Scene Recognition by Hyperspectral Ratio Indexing: How Many Channels Are Necessary?

Nsikak Ekpenyong and David H Foster; School of Electrical & Electronic Engineering, University of Manchester; Manchester, UK

Abstract

The problem of object or scene recognition is often addressed by seeking geometric image properties that are invariant under changes in viewing conditions. An alternative, non-geometric, ratio method was described by Funt and Finlayson (IEEE Trans. Pattern Anal. Mach. Intell. 17,522, 1995) in which histograms of spatial ratios of colour RGB triplets from neighbouring image regions were used to recognise objects under changes in viewpoint and illumination. In this study, ratio indexing was extended from RGB images to hyperspectral images with a variable number of sensor channels distributed over 400-720 nm. Fifty natural scenes were used to generate test and reference images. For each number of sensors, independent random samples were drawn from each test image of a scene under either a daylight of correlated colour temperature of 25000 K or of 4000 K and matched against independent random samples drawn from each reference image of the scenes under a daylight of correlated colour temperature 6500 K. Matching was based on the intersection of multi-dimensional histograms of ratios of sensor signals in these samples Differences between match hit and false-alarm rates provided a measure of recognition performance. Results suggest that for small samples, indexing with five sensor channels has advantages over indexing with three sensor channels for the recognition of natural scenes.

Introduction

Object recognition methods are predominantly based on geometric image properties that, in principle, are invariant under changes in viewpoint. By contrast, approaches to recognition based on colorimetric properties depend little on viewpoint. One such method—colour indexing—was developed by Swain and Ballard [1], who used colour histograms and histogram intersection to determine matches between test and reference images obtained under different viewing conditions. The colour axes used for the histograms were opponent and non-opponent combinations of the red, green, and blue components of the triplets (r, g, b) at each point. The method was generally robust to variations in viewpoint and scene background, but had limited invariance to changes in illumination, as the red, green, and blue components were simply normalized by their sum.

Funt and Finalyson [2] improved the illumination invariance of colour indexing by replacing the red, green, and blue components of the triplet (r, g, b) at a point by the corresponding triplet of spatial ratios defined across adjacent points; that is, (r_1/r_2, g_1/g_2, b_1/b_2) for points 1 and 2 (they actually used a Laplacian or first directional derivatives of the logarithm of the colours). Such spatial ratios are relatively stable under changes in illumination [3], although not exactly invariant. Funt and Finalyson [2] noted that if the sensor spectral sensitivities were broad, as with the cone photoreceptors of the eye, then indexing performance was worse, but by transforming spectral sensitivities so that they were spectrally narrower or sharper [4], almost perfect indexing performance could be obtained with their test and reference images. These were Mondrian-like, abstract coloured patterns under different illuminations. Somewhat lower performance was obtained with images of real objects [2].

Whether spectral sensitivities are broad or narrow, however, there is a more general problem with using three sensor spectral sensitivities, in that according to some behavioural measures [5, 6], reliable surface identification by spectral sampling requires more than three degrees of freedom, in particular with natural scenes of the kind illustrated in Fig. 1.

In principle, increasing the number of sensor classes over the available wavelength range should increase the reliability of the colour signal by reducing the number of false matches, and therefore produce better recognition performance. On the other hand, more sensor classes might reduce the number of correct matches and increase the level of noise.

