Texture, human perception, and information retrieval measures

Janet S. Payne
Dept of Computing, Buckinghamshire Chilterns University College, High Wycombe, HP11 2JZ, UK
email: janet.payne@bcuc.ac.uk

L. Hepplewhite and T. J. Stonham
Dept of Electronic and Computer Engineering, Brunel University, Uxbridge, Middx, UB8 3PH, UK
email: john.stonham@brunel.ac.uk

Abstract

Content-based image retrieval (CBIR) uses image properties such as colour, texture and shape. Although colour has been used successfully, texture can be even more significant, both for human perception and for CBIR. We present the results from a human study with 30 volunteers that ranks the "most like" images for each of the Brodatz textures. In all cases, only a limited number of the other textures were considered to be in any way similar. This perceptually-derived ranking is then used as a basis for relevance judgements, and retrieval performance measures calculated for the ranking given by each of ten computational texture-based methods. Precision, recall and Faloutsos' AVRR measure are calculated, and the averages plotted for each method. These plots provide some guidance on the optimum number of retrievals. However, it is also necessary to consider the relative ordering of the retrievals, and the nature of the non-relevant images retrieved in each case, when considering system usability.

1. Introduction

The rise of the Internet and the World Wide Web, and the increasing power and decreasing cost of multimedia computer systems, have vastly increased the number of images available to individual computer users. Information retrieval for text is difficult enough – multimedia information retrieval adds yet more complexity! Colour, texture and shape have all been used for image retrieval [5], individually and in combination by systems such as QBIC [6] and Virage [7]. Texture plays a major role in the human visual system, and one study which used both colour and texture patterns found that texture was often more significant [15]. A number of studies have focused on relating texture measures to human perceptions [23], [1], [20], and most have used the Brodatz photographic album of textures [2], consisting of 112 photographs labelled D1 through to D112. (A later volume of photographs by Brodatz, of "Wood and Wood Grains", labelled WW1 to WW112, has not achieved the same popularity in texture analysis.) Although the Brodatz textures were never intended to be used in this fashion, and have been criticised as providing an unrealistic test set, they remain the nearest thing to a standard. Other sets of textured images have been developed, but tend to be used mainly by their developers.

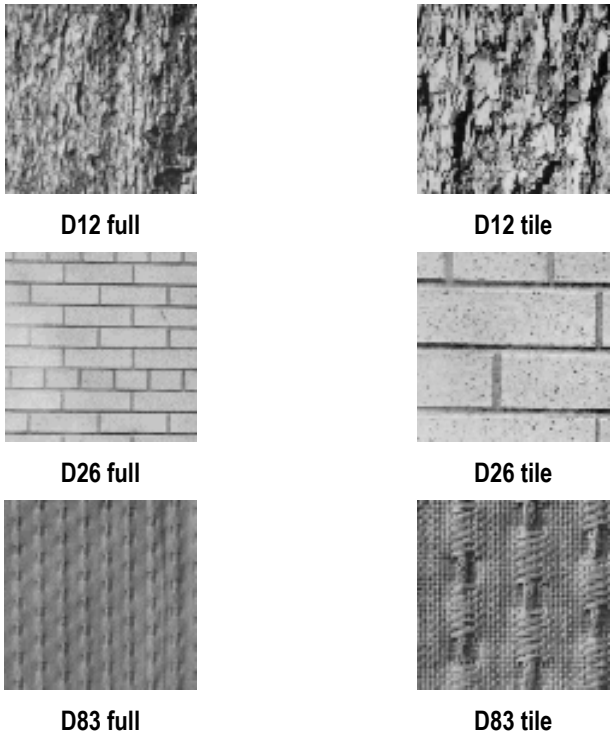

Figure 1 Some examples of textures from the Brodatz album

The original Brodatz album is a photographic one, with high quality "black and white" reproduction of the individual textures, each one occupying most of a 7¾ by

10½ inch page (with a picture area of 19.5 by 24 cm). However, the various studies have worked with digitised images, using varying numbers of pixels and grey levels – most commonly 384 by 384 pixels, with 256 grey levels [1], but 256 by 256 pixels, with 64 grey levels [23], and 512 by 512 pixels, 8 bit grey levels, have also been used. In all cases, square images are used, usually by omitting a strip at the bottom. There are also variations in lighting and method of digitisation (eg photography, or scanning, with various resolutions), which make it harder to compare the different studies. The loss of quality in going from the original page to a computer screen or a 300 or 600 dpi laser printed page is very noticeable. For comparability, we digitised all 112 Brodatz images, using a scanner, to 384 by 384 pixels, 8 bit grey scale, and used this both for the human study, and for each of the image retrieval methods described below [18]. Each image was divided into nine non-overlapping tiles, of 128 by 128 pixels. For our human study, only the top left hand tile was used from each texture. As the majority of the textures are homogeneous, this reduction did not affect the comparability of the images.

