TEXTURE CLASSIFICATION BASED ON HYPOTHESIS TESTING APPROACH

Mark J. Carlotto
The Analytic Sciences Corporation
1 Jacob Way
Reading, Mass. 01867 USA

An approach to texture classification formulated as a hypothesis test on the local sample distribution is described. Texture is measured by the local histogram computed in a window that slides over the image. For each shift of the window, the local histogram is updated, and the resulting sample distribution is compared to a set of pre-computed sample distributions for textures of interest. The two-sample Kolmogorov-Smirnov test is used to decide if two histograms are similar by comparing the maximum difference between the sample distributions to a threshold which is a function of the sample size (i.e., the size of the sliding window) and the level of significance of the test. The method is relatively easy to implement and is computationally feasible. Accuracy of 87-100% correct classification is demonstrated on natural textures from Brodatz's photographic album.

1. INTRODUCTION

Considerable attention has been devoted to the modeling, representation, and classification of texture for applications ranging from medicine to remote sensing. In general, the problem of classfying image textures has been approached by first developing a texture representation (or model), and then defining a metric for comparing and discriminating between textures. Wong[1] and Nagao[2] are among those who have proposed texture measurements based on local histograms. Since textures often possess a high degree of variability, it is not always possible to classify textures reliably by simply assigning the texture to the "nearest" class.

In this paper, the classification of textures is formulated as a hypothesis test. Texture classification may then be viewed as texture detection where the texture sampled at a particular point in the image is tested against a set of pre-computed references, each representing a particular texture type. The Kolmogorov-Smirnov two-sample test[3], a non-parametric statistical test used to determine if two random samples were drawn from the same statistical population, is used here for texture classification. In short, if the difference between the two sample distribution functions is less than a threshold, which is a function of the significance level and the sample size, the two samples may be said to have been obtained from the same statistical source, in other words, the textures are the same.

The organization of this paper is as follows: In Section 2, the basic assumptions of the method are stated and the ability to measure texture with local histograms is discussed. The use of the two-sample Kolmogorov-Smirnov test to discriminate between textures is described in Section 3. Finally, the performance of the texture classifier on textures from Brodatz's photographic album is presented in Section 4.

2. TEXTURE MEASUREMENT

It is assumed that a texturally homogeneous region in the image may be characterized by its underlying continuous probability density over one or more image features. Image features may include intensity, gradient, and other measurements derived from the image.

Let the input image array be denoted $\underline{x}=\{x(i,j)\}$ for $0 \le i < M$ and $0 \le j < N$. The texture in a rectangular region in the image is measured by the sample histogram computed over that region. Since the sample histogram is an estimate of the continuous density, it can be used to approximate the probability distribution for hypothesis testing. Let $W(i,j)$ be the PxQ pixel window centered at (i,j) in $\underline{x}$ which defines the region of support: $i-P/2 \le p < i+P/2$ and $j-Q/2 \le q < i+Q/2$. The population of the $k-th$ bin in the local histogram $\underline{f}(i,j)=\{f_k\}$ is

$$f_k = \#[x_k \le x(i,j) < x_{k+1}] \tag{1}$$

for $(i,j) \in W(i,j)$ where K is the number of bins, $\#[A]$ denotes the number of occurrences of A, and the x_k levels divide the range into K intervals. The PxQ texture block is thus transformed in dimensionality with loss of information from PQ to K. Note that the entire histogram is used as a texture measurement rather than selected statistics (as computed from the co-occurrence matrix, for example).

