby, bush, camel, crab, frost, hair, hedgehog, kangaroo, lightning, milk, mole, moth, mouse, pebbles, snow, spinach, stones, trees, zebra, bear, cow, donkey, ducks, eye, forest, grapes, penguin, seaside, seasons, squirrel, tadpole, wood, beach, beans, birds, butterfly, cherry, clouds, cubs, ears, feathers, ladybird, monkey, mouth, parrot, peach, pigeon, rice, seal, snail, strawberry, apple, beehive, children, dolphin, elbow, farmer, gorilla, grapefruit, guinea pig, hippopotamus, iceberg, lemon, owl, rain, smoke, thumb, water, wolf, carrot, face, fisherman, hay, lambs, paws, pelican, plum, seagull, snake, tortoise, waves, weather, beaver, bison, canary, chalk, cheek, chocolate, cockerel, crocodile, ducklings, elephant, field, geese, hens, lettuce, melon, nuts, panda, pepper, pineapple, pony, raspberry,

rhinoceros, straw, stream, sun, vegetables, wing, winter, badger, celery, chin, cucumber, dancers, deer, families, girl, grass, hand, haystack, hill, horns, leek, lips, pumpkin, shepherdess, shoulders, starfish, tomato, tongue, waterfall, wind, banana, bat, beaker, boy, calf, clementine, hamster, horse, light, logs, moon, mud, neck, planet, shark, sheep, sheepdog, arm, box, cheese, cobweb, cone, dog, duck, eggs, fried egg, frog, head, hedge, honey, leopard, nails, people, potatoes, puppy, rabbit, reindeer, sandwich, space, spring, sticks, tea, wasp, bride, brush, budgerigar, cauliflower, fox, fruit, knee, leg, toes, toilet paper, woman, artist, bucket, chef, chicks, clothes, country, cream, farm, fireman, foot, meat, onion, orchard, pets, pigs, plank, rope, soap, sponge, stable, straw bales, tummy, wall, beads, candle, circle, eyebrow, fence, fingers, flippers, flower bed, garden, iron, omelette, pear, salad, sandpit, singers, soup, sugar, supper, swans, walking stick, acrobats, bottles, broom, drink, food, hoe, lamp, mat, oval, plates, plough, rectangle, sauce, sawdust, tail, teeth, towel, bonfire, bow, chips, cushion, fishing boat, home, house, jam, kennel, mashed potatoes, masks, oar, pet, plaster, ring, sailor, steps, stool, tightrope walker, ball, butcher, canoe, chicken, cowshed, floor, fruit juice, goldfish, nose, popcorn, rolls, rowing boat, rubbish, school, shower, skirt, toast, tourists, trowel, village, basket, bench, breakfast, bridge, cap, drawing, duster, face paints, flour, ham, hot-air balloon, jars, paint, pancakes, pegs, piglets, present, ruler, sandcastle, scales, scarf, shapes, shavings, signpost, sink, spade, square, string, teacher, train driver, trunk, waiter, aquarium, ceiling, chair, crayons, crescent, crossing, dress, dustpan, farmhouse and football.