If CBIR systems are intended for use by humans, then humans should be the judge of how effective or otherwise the retrievals are. We compare the retrieval performance of the ranking derived from a human study, described in section 2, with those from a range of computational methods, outlined in section 4. There are various ways of measuring effectiveness of a retrieval – we discuss some of them in section 3. One approach is to divide each of the Brodatz images into nine tiles, and using one as the query, measure how many retrievals are required to obtain the other eight tiles. We have not considered that here, since it is not well suited to a human study, requiring the comparison of 1008 rather than 112 samples with each of the others, and in the case of the non-homogeneous textures, can be misleading. Section 5 shows the comparisons for the chosen methods.

Figure 2 Screen-based version, awaiting user's selection

2. Human study

We have carried out two sets of human studies, where volunteers were asked to select which images, in order, they considered to be most like a given target image. As stated above, the full Brodatz dataset of 112 images was used in each case, using only the top left hand tile from each image. The initial study used printed A4 sheets, so that each person could view all the images at once (on the equivalent of an A2 sheet of paper), and volunteers were asked to list, in order, up to four textures that they considered most like each of the samples in turn, using a form giving the image numbers which matched the printed sheets. However, due to the lengthy nature of this process, only six people took part in this first stage. The results were used to produce a computer-based version, shown in *Figure 2*.

Although, due to the limitations of the screen, only 15 images were displayed, this included all the images that anyone had selected in the first stage, plus others at random to make it up to the number where necessary.

Each of the 112 images is shown in turn (in a randomised order) as the centre image in three rows of five, and the volunteer is asked to select, in order, up to four images which they consider most like this target image. They are instructed to do this "as quickly as possible", and most people have taken about 40 to 45 minutes to work through the full set. So far, 30 people have taken part in this second stage, and the selections made by each person have been used to calculate an overall ranking [18]. Of the 30 volunteers, ten were female, twenty male, and the age range was from 18 to 54, with a median age of 26.5. With one exception (one of the authors, JSP) different volunteers took part in each study.

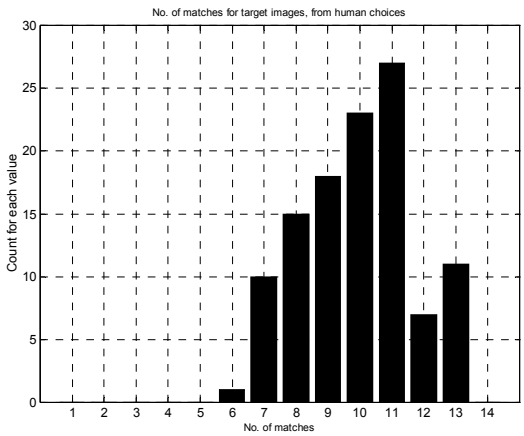

Fig 3 Distribution of choices made for each texture

Although each texture could have up to 14 others considered as "similar", depending upon the selections made, we found that 10 or 11 was the most usual, as shown in *Figure 3*. In one case – D31, pebbles – there was strong agreement about which images most resembled it, and no other selections were made. (In fact, only one person chose D10, and that as their only choice). *Figure 4* shows the textures considered to be similar, in order of first preferences, together with the number of people (out of the total of 30) who included it among their choices. Section 3 below uses this texture as an example to calculate the correlation coefficient between individual and aggregate ranking. These results correspond to what we found in the first human study.