Local histograms of image grey-levels capture the grey-level (tonal) variations in the image but do not adequately represent textural properties. Qualitatively, texture can be described in terms of its coarseness, roughness, directionality, and uniformity. To provide additional textural information, local histograms over other image measures such as the gradient magnitude and gradient angle can be used. To illustrate the use of local histograms to describe texture at least in qualitative terms, consider the following situations. A texture with a strong directional component (i.e., with edges lined up ϕ relative to the x-axis) gives rise to an angular histogram with strong lines (large bin populations) at $\phi \pm 90$ degrees. The roughness and uniformity of the texture is captured in the local gradient magnitude histogram. The position of the major mode will be indicative of the average roughness. The

CH2046-1/84/0000/0093$01.00©1984 IEEE

uniformity (variation in the roughness) will be proportional to the spread in the histogram. The relative balance between pixels with some directionality and those with little or no directionality will be indicative of the coarseness of the texture since large texture elements will have fewer edges per unit area than small texture elements.

3. HYPOTHESIS TESTING

The above texture measure serves as the basis for the hypothesis testing approach to texture classification. Initially in a training data set, a typical region for each texture type is selected and the sample histogram computed. Large regions are desirable in order to obtain a good statistical sample. For certain types of features (e.g, gradient angle), long term variations or inhomogeneities in the texture (e.g., rotations of the basic repition pattern) can result in poor texture references if the region is too large. A set of P reference distributions $\underline{G}(p)$ are computed from the sample histograms accumulated in the above regions. At each point in the image, the local histogram $\underline{f}(i,j)$ is used to compute the sample distribution $\underline{F} = \{F_k\}$ by

$$F_k = \sum_{m=0}^{k} f_m . \tag{2}$$

Again, a large window is desirable to obtain a good estimate for the distribution, however, in order to obtain good spatial resolution, the window should be small. In general, the window should be large enough to capture the basic repitition pattern (macro-structure) of the texture but not too large so as to pick up the long term variations in the texture.

Thus at each point in the image $(P+1)$ hypotheses are considered:

- The unknown texture in the PxQ region centered at (i,j) is an instance of one of the P textures previously identified, or
- The texture is not an instance of any of these P textures.

If the texture cannot be positively identified as being an instance of at least one of the P textures, the null hypothesis that it is not any of these texture is true. Also if more than one hypothesis is true, it may be desirable to choose the null hypothesis.

In order to achieve invariance to texture rotation, a composite hypothesis test[4] must be performed. Rotational invariance means that we care only about the relative orientations of the texture elements in the window. The test involves circularly shifting the top $(K-1)$ bins (i.e., the directional components) of the gradient angle histogram $mod(K-1)$, and testing the resulting sample distributions against each of the reference distributions. Thus, $P(K-1)$ simple tests must be performed at each point.

The maximum difference between $\underline{G}$ and $\underline{F}(p)$

$$d_p = || \underline{G} - \underline{F}(p) ||_\infty \tag{3}$$

is used as the measure of similarity between the local (unknown) texture and the $p-th$ texture. The two-sample Kolmogorov-Smirnov test may then be used to accept the hypothesis that the local texture is an instance of the $p-th$ texture at a level of significance $H(t)$ if:

$$d_p < \frac{t}{N} \tag{4}$$

where $N = [N_L N_R / (N_L + N_R)]^{\frac{1}{2}}$, $N_L = PQ$ and N_R are the local and reference histogram sample sizes. Values of t to achieve a given $H(t)$ are tabulated[3].

The effect of varying $H(t)$ on the detection and false alarm rates is now considered. For finite sample sizes, the significance level of the two-sample Kolmogorov-Smirnov test is only approximately equal to the probability that the maximum difference between distributions d_p is less than the threshold t/N given that the two distributions were drawn from the same statistical source. If we define $P[D_p | H_p]$ to be the probability of deciding that hypothesis H_p is true given that it is true, then $H(t) \rightarrow P[D_p | H_p]$ as $N \rightarrow \infty$. The average error rate is

$$\begin{aligned} p[e] &= \sum_{p=0}^{P-1} \sum_{q \neq p} P[D_q | H_p] P[H_p] \\ &= 1 - \sum_{p=0}^{P-1} P[D_p | H_p] P[H_p] \end{aligned} \tag{5}$$