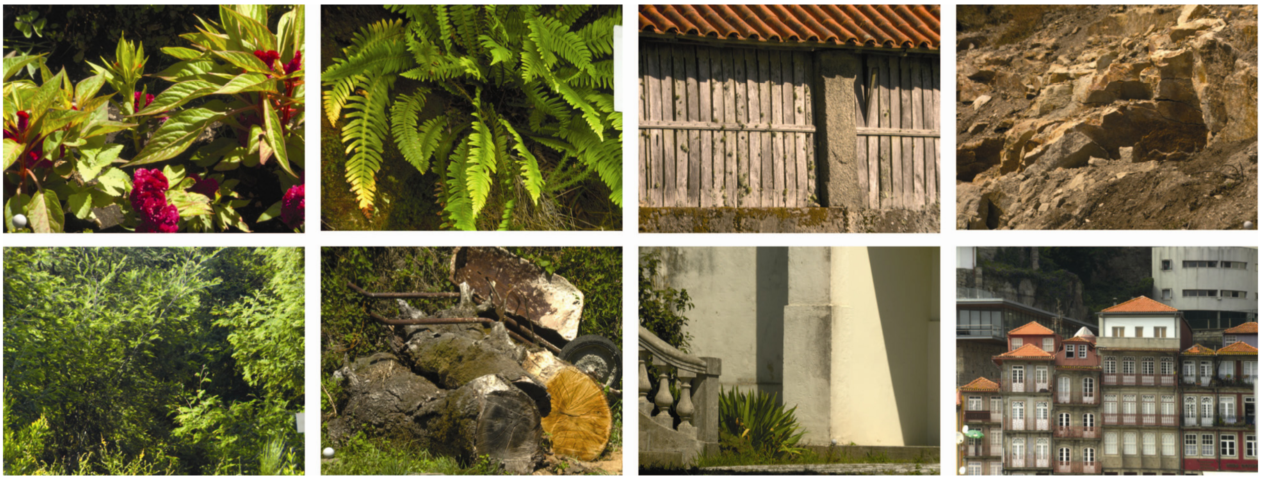

***Figure 1**. Eight of the 50 natural scenes used in this study (adapted from Fig. 2 of Ref. [8]).*

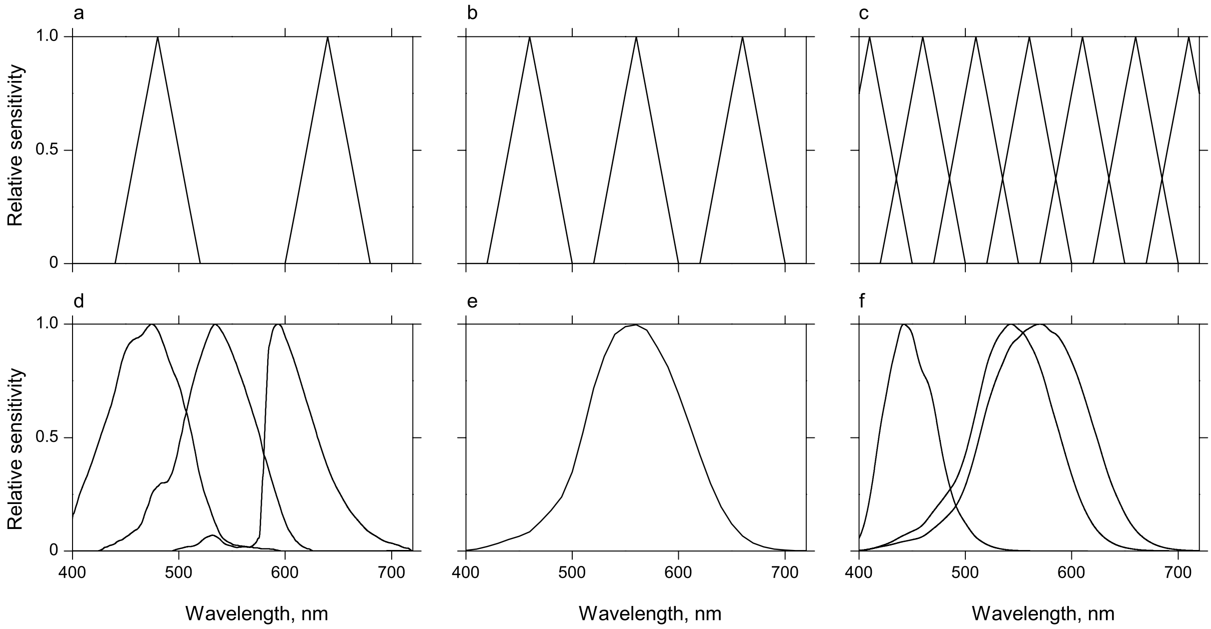

Figure 2. *Examples of spectral sensitivities of simulated variable-channel system with (a) 2, (b) 3, and (c) 7 sensor channels, and of the spectral sensitivities of (d) a Nikon D1 camera, (e) the CIE photopic luminance function, and (f) the spectral sensitivities of the cone fundamentals.*