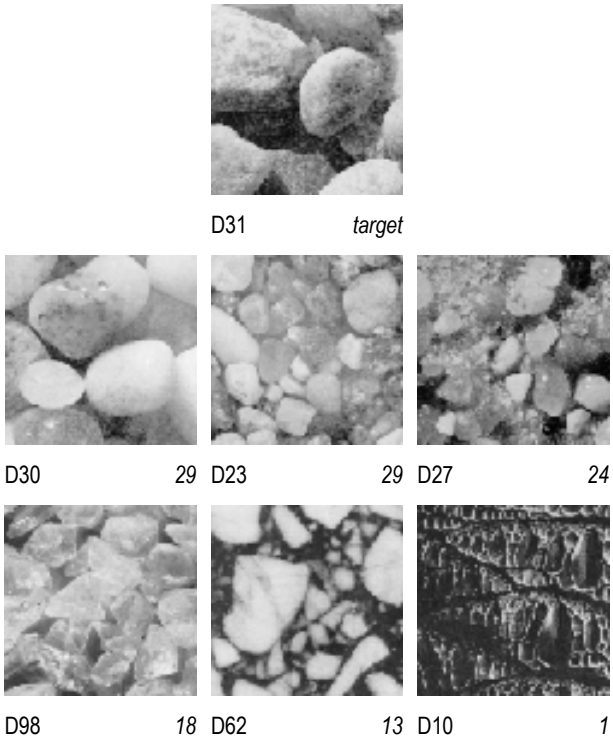

Figure 4 textures ranked as "similar to" D31, showing the number of people who chose each one

At the other extreme, some images proved harder to agree on, and these show weaker correlation between each individual's choices and the combined ranking. These images are: D13, D14, D18, D19, D21, D37, D38, D39, D46, D48, D49, D54, D61, D62, D66, D86, D87, D93.

For example, D86, as discussed in [19] shows a poor correlation between individual rankings. *Figure 5* shows the combined ranked selections, again ordered by the number of first choices, and showing the number of people who included it among their choices.

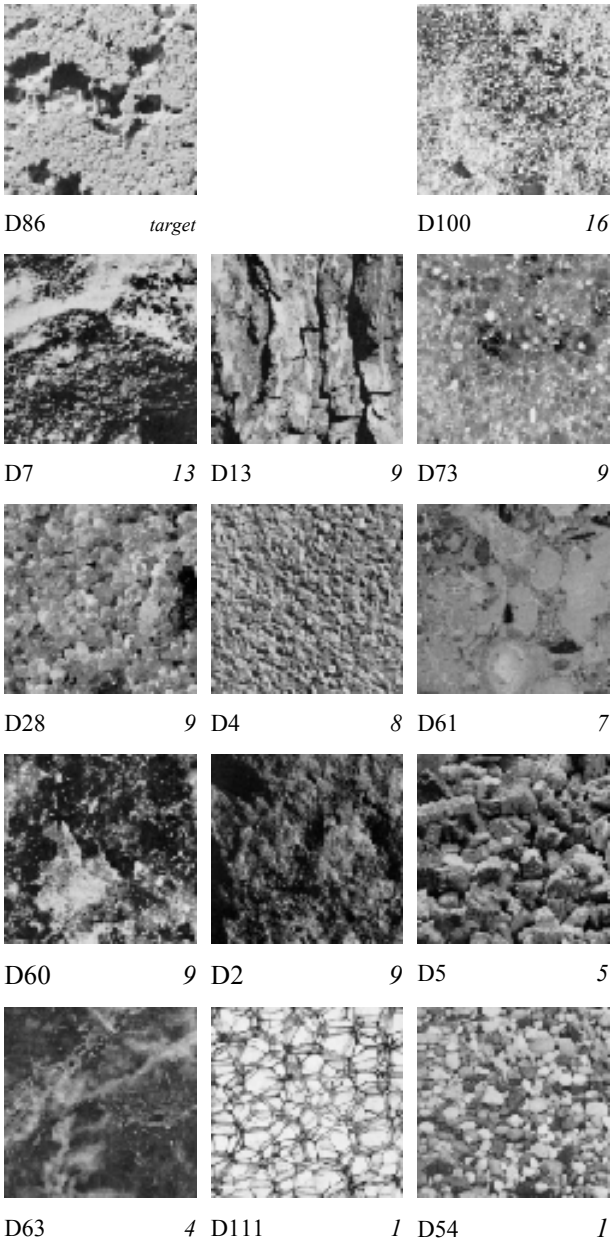

Figure 5 textures ranked as "similar to" D86, showing the number of people who chose each one

The combining ranking of "similar" images, derived from the individual classifications for each of the 112 Brodatz textures, provides a standard set of perceptually derived "relevant" images for use in evaluation of retrieval performance. However, since in all cases, only a limited number of other images were considered relevant (from a minimum of 6, to a maximum of 13 as shown in *figure 3*), there are certain implications when carrying out this evaluation. For any computational method, similarity is generally considered in terms of distance in feature space – and the full set of other images can be ranked as easily as a subset. In practice, in many CBIR systems, we usually consider some smaller number – such as 10, or 15, or 20. From our findings for textures, even this exceeds the maximum possible number of relevant images.