A.2. The manufactured set

Globe, gloves, hammer, juggler, jumper, painter, park, pen, picnic, pigsty, pole, rake, roller, roundabout, saddle, shoes, target, waitress, yoghurt, alphabet, arrows, axe, band, bath, bolts, boot, bowls, candy floss, coat, comb, crutches, drums, fishing rod, hat, hot chocolate, letters, marbles, market, paddle, paper, policewoman, road, roof, rope ladder, saucepans, shed, skipping rope, sport, toilet, balloon, bed, belt, boat, bricks, cage, carrier bag, cart, chimney, coffee, cupboard, curtain, dinner, duvet, greenhouse, ladder, lunch, ribbon, rubber, sandals, signals, skip, tents, whistle, barge, buttons, canal, cards, cooker, crisps, door, glasses, hole, jeans, judo, kite, medicine, nurse, petrol, pillow, playground, salami, teddy bear, tightrope, triangle, trousers, apron, barrel, bathroom, birthday cake, changing room, Christmas tree, cricket, cups, deck chair, door handle, father Christmas, frying pan, hoop, jacket, mirror, nappy, photographer, pilot, platform, pockets, pudding, rifle range, scarecrow, shorts, sunhat, swings, tape measure, teapot, teaspoons, tins, tissues, tray, boots, calendar, clown, comic, cube, desk, doctor, easel, fireworks, gate, hall, handkerchief, kettle, kitchen, knives, living room, money, necklace, newspaper, nightdress, party, pavement, pictures pizza, postman, pyjamas, saw, shop, slippers, soldiers, tunnel, vice, zip, barn, bedroom, bridegroom, button holes, cardigan, crane, flats, fork, guitar, hotel, judge, lock, pencil, pills, rocking horse, runway, sausage, screwdriver, sheet, shoelace, spoons, switch, table, thermometer, tights, tool box, waiting room, washbasin, washing powder, watering can, window, badge, bonnet, bulb, glue, hamburger, hen house, karate, map, mop, pram, railings, sailing boat, shirt, ski pole, socks, stairs, sumo wrestling, swimsuit, tie, trapeze, trolley, wheel, books, buffers, dice, dustbin, flag, handbag, helter-skelter, ice skates, loft, pipes, presents, puppets, ring master, scissors, screws, ship, slide, sweatshirt, tea towel, toothpaste, tractor, trailer, upstairs, windmill, yacht, downstairs, drill, file, forks, jack-in-the-box, matches, paint pot, paints, plane, safety net, spaghetti, street, sweet, tennis, tiles, top hat, café, carpet, castle, cinema, circus, dressing gown, fancy dress, ghost train, helicopter, helmet, locker, money box, purse, racket, robot, spaceship, table tennis, toothbrush, toyshop, trainers, water–skier, wedding day, wheelchair, workshop, backpack, big dipper, camera, car wash, digger, factory, hairdresser, jigsaw, lamp post, parachute, radio, sandpaper, saucers, snowboarding, sofa, sprinkler, syringe, ticket machine, tricycle, vacuum cleaner, wheelbarrow, airport, basketball, bicycle, big wheel, bus, garage, gym, motor–boat, notebook, penknife, sleigh, suitcase, swimming pool, television, toys, workbench, dodgems, fire engine, goods train, lawn mower, racing car, roller blades, umbrella, video, wardrobe, washing machine, cassette tape, checkout, dentist, engine, fridge, headlights, key, oil tanker, photographs, radiator, recorder, spacemen, tanker, tap, telephone, train, train set, video camera, caravan, chairlift, chest of drawers, dolls, drawer, ironing board, lift, piano, spanner, taxi, tyre, van, ambulance, american football, badminton, birthday card, bus driver, carriages, tablecloth, telescope, compact disc, railway station, trumpet, bow tie, control tower, doll's house, fairground, hospital, lighthouse, police car, lorry, rocket, battery, submarine, railway track, traffic lights.

REFERENCES

Aherne, F. J., Thacker, N. A. & Rockett, P. I. 1997 The Bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika* **32**, 1–7.

Amery, H. 1997 *First 1000 words sticker book*. London: Usborne Publishing Ltd.

Berlin, B. & Kay, P. 1969 *Basic color terms: their universality and evolution*. Berkeley: University of California Press.

Brown, D. E. 1991 *Human universals*. New York: McGraw-Hill.

Buchsbaum, G. & Bloch, O. 2002 Color categories revealed by non-negative matrix factorization of Munsell color spectra. *Vis. Res.* **42**, 559–563. (doi:10.1016/S0042-6989(01)00303-0)

Chazelle, B. 1993 An optimal convex hull algorithm in any fixed dimension. *Discrete & Computational Geometry* **10**, 377–409.

Ciocca, G. & Schettini, R. 1999 A relevance feedback mechanism for content-based image retrieval. *Inf. Process. Manag.* **32**, 1685–1695.