The aim of this work was to determine how many sensor channels are needed for the reliable recognition of scenes under different illuminations when test and reference images are sparsely and independently sampled. Sparse independent sampling was used to capture the spatial uncertainties that, under other imaging conditions, could arise by occlusion or change in viewpoint.

The analysis was based on 50 natural scenes, represented as hyperspectral images. Each scene was simulated under daylight illuminants with different correlated colour temperatures (CCTs). Unlike the procedure adopted in [2], where spatial ratios were drawn from neighbouring points in the scene, spatial ratios were here obtained by taking signals from pairs of points chosen at random across the scene. It was found that with small sample sizes, recognition performance increased with the number of sensor classes, but reached a maximum with five classes.

Methods

Scenes

Eight of the 50 natural scenes are shown in Fig. 1 and thumbnail illustrations of some of the larger set are available in Ref. [7]. Details of how the hyperspectral data were obtained and of their accuracy are given in Ref [8]. Each hyperspectral image had spatial dimensions ≤ 1344 × 1024 pixels and spectral range 400–720 nm sampled at 10-nm intervals, thereby providing a discrete representation of an effective spectral reflectance $R(\lambda; x, y)$ at each wavelength λ and position (x, y) in the scene. The effect of illuminating the scene by a particular illuminant with spectrum $E(\lambda)$ was simulated by multiplying $R(\lambda; x, y)$ at each point (x, y) by $E(\lambda)$. The assumptions and approximations involved in this approach have been discussed in Ref [8, Appendix A]. Because of the approximately 1.3-pixel line spread function of the camera system used to acquire the hyperspectral data [8], only non-adjacent pixels were spatially sampled.

Daylight spectra were simulated from those described by the CIE [9] with CCTs of 4000 K, 6500 K, and 25000 K, characteristic of the sun and sky at different times of the day.

Spectral Sampling

A sensor system with a variable number *n* of sensor channels was simulated by taking the average bandwidth of a commercial RGB sensor (a Nikon D1 digital camera [10]), and then replicating a triangular spectral sensitivity with this bandwidth at evenly spaced points over the visible spectrum, as illustrated in Fig. 2 for three examples, with (a) two, (b) three, and (c) seven sensor channels. The maximum number of sensor channels possible in the present simulation was limited to seven, owing to limits on computer calculations with histograms of more than seven dimensions. No attempt was made in this analysis to optimize the spectral locations of the sensors according the characteristics of the scene being sampled. For comparison, the sensors of (d) a Nikon D1 camera, (e) the CIE photopic luminance function [9], and (f) the spectral sensitivities of the cone photoreceptors, i.e. the cone fundamentals [9], were also used.

Spatial Sampling

Spatially random samples of size $N = 10$ and $N = 100$ points were taken from images of scenes under a daylight of CCT 4000 K or 25000 K to act as test sets and from images of scenes under a daylight of CCT 6500 K to act as the reference set. Critically, the spatial samples in the test and reference sets were drawn independently.