3. Measures of performance

In traditional (document) information retrieval, performance is often measured by using recall and precision [22], [21]. Recall measures the ability of the system to retrieve useful documents, while precision measures its ability to reject useless documents. The two measurements are inter-related, and are defined by:

recall = (retrieved & relevant) / (total relevant)
precision = (retrieved & relevant) / (total retrieved)

These measures can also be applied to image retrieval [5], and have been adapted by eg QBIC [6], to provide a normalized recall. If there are T relevant images in the database, then for an ideal retrieval, all T relevant items occur in the first T retrievals (in any order). Faloutsos *et al* define this as the IAVRR, the ideal AVRR (average rank of all relevant, retrieved images). There is of course the problem of defining what is meant by relevant, whatever measures are used.

For example, using the Brodatz textures, if the relevant images for D47 are defined as:

D46 D18 D101 D52 D35 D102

so that $T = 6$ (taking only the first six, to simplify this example), and a retrieval system returns, in order:

D102	D109	D50	D18	D74	D46	D52
D57	D17	D35	D63	D16	D58	D101

then relevant items appear at 0, 3, 5, 6, 9 and 13 (where the

first position is the 0th). The AVRR for this is therefore (0 + 3 + 5 + 6 + 9 + 13)/6, giving 36/6 = 6. The IAVRR would be (0 + 1 + 2 + 3 + 4 + 5)/6, ie 2.5. The ratio of the AVRR to the IAVRR gives a measure of the effectiveness of the retrieval. *Figure 6* shows the retrieved textures for D47.

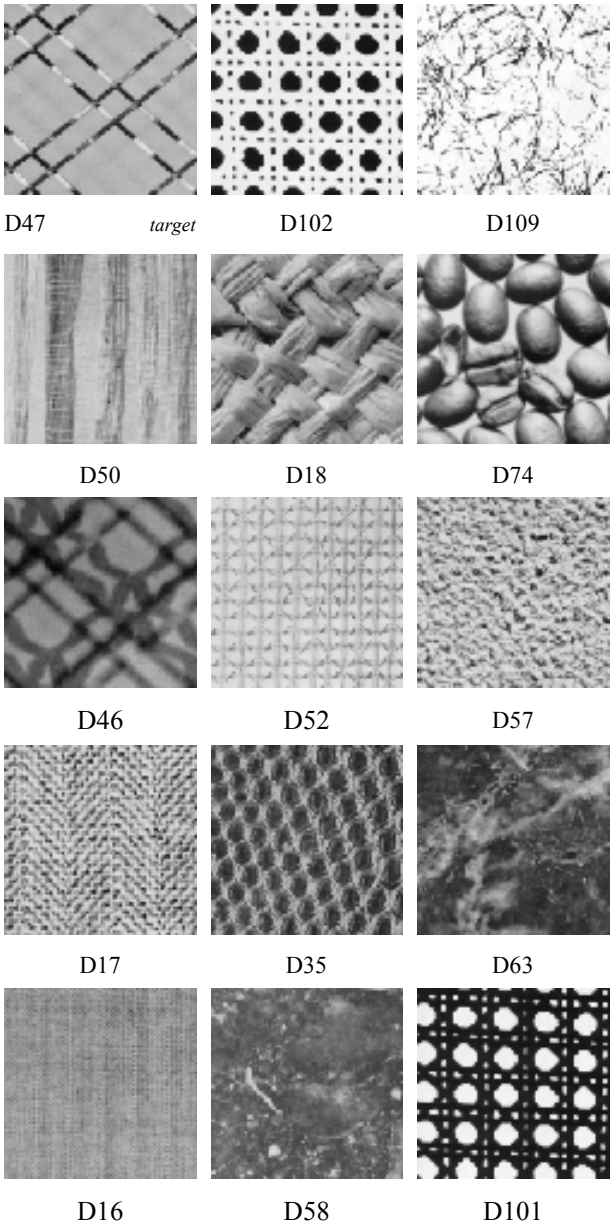

Figure 6 textures ranked as "similar to" D47, by a retrieval system using BTCS

If the order of retrieval matters, which it may well, there are a variety of statistical techniques which provide measures of association, such as Spearman's or Kendall's coefficient of rank correlation [4]. Kendall's tau is perhaps clearer to understand and apply [12], and although the calculated numerical values of the coefficients will be different, both will produce nearly identical results in most cases, and both can be used to test the significance of the association between two sets of data. We have used Kendall's tau to correlate the results of our human study and a number of computational methods [19].