Davidoff, J., Davies, I. & Roberson, D. 1999 Colour categories in a stone-age tribe. *Nature* **398**, 203–204. (doi:10.1038/18335)

Dowman, M. 2002 Modelling the acquisition of colour words, in Al 2002. *Adv. Artif. Intell.*, 259–271.

D'Zmura, M. & Knoblauch, K. 1998 Spectral bandwidths for the detection of color. *Vis. Res.* **38**, 3117–3128. (doi:10.1016/S0042-6989(97)00381-7)

Ellison, T. M. 2001 Induction and inherent similarity. In *Similarity and categorization* (ed. U. Hahn & M. Ramscar), pp. 29–49. Oxford: OUP.

Gagliardi, I. & Schettini, R. 1997 A method for the automatic indexing of colour images for effective image retrieval. *N. Rev. Hypermedia Multimedia* **3**, 201–224.

Gärdenfors, P. 2000 *Conceptual spaces: the geometry of thought*. Cambridge, MA: MIT Press.

Gibson, J. J. 1979 *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gong, Y. H., Chuan, C. H. & Guo, X. Y. 1996 Image indexing and retrieval based on color histograms. *Multimedia Tools Appl.* **2**, 133–156.

Griffin, L. D. 1999 Partitive mixing of images: a tool for investigating pictorial perception. *J. Opt. Soc. Am. A* **16**, 2825–2835.

Griffin, L. D. 2001 Similarity of pyschological and physical colour space shown by symmetry analysis. *Color Res. Appl.* **26**, 151–157. (doi:10.1002/1520-6378(200104)26:2<151::AID-COL1006>3.0.CO;2-G)

Griffin, L. D. & Sepehri, A. 2002 Performance of CIE94 for non-reference conditions. *Color Res. Appl.* **27**, 108–115. (doi:10.1002/col.10029)

Hering, E. 1920 *Outlines of a theory of the light sense*. Harvard: Harvard University Press.

Humphrey, G. K. 1994 The role of surface information in object recognition: studies of a visual form agnosic and normal subjects. *Perception* **23**, 1457–1481.

Hurlbert, A. 1998 Illusions and reality checking on the small screen. *Perception* **27**, 633–636.

Hurvich, L. M. & Jameson, D. 1957 An opponent-process theory of color vision. *Psychol. Rev.* **64**, 384–404.

Ingemar, J. C. *et al.* 2000 The Bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments. *IEEE Trans. Image Process.* **9**, 20–37. (doi:10.1109/83.817596)

ITU 1990 *Parameter values for the HDTV standards for production and international programme exchange, ITU-R BT.709-3*. Geneva: International Telecommunications Union.

Jameson, K. A. In press. Culture and cognition: what is universal about color experience? *J. Cogn. Culture*.

Kay, P. 1999 Color. *J. Linguistic Anthropol.* **1**, 29–32.

Kay, P. & Maffi, L. 1999 Color appearance and the emergence and evolution of basic color lexicons. *Am. Anthropol.* **101**, 743–760. (doi:10.1525/aa.1999.101.4.743)

Kay, P. & Regier, T. 2003 Resolving the question of color naming universals. *Proc. Natl Acad. Sci. USA* **100**, 9085–9089. (doi:10.1073/pnas.1532837100)

Kay, P. *et al.* 2005 The World Color Survey. Center for the Study of Language and Information.

Kelly, K. L. & Judd, D. B. 1976 Color: universal language and dictionary of names. National Bureau of Standards 189.

Kim, C. E. & Rosenfeld, A. 1982 Digital straight lines and convexity of digital regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **4**, 149–153.

Koenderink, J. J. & van Doorn, A. J. 2003 Perspectives on color space. In *Colour perception: mind and the physical world* (ed. R. Mausfield & D. Heyer), pp. 1–56. Oxford: Oxford University Press.