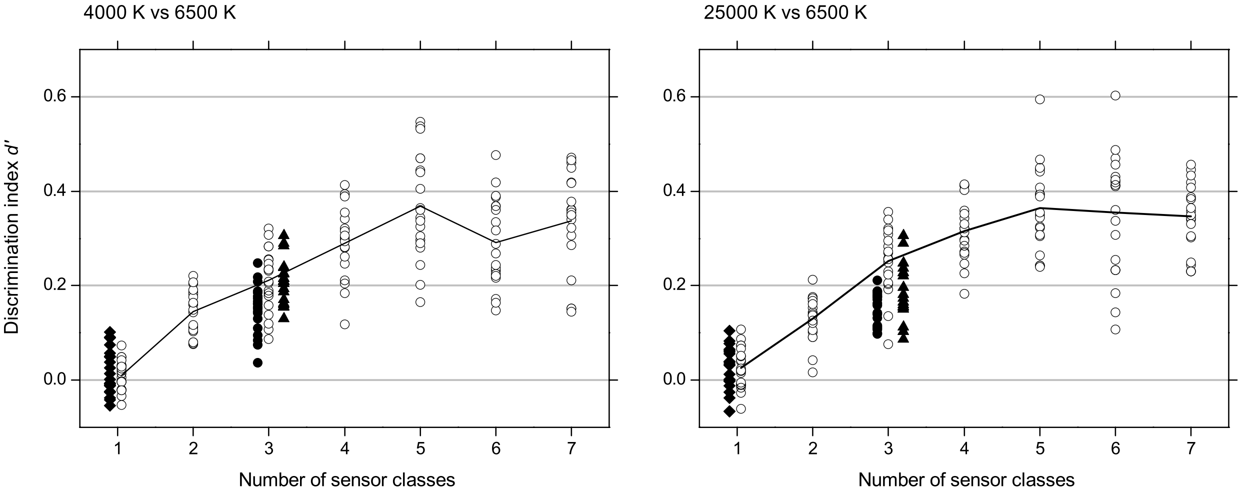

***Figure 3.** Recognition performance as a function of the number of sensor classes. Values of the discrimination index d' [11] are shown for the simulated variable-channel system (open circles); for Nikon D1 sensors (filled circles); the CIE photopic luminance function (diamonds); and the cone fundamentals (triangles). For each number of sensors, independent random samples were drawn from each image of a scene under either a daylight of CCT 25000 K or 4000 K and matched against independent random samples drawn from each image of the scene under a daylight of CCT 6500 K. Matching was based on the intersection of multi-dimensional histograms of ratios of sensor signals in these samples. The data points show the results of 20 such samples, each averaged over 50 natural scenes. The continuous curve represents mean d' over the 20 samples for the variable-channel system. Data based on 10 image points per sample.*

Ratios of sensor signals were obtained as follows. At each scene point $i = 1, 2, \ldots, N$ of the sample, let $\boldsymbol{q}_i = (q_i^1, q_i^2, \ldots, q_i^n)$ be the n-tuplet of sensor responses in classes 1, 2, …, n, and let $(\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N)$ be the N-vector of these n-tuplets. Let σ be a random permutation of the points 1, 2, …, N. Then the set of sample ratios consists of the (unordered) set of N values $\{\boldsymbol{q}_1/\boldsymbol{q}_{\sigma(1)}, \boldsymbol{q}_2/\boldsymbol{q}_{\sigma(2)}, \ldots, \boldsymbol{q}_N/\boldsymbol{q}_{\sigma(N)}\}$, where each of the quotients $\boldsymbol{q}_i/\boldsymbol{q}_{\sigma(i)}$ is given by $\left(q_i^1/q_{\sigma(i)}^1, q_i^2/q_{\sigma(i)}^2, \ldots, q_i^n/q_{\sigma(i)}^n\right)$.

Histograms H were formed from these sets of ratios, but with unequal bin sizes to accommodate the nonuniform distribution of ratios from a uniform distribution of colours, as in Ref. [2].

Histogram Intersection

Let H_a be the histogram based on N points from a test image of scene a under a daylight of CCT 4000 K and let H_b be the histogram based on a different set of N points from a reference image of a scene b under a daylight of CCT 6500 K, where a may or may not coincide with b. The goodness of a match is defined [1] by their intersection $I\,(H_a, H_b)$; that is,

$$I(H_a, H_b) = \frac{\sum_j \min\{H_a(j), H_b(j)\}}{\min\left\{\sum_j H_a(j), \sum_j H_b(j)\right\}}, \tag{1}$$

where j indexes the bins used to form the histograms. Necessarily, $0 \le I(H_a, H_b) \le 1$.