Kendall's tau can be viewed as a coefficient of disorder. For example, consider the following two rankings, where both have selected the same four images, but have placed them in a different order:

1 2 3 4
2 1 4 3

That is, the first person's first choice is ranked second by the other person, and so on. Tau is calculated as

$$\frac{\text{no of pairs in order} - \text{no. of pairs out of order}}{\text{total no. of possible pairs}}$$

For this example, 2 in the bottom row is followed by 1, 4, and 3. 2-1 is out of order, scoring −1, and 2-4, 2-3 are in order, scoring +1 each. Similarly, 1 is followed by 4 and 3. Both are in order, scoring +1 each. Finally, 4 is followed by 3, scoring −1. The number of in-order pairs is four, and out-of-order pairs is two, therefore the total is +2, divided by the maximum number of in-order pairs, N(N − 1) / 2, which here is 6, since N = 4. The value of tau is therefore 2/6, or 0.3333. This gives a measure of the "disarray", or difference in ranking, between the two. It ranges from −1, which represents complete disagreement (a choice of 4 3 2 1 in this example), through 0 (no correlation), to +1, complete agreement (1 2 3 4 in this example). For the number of retrievals used in our experiments, the one-tailed 1% significance level is 0.8, and the 5% significance level is 0.6.

For example, calculating the correlation coefficients for each individual against the combined responses gives a value of 0.8264 for the mean tau for image D31, where individuals showed significant agreement *(figure 4)*, whereas it is only 0.3682 for image D86 *(figure 5)*.

Taking the composite choices for texture D31:

D30	D23	D27	D98	D62	D10

with a rank order of

1 2 3 4 5 6

by definition, and three individuals' choices:

"A" D30 D23 D27 D62 ranking 1 2 3 5

"B" D30 D23 D98 D27 ranking 1 2 4 3

"C" D10 *(no other choices)* ranking 6 7 7 7

(the value 7 indicates that the other choices do not occur within the comparison group, and are tied.)

For "A", the number of in-order pairs is 6, and out-of-order pairs is 0. The maximum number of in-order pairs is also 6, as shown in the example above, and

$$tau = (6 - 0) / 6 = 1$$

as expected, showing perfect agreement.

For "B", the number of in-order pairs is 5, out-of-order 1, with the same maximum, so

$$tau = (5 - 1) / 6 = 0.67$$

still showing significant agreement.

For "C", only one choice was made, and therefore we have to take ties into account [12]. The number of in-order pairs is 3, out-of-order 0, and for the same maximum,

$$tau = (3 - 0) / 6 = 0.5$$

less than 5% significance level.

In section 5, we consider the values of recall, precision, and AVRR, as defined above, for each of the ten computational methods described in section 4 below, using the perceptually-based composite ranking of Section 2 to define relevance.

4. Computational techniques used

A range of computational methods was selected for comparison, covering statistical, Fourier, and spatial approaches, as discussed in [19]. The following ten methods were implemented as in the relevant reference; any specific parameters are detailed below.

Statistical, using co-occurrence matrices:

- **!** **Haralick** implemented using the matrix itself as the feature vector. The number of grey levels in the texture image has been reduced to 16 and a single displacement vector used [8].
- **!** **GLCM** as above, but with 64 levels of intensity. and the commonly cited matrix features of energy, entropy, correlation, homogeneity and inertia were extracted, using four displacement vectors [3].
- **!** **TUTS** implemented as local binary patterns, reducing the feature space dimensionality to N=256, from N=6561 [9].
- ! **BTCS** using binary (thresholded) images with n-tuple size of n=4, and interpixel spacing, t=1, and a global threshold level [16].
- **!** **GLTCS**, using grey scale images, with n=4 and t=1 as for BTCS [17].
- **!** **SRank** again using grey scale images, and n=4 and t=1 as in GLTCS and BTCS but with "roughly equal to" band of " 5 levels [10].