Kuehni, R. G. 2004 Variability in unique hue selection: a surprising phenomenon. *Color Res. Appl.* **29**, 158–162. (doi:10.1002/col.10237)

Lin, H. *et al.* 2001 A cross-cultural colour-naming study. Part I. Using an unconstrained method. *Color Res. Appl.* **26**, 40–60. (doi:10.1002/1520-6378(200102)26:1<40::AID-COL5>3.0.CO;2-X)

Lucy, J. A. 1997 The linguistics of color. In *Color categories in thought and language* (ed. C. L. Hardin & L. Maffi), pp. 320–346. Cambridge: Cambridge University Press.

Manly, B. F. J. 1997 *Randomization, bootstrap and monte carlo methods in biology*. London: Chapman & Hall.

Ostergaard, A. L. & Davidoff, L. B. 1985 Some effects of colour on naming and recognition of objects. *J. Exp. Psychol.* **11**, 579–587.

Paramei, G. V. 2005 Singing the Russian blues: an argument for culturally basic color terms. *Cross-Cult. Res.* **39**, 10–38.

Riemann, A. 1854 *On the hypotheses which lie at the foundations of geometry*. A Source Book of Mathematics. New York: Dover.

Roberson, D. 2005 Color categories are culturally diverse in cognition as well as in language. *Cross-Cult. Res.* **39**, 56–71. (doi:10.1177/1069397104267890)

Roberson, D., Davies, I. & Davidoff, J. 2000 Color categories are not universal: Replications and new evidence from a stone-age culture. *J. Exp. Psychol.-Gen.* **129**, 369–398. (doi:10.1037//0096-3445.129.3.369)

Saunders, B. 2000 Revisiting basic color terms. *J. R. Anthropol. Inst.* **6**, 81–99.

Saunders, B. & van Brakel, J. 1997 Are there non-trivial constraints on color categorization? *Behav. Brain Sci.* **20**, 167–228. (doi:10.1017/S0140525X97531426)

Schirillo, J. 2001 Tutorial on the importance of color in language and culture. *Color Res. Appl.* **26**, 179–192. (doi:10.1002/col.1016)

Sclaroff, S., Taycher, L. & La Cascia, M. 1997 Image-Rover: a content-based image browser for the world wide web. In *Proc. IEEE Workshop on Content-based Access Image and Video Libraries*, pp. 2–9. Los Alamitos: IEEE Computer Press.

Smith, J. R. & Chang, S.-F. 1995 Single color extraction and image query. *Proc. ICIP*, vol. 3, pp. 528–531. Los Alamitos: IEEE Computer Press.

Steels, L. & Belpaeme, T. In press. Coordinating perceptually grounded categories through language. A case study for colour. *Behav. Brain Sci.*

Swain, M. J. & Ballard, D. H. 1991 Color indexing. *Int. J. Comput. Vis.* **7**, 11–32. (doi:10.1007/BF00130487)

Tanaka, J., Weiskopf, D. & Williams, P. 2001 The role of color in high-level vision. *Trends Cogn. Sci.* **5**, 211–215. (doi:10.1016/S1364-6613(00)01626-0)

von Kries, J. 1902 Chromatic adaptation. In *Sources of colour vision* (ed. D. L. MacAdam), pp. 109–119. Cambridge, MA: MIT Press.

Webster, M. A. 2000 Variations in normal color vision. II. Unique hues. *J. Opt. Soc. Am. A* **17**, 1545–1555.

Webster, M. A. & Mollon, J. D. 1994 The influence of contrast adaptation on color appearance. *Vis. Res.* **34**, 1993–2020. (doi:10.1016/0042-6989(94)90028-0)

Wurm, L. H. *et al.* 1993 Color improves object recognition in normal and low vision. *J. Exp. Psychol.: Hum. Percept. Perform.* **19**, 899–911. (doi:10.1037//0096-1523.19.4.899)

Yendrikhovskij, S. N. 2001 Computing color categories from statistics of natural images. *J. Imaging Sci. Technol.* **45**, 409–417.