Discrimination index

With 50 scenes, there are 50 possible correct matches, i.e. the test and reference samples come from images of the same scene, and 50×49 = 2450 false matches, where the test and reference samples come from images of different scenes. Let HR be the match hit rate defined by the mean of $I(H_a, H_b)$ over the 50 correct matches and let FAR be the match false-alarm rate defined by the mean of $I(H_a, H_b)$ over the 2450 false matches. As intimated earlier, both HR and FAR were expected to vary with the number of sensor channels. Thus, as the number of sensor channels increases, the conditions for a match become more demanding, and so the hit rate should decrease but so also should the false-alarm rate. The true level of recognition depends on the difference between the two, although this needs to be expressed on a scale that takes into account the limitations of the measure, i.e. intersection, which as a proportion varies between 0 and 1.

One common approach is to summarize the difference between HR and FAR by the discrimination index d' from signal-detection theory [11]; that is, $d' = \Phi^{-1}(\mathrm{HR}) - \Phi^{-1}(\mathrm{FAR})$, where Φ is the normal cumulative distribution function. This index has the advantage of both linearizing proportions and reducing the effects of bias.

Results

Figure 3 shows discrimination index d' plotted against the number of sensor classes of each type. The separate data points represent results from 20 independent samples of 10 points drawn randomly from the images. The continuous curve represents the mean d' over the 20 samples for the variable-channel system. The two plots are for test images obtained under a daylight of CCT 4000 K and under a daylight of CCT 25000 K matched against reference images obtained under a daylight of CCT 6500 K.

Figure 4 shows similar results for data from independent samples of 100 points drawn randomly from the images.

The interpretation of differences in d' values with different numbers of sensor classes is complicated by the constraints imposed by the number of scenes being sampled (the more scenes there are, the greater FAR even though HR remains constant). Importantly, however, the dependence of mean d' on the number of sensor channels in the variable-channel system appears to peak with five sensor channels, after which it levels off and possibly declines with six and seven sensor channels.

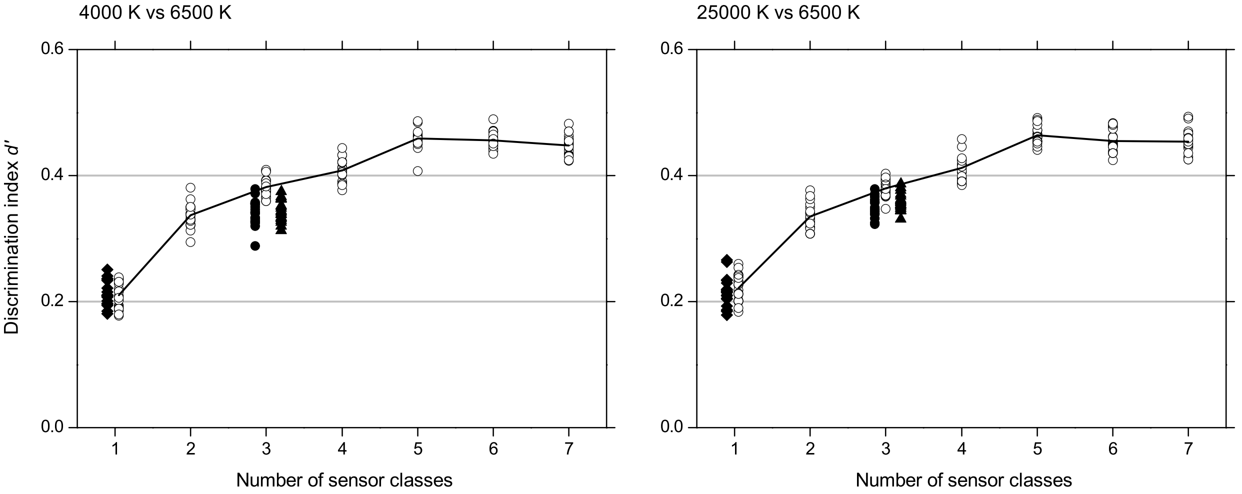

***Figure 4.** Recognition performance as a function of the number of sensor classes. Details as for Fig. 3, but data based on 100 image points per sample.*