Fourier:

- **!** **R&W** implemented as in [24] with four Ring features and four Wedge features.
- **!** **LSF** Liu's Spectral Features [14] implementing six computationally efficient and optimal features.

Spatial:

- **!** **LTE** Laws' Texture Energy method using nine 3*3 masks and a 5*5 moving window standard deviation estimate [13].
- ! **Gabor** Gabor filter energy with four orientations and up to four scales, depending on the window size. The features extracted are quadrature filter pair mean and standard deviation of energy [11].

5. Comparisons of retrieval performance

A standard way of evaluating retrieval performance is to plot an average recall-precision graph [22]. *Figure 6* shows the results, for each of the ten methods used, with retrievals from 1 to up to 14, averaged over all textures. For the case

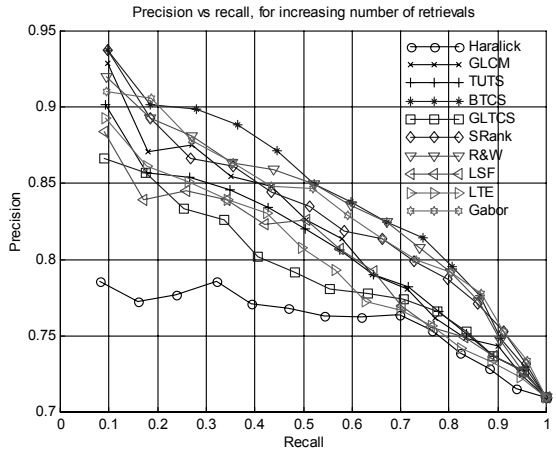

Figure 6 – precision plotted against recall

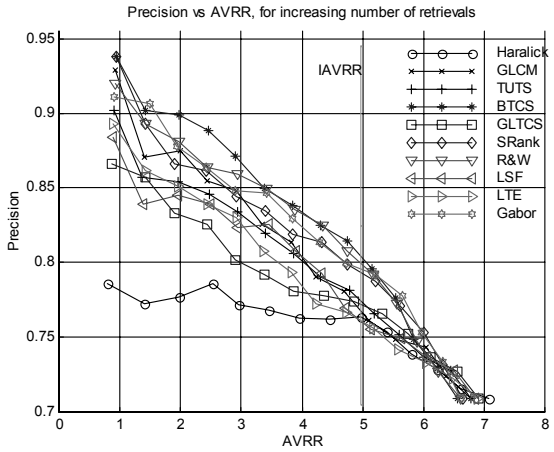

Figure 7 – precision-AVRR plot

of 14 retrievals, all the relevant images in the comparison set were retrieved, for all methods, and the recall is therefore 1.

Table 1 – values for BTCS -- recall

retrievals	min	max	mean	std
1	0	0.1667	0.0971	0.0316
5	0.125	0.7143	0.4466	0.1059
10	0.625	1	0.8068	0.0944
14	1	1	1	0

Table 2 – values for BTCS -- precision

retrievals	min	max	mean	std
1	0	1	0.9375	0.2431
5	0.2	1	0.8714	0.1778
10	0.5	1	0.7955	0.1358
14	0.4286	0.9286	0.7092	0.1245

Table 3 – values of AVRR for BTCS

retrievals	min	max	mean	std
1	0	1	0.9375	0.2431
5	1	4	2.9055	0.3843
10	3	6.8333	5.1529	0.6264
14	4.5714	8.7778	6.5583	0.8843

As can be seen from the graph of *Figure 3*, in most cases the comparison set contained a smaller number of relevant images; the precision therefore decreases. It may be noted that statistical methods produce one of the best performances (BTCS), as well as the poorer ones (GLTCS, Haralick). Even so, the overall results for all methods seem quite effective, on this measure. The averaging does of course disguise differences across the various textures, and Tables 1, 2 and 3 show the range of values for BTCS for recall, precision, and AVRR, respectively.

Figure 7 plots precision against AVRR, for the same ten methods over all 112 textures, using the composite ranking as the standard for comparison. The value of IAVRR, the ideal AVRR, is also plotted, as a vertical line with a value of 4.9643. Note that the shape of the graph is very similar to that for recall alone, irrespective of ranking.