which can, in principle, be made to approach zero as $H(t) \rightarrow 1$ for infinite sample sizes. Since $H(t)$ only approximates the true significance level for finite sample sizes, increasing $H(t)$, and hence t to increase the detection rate will, beyond a point, increase the false detection and false alarm rates as well. This is easy to see, since by increasing t, we make the test less stringent in that greater within-class variability is tolerated. Statistical rules of thumb are to use high significance levels, $H(t) > 0.95$, if it is important not to miss detecting the event, and lower levels, $H(t) < 0.95$, if it is important to reduce the false alarms. Also, by choosing the null hypothesis when the texture is assigned to more than one class, the error rate may be reduced by simply not classifying those regions. This sort of graceful degradation is desirable when the textures are not distinct enough to allow reliable classification.

4. EXPERIMENTAL RESULTS

To demonstrate and evaluate the performance of the texture classifier, four textures from Brodatz's photographic album[5] were digitized and pasted into a mosaic (Fig. 1). The textures D18, D66, D84, and D20 were selected and are refered to as textures B1 through B4 hereafter.

The grey-level (GL), gradient-magnitude (GM) and gradient-angle (GA) images were used in the experiments. The later two images were computed from the original grey-level image using the Sobel operator. The GL and GM images were quantized to 32 levels and the GA was quantized to 9 levels. Bins $K = 1$ through 8 in the GA histogram are the orientations of edge-type pixels in 45 degree increments; bin $K = 0$ contains the number of pixels which have no net directionality. Histograms were computed over typical regions (33x33 in size) in the GL, GM, and GA images for each texture (Fig 2).

First, the textures were classified on the basis of their GL information. The GL histograms for B2 and B4 (Fig. 2d and 2j) are distinct from each other while the GL histograms for B1 and B3 (Fig. 2a and 2g) are relatively similar. We would thus expect B2 and B4 to be classified better then B1 and B3 using GL information. The results summarized in Table 1 confirms our intuition. (The percentage of pixels not classified are denoted "NC"). Due to illumination variations over the light table used in digitizing the textures, the GL histograms were not invariant over the image. Approximately 84% of the pixels were not classified either because the differences between the local and reference distributions exceeded the detection threshold or multiple textures were assigned to the same region (i.e., the classes were not well-separated). Yet, if a pixel was classified, the probability that it was correctly classified was about 90% overall. The large fraction of unclassified pixels suggests that the grey-level was not a good feature to use for these particular textures.

Table 1 GL Texture Classification Results

Actual/Classified	B1	B2	B3	B4	NC
B1	.053	.000	.072	.000	.875
B2	.000	.104	.000	.000	.896
B3	.020	.000	.182	.000	.798
B4	.000	.000	.000	.230	.770

The four textures were then classified using the GA information. The GA histograms for textures B1 and B2 are fairly distinct while those for B3 and B4 are quite close to one other. As a result, B1 and B2 were classified quite well, whereas B3 and B4 were left largely unclassified due to their similarity to one another (Table 2). Approximately 38% of the pixels were classified, and of those classified, 87% were correctly classified.

Table 2 GA Texture Classification Results

Actual/Classified	B1	B2	B3	B4	NC
B1	.564	.002	.099	.002	.333
B2	.087	.581	.004	.011	.317
B3	.000	.000	.043	.000	.957
B4	.000	.000	.001	.106	.893

Next, GM information was used for classifying the textures. The results using a 33x33 window are summarized in Table 3. Using the gradient magnitude, the effect of changing the size of the training regions and the processing window were also investigated. Increasing the window size to 65x65 increased the portion of pixels not classified from 38% to 62%. However, all the pixels that were classified were classified correctly. Decreasing the window size to 17x17 decreased the number of pixels not classifed from 38% to 27%, but the classification accuracy decreased from 99% to 93%. Lowering the level of significance decreased the number of pixels classified since the test became more stringent, i.e., less intra-class variability was tolerated. As a result, the classification accuracy for those classified increased.