Discussion

With just one sensor channel of the simulated variable-channel system, there was little difference in performance between it and the CIE photopic luminance function, both yielding chance levels of scene recognition. But as expected, as the number of channels in the variable-channel system increased, recognition performance increased. With three sensor channels, there was little difference between its performance and that of the Nikon sensors and of the cone fundamentals. As the number of channels in the variable-channel system increased beyond three, performance continued to increase but reached a maximum with about five channels. The failure to increase further may be due to several factors. One possibility alluded to earlier is a decreased signal-to-noise ratio with more channels; another possibility is the potential confound introduced by summarizing recognition performance by a single measure when both match hit rate and match false-alarm rate are varying. In any event, with small samples, it seems that indexing with five sensor channels has advantages over indexing with three sensor channels for the recognition of natural scenes.

As noted in the Introduction, spectral sampling with any set of sensors can be improved by spectral sharpening [2], which can of course be extended to four or more sensor classes. In so far that spectral sharpening narrows spectral sensitivities, and thereby increases invariance to changes in illumination, any improvement in recognition performance should persist with four or five sensor channels.

Acknowledgements

Part of the work described here was reported at the AVA/BMVA Spring Meeting, Cardiff, May 2011. We thank K. Amano and G. Feng for comments on the manuscript.

References

[1] M. J. Swain and D. H. Ballard, "Color indexing," Int. J. Comput. Vis., vol. 7, pp. 11-32, 1991.

[2] B. V. Funt and G. D. Finlayson, "Color constant color indexing.," IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, pp. 522-529, 1995.

[3] S. M. C. Nascimento, F. P. Ferreira, and D. H. Foster, "Statistics of spatial cone-excitation ratios in natural scenes," J. Opt. Soc. Am. A - Opt. Image Sci. Vis., vol. 19, pp. 1484-1490, 2002.

[4] G. D. Finlayson, M. S. Drew, and B. V. Funt, "Spectral sharpening: sensor transformations for improved color constancy.," J. Opt. Soc. Am. A - Opt. Image Sci. Vis., vol. 11, pp. 1553-1563, 1994.

[5] S. M. C. Nascimento, D. H. Foster, and K. Amano, "Psychophysical estimates of the number of spectral-reflectance basis functions needed to reproduce natural scenes," J. Opt. Soc. Am. A - Opt. Image Sci. Vis., vol. 22, pp. 1017-1022, 2005.

[6] E. K. Oxtoby and D. H. Foster, "Perceptual limits on low-dimensional models of Munsell reflectance spectra," Perception, vol. 34, pp. 961-966, 2005.

[7] J. M. M. Linhares, P. D. Pinto, and S. M. C. Nascimento, "The number of discernible colors in natural scenes," J. Opt. Soc. Am. A - Opt. Image Sci. Vis., vol. 25, pp. 2918-2924, 2008.

[8] D. H. Foster, K. Amano, S. M. C. Nascimento, and M. J. Foster, "Frequency of metamerism in natural scenes," J. Opt. Soc. Am. A - Opt. Image Sci. Vis., vol. 23, pp. 2359-2372, 2006.

[9] CIE, "Colorimetry, 3rd Edition," CIE Central Bureau, Vienna CIE Publication 15:2004, 2004.

[10] J. M. DiCarlo, E. Montgomery, and S. W. Trovinger, "Emissive chart for imager calibration," in 12th Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications, Scottsdale, AZ, 2004, pp. 295-301.

[11] N. A. Macmillan and C. D. Creelman, Detection Theory: A User's Guide (2nd Edition). Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.

Author Biography

Nsikak Ekpenyong received his B.Sc.in Physics from the University of Uyo, Nigeria, and his M.Sc. in Artificial Intelligence from the University of Edinburgh, Edinburgh, UK. He is currently working toward his Ph.D. in the School of Electrical and Electronic Engineering, University of Manchester, UK. His research interests include image processing, computer vision, robotics, and pattern recognition