6. Discussion

The results from the human study described in section 2 provide a basis for comparison with any computational or other retrieval results for the Brodatz textures. It is noticeable that only a limited number of images are considered to be "similar", and this low number of relevant images affects the measures for precision, recall, AVRR and correlation calculations.

Plotting recall-precision, and AVRR-precision graphs for the average for each technique produces similar curves, showing apparently good performance characteristics in each case. The AVRR plot discriminates slightly better, but not by much. These plots provide some guidance on the optimum number of retrievals. However, it is also necessary to consider the relative ordering of the retrievals, and the nature of the non-relevant images retrieved in each case, when considering system usability.

References

[1] M. Amadasun and R. King, Textural Features Corresponding to Textural Properties, IEEE SMC vol 19 No5 p1264-1274, 1989

[2] P. Brodatz, Textures - a photographic album for artists & designers, Dover, New York, 1966.

[3] R. W. Conners and C. A. Harlow, A theoretical comparison of texture algorithms, IEEE Trans. PAMI, vol 2 no.3, p204-222. 1980.

[4] W. J. Conover, Practical Nonparametric Statistics, 3rd ed, John Wiley & Sons, New York, 1999

[5] A. Del Bimbo, Visual information retrieval. Morgan Kaufmann Publishers, San Francisco, 1999.

[6] C Faloutsos and M. Flickner and D. Petcovic and W Niblack and W. Equitz and R. Barber, Efficient and effective querying by image content, Tech Report, IBM, 1993.

[7] A. Gupta and R. Jain, Visual information retrieval. Comm ACM vol 40 no. 5, p70-79, 1997.

[8] R.M. Haralick and K. Shanmugam and I. Dinstein, Textural features for image classification, IEEE Trans. SMC, vol 3 no 6, p610-621. 1973.

[9] D.C. He and L. Wang, Texture unit, texture spectrum and texture analysis, IEEE Trans. Geoscience and Remote Sensing, vol 28 No 4, pp 509-512, 1990

[10] L. Hepplewhite and T.J. Stonham and R.J. Glover, Automated visual inspection of magnetic disk media, Proc. of 3rd ICECS, vol 2, p 732-735. 1996

[11] A. K. Jain and F. Farrokhnia, Unsupervised texture segmentation using Gabor filters, Pattern Recognition, 24(12):1167-1186. 1991.

[12] M. Kendall and J. Dickinson Gibbons, Rank Correlation Methods, 5th ed, Edward Arnold, London, 1990

[13] K.I. Laws, Texture image segmentation, PhD thesis, University of Southern California, 1980.

[14] S.S. Liu and M.E. Jernigan, Texture analysis and discrimination in additive noise, CVGIP, vol 49, p52-67, 1990

[15] A. Mojsilovic, J. Kovacic, J. Hu, R. J. Safranek and S. K. Ganapathy, Matching and retrieval based on the vocabulary and grammar of color patterns, Proc IEEE ICMCS99, Vol I p189-194

[16] D. Patel and T.J. Stonham, Low level image segmentation via texture segmentation, Proc SPIE vol 1606, p621. 1991.

[17] D. Patel and T.J. Stonham, Unsupervised / supervised texture segmentation and its application to real world data, Proc. SPIE Visual Comms. and Image Processing, vol 1818. 1992.

[18] J. S. Payne, L. Hepplewhite and T. J. Stonham, Evaluating content-based image retrieval techniques using perceptually based metrics Proc SPIE, Vol 3647, p122-133. 1999.

[19] J. S. Payne, L. Hepplewhite and T. J. Stonham, Perceptually based metrics for the evaluation of textural image retrieval methods. Proc IEEE ICMCS99, Vol II p793-797, 1999

[20] A. R. Rao, N. Bhushan and G. L. Lohse, The relationship between texture terms and texture images: a study in human texture perception, Proc SPIE, vol 2670, p206-214. 1996.

[21] C. J. van Rijsbergen, Information Retrieval. Butterworths, London, 1979.

[22] G. Salton and M. J. McGill, Introduction to modern information retrieval. McGraw-Hill, New York, 1983.

[23] H. Tamura and S. Mori and Y. Yamawaki, Textural features corresponding to visual perception, IEEE Trans. on SMC, vol 6 no 6, p460-473. 1978.

[24] J. S. Weszka and C. R. Dyer and A. Rosenfeld, A comparative study of texture measures for terrain classification, IEEE Trans SMC, vol 6 no. 4, p269-285. 1976.