Table 3 GM Texture Classification Results

Actual/Classified	B1	B2	B3	B4	NC
B1	.583	.000	.002	.000	.415
B2	.004	.585	.000	.000	.411
B3	.000	.000	.693	.004	.303
B4	.000	.000	.000	.633	.367

Since the GM worked well in detecting B3 and B4, and the GA worked well in detecting B1 and B2, by using GM to detect only B3 and B4 and GA to detect only B1 and B2, a classification rate of 96% was achieved over 69% of the pixels.

5. SUMMARY

The texture classifier described in this paper is based on the use of histograms to represent the texture in a sliding window, and statistical tests to discriminate between textures on the basis of their histograms. An evaluation of the classifier on known textures, using several features, varying the window size and the significance level of the test resulted in classification accuracies between 87% and 100%. By reformulating texture classification as texture detection, the performance of the classifier can be controlled. As a result, textures which cannot be classified reliably at the specified level of significance (either because they are not dinstinct enough from each other or because there is too much variability within the texture) are not classified.

REFERENCES

1. M.Nagao, H.Tanabe, K.Ito, "Agricultural Land-Use Classification of Aerial Photographs by Histogram Similarity Method," *Proceedings of the Third International Joint Conference on Pattern Recognition, 1976*
2. H.Shen, A.Wong, "Generalized Texture Representation and Metric", *Computer Vision, Graphics, and Image Processing, Vol. 23, pp. 187-206, 1983*
3. M.DeGroot, *Probability and Statistics*, Addison-Wesley, Reading, Mass., 1975.
4. A.Whalen, *Detection of Signals in Noise*, Academic Press, New York, 1971.
5. P. Brodatz, *Textures*, Dover, New York, 1966.

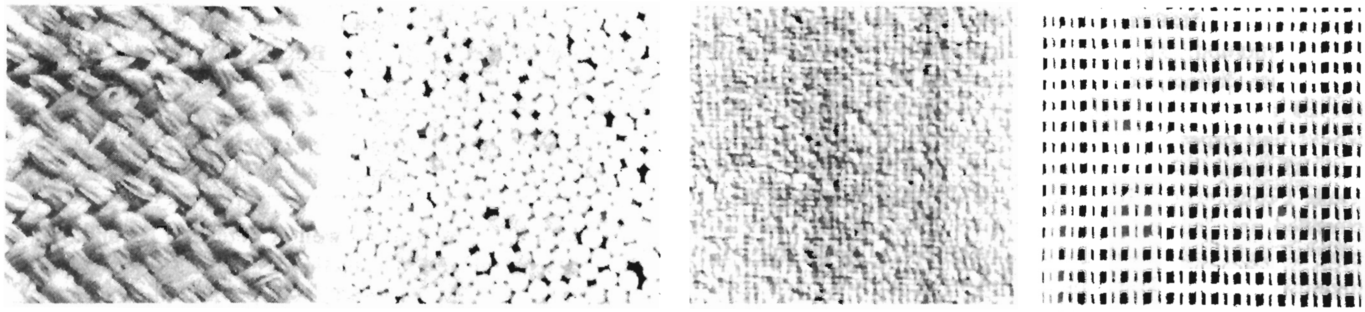
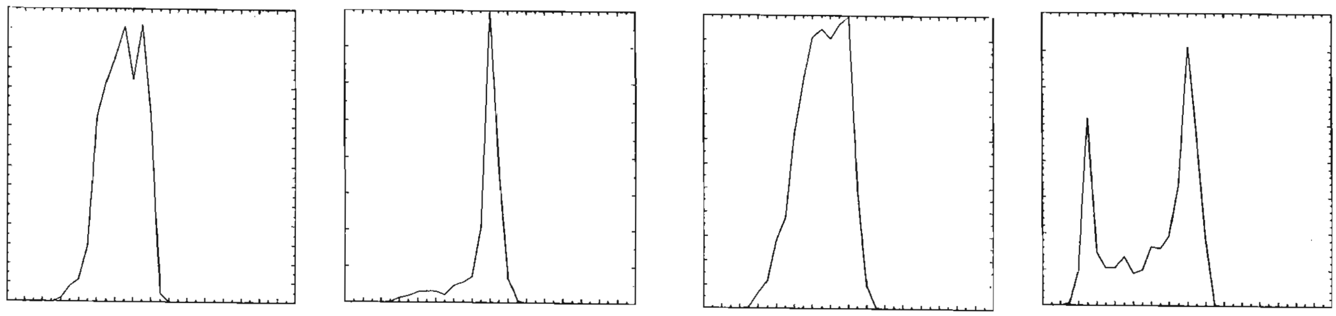
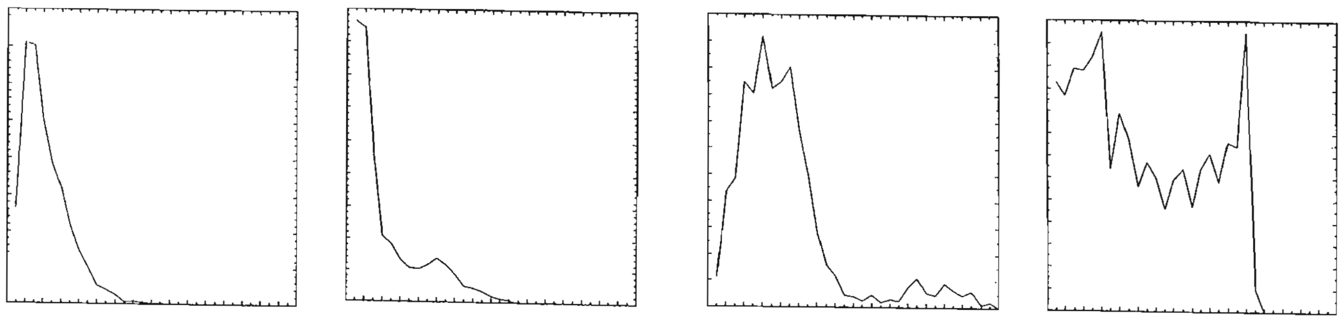


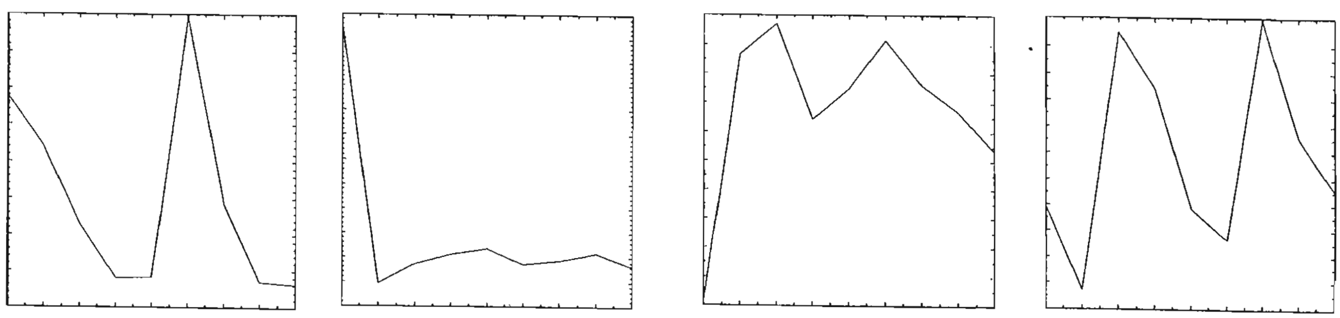
Fig 1 Textures from Brodatz's Photographic Album



Gray-Level (a,d,g,j)



Gradient-Magnitude (b,e,h,k)



Gradient-Angle (c,f,i,l)

Fig 2 Reference Histograms