A Maximum Likelihood Investigation Into Texture Classification

N. Sebe M. Lew D.P. Huijsmans

Leiden Institute of Advanced Computer Science

Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands

{nicu mlew huijsman}@wi.leidenuniv.nl

ABSTRACT

Textures are one of the basic features in visual searching and computional vision. In literature, most of the attention has been focussed on the texture features with minimal consideration of the noise models. In this paper we investigate the problem of texture classification from a maximum likelihood perspective. We take into account the texture model, the noise distribution, and the inter-dependence of the texture features. Our investigation shows that the real noise distribution is closer to an exponential than a Gaussian distribution, and that the L_1 metric has a lower misdetection rate than L_2 or the Kullback discriminant. Furthermore, we provide a direct method for deriving an optimal distortion measure from the real noise distribution, which experimentally provides consistently improved results over the other metrics. We conclude with results and discussions on an international texture database.

Keywords: Maximum likelihood, texture classification, noise distribution

1 INTRODUCTION

Texture analysis is important in many applications of computer image analysis for classification, detection or segmentation of images based on local spatial patterns of intensity or color. Textures are replications, symmetries and combinations of various basic patterns or local functions, usually with some random variation. Textures have the implicit strength that they are based on intuitive notions of visual similarity. This means that they are particularly useful for searching visual databases and other human computer interaction applications. However, since the notion of texture is tied to the human semantic meaning, computational descriptions have been broad, vague and sometimes conflicting.

There are several problem spaces associated with texture classification. In texture segmentation, the goal is to segment the image into regions which have similar textures. The textures may not be known apriori and the major difficulties are in deciding what the dominant region textures are, and locating the region boundaries.

In our problem space, the texture classes are known apriori, and the goal is to classify a new texture into one of the known classes. It is important to extract the features and metrics which give the most discriminatory power towards minimizing the misdetection rate.

Many methods have been proposed to extract texture features either directly from the image statistics, e.g. co-occurrence matrix, or from the spatial frequency domain [13]. Ohanian and Dubes [8] studied the performance of four types of features: Markov Random Fields parameters, Gabor multi-channel features, fractal-based features and co-occurence features. Comparative studies to evaluate the performance of some texture measures were made in [10], [9]. Most of these previous studies have focussed on the features, but not on the metric, nor on modeling the noise distribution. Using the previously introduced texture distribution features, we study the effect of the noise and the metric and their interrelationship within the maximum likelihood paradigm.

1.1 Texture Distribution Models

In this paper, we focus on the texture distribution methods. These methods can be categorized into the following classes: gray-level differences; Laws' texture measures; center-symmetric covariance measures; and local binary patterns. We briefly describe them and give references to the original papers.

Gray-level differences

The class of gray-level differences were used by Unser [12]. These methods capture the distribution of local contrast in different directions. Since they rely on the differences, they provide reduced dependence on intensity. In our implementation, we used four measures based on the gray-level difference method. By accumulating the differences of the adjacent gray levels in the horizontal and vertical directions, we create histograms DIFFX and DIFFY. When we accumulate the absolute differences in both horizontal and vertical directions, we arrive at DIFF2, and in DIFF4, we accumulate the absolute differences in all four principal directions, which also gives rotational invariance.

Laws' texture energy measures

Beyond gray-level differences, we examine larger convolution masks which measure the energy of local patterns. From Laws work [7] on texture energy measures, we used four Laws' 3×3 operators: 2 perform edge detection in vertical (L3E3) and horizontal directions (E3L3) and the other ones are line detectors in these two orthogonal directions (L3S3 and S3L3).

Center-symmetric covariance measures

We also consider statistical concepts of symmetry and covariance. Harwood [3] introduced measures for gray-level symmetry (positive) and anti-symmetry (negative) by computing local auto-covariances or auto-correlations of center-symmetric pixel values of suitably sized neighborhoods. We implemented a local center-symmetric auto-correlation measure based on neighborhood rank-order (SRAC) and a related covariance measure (SCOV).

Local binary patterns and trigrams

Another way of analyzing local patterns is to binarize the local pattern information and measure the distribution of these patterns in the texture. Ojala [9] proposed a texture unit represented by eight elements, each of which has two possible values (0,1) obtained from a neighborhood of 3×3 pixels. These textures units are called local binary patterns (LBP) and their occurrence of distribution over a region form the texture spectrum. The LBP is computed by threshold each of the noncenter pixels by the value of the center pixel, resulting in 256 binary patterns. The LBP method is a gray-scale invariant and can be easily combined with a simple contrast measure by computing for each neighborhood the difference of the average gray-level of those pixels which have the value 1, and those which have the value 0, respectively.

Another texture unit called **trigram** was introduced by Huijsmans [5]. This texture unit is represented by 9 elements each of which has two possible values (0,1) obtained from a neighborhood of 3×3 pixels. The value 0 or 1 associated with each element is calculated by applying a threshold in gradient space. If the pixel value is greater than the threshold then, the assigned value of the corresponding trigram element is 1, otherwise 0. This results in 512 trigrams which are accumulated in a histogram. Note that for the trigrams it is important to select the threshold properly.

Complementary feature pairs

In many cases a single texture measure cannot provide sufficient information about the spatial and frequency oriented structure of the local texture. Better discrimination of textures can be obtained considering joint occurrences of two or more features. Therefore, we consider pairs of features which provide complementary information.

The center-symmetric covariance measures provide robust information about the local texture, but little about the exact local spatial pattern. This suggests that as complementary features we should consider measures that provide spatial patterns such as LBP, trigrams or any difference measure. We consider two different features combined with LBP. LBP/C is based on the contrast measure already introduced and the other pair is LBP/SCOV. Laws' masks perform edge and line detections in horizontal and vertical directions. Since these patterns can be in arbitrary directions, the joint use of edge and line detectors in the orthogonal directions should be considered. Similarly, the joint use of histograms of differences between neighboring pixels computed in the horizontal and vertical directions should provide useful information for texture discrimination. The pair L3E3/E3L3 corresponds to edge detection, L3S3/S3L3 to line detection and DIFFX/DIFFY to absolute gray-level in the horizontal and vertical direction, respectively, DIFFY/SCOV combines absolute gray-scale differences in the vertical direction with the center-symmetric covariance measure.

1.2 Texture Classification

In our experiments we used a nearest neighbor classifier. The classification of a sample was based on comparing the sample distribution of feature values to several pre-defined model distributions of feature values with known true-class labels. The sample was assigned the label of the model that was found to be more similar, using a certain similarity measure. Here, we use the term **distortion measure** to denote a measure of dissimilarity. This is a necessary distinction because some distortion measures are not metrics.

Consider that s and m are the sample and the model distributions, n is the number of bins and s_i, m_i are respective sample and model probabilities at bin i. In this context we can define the L_1 metric:

$$d_{L_1}(s,m) = \sum_{i=1}^{n} |m_i - s_i|$$

Similarly, the L_2 metric will be

$$d_{L_2}(s,m) = \sqrt{\sum_{i=1}^{n} (m_i - s_i)^2}$$

The model distribution for each class was obtained by scanning the gray-scale 256×256 texture image with the local texture operator. The number of bins used in quantization of feature space is important. Histograms with small number of bins will not provide enough discriminative information about the distributions. Furthermore, if histograms have too many bins and the average number of entries per bin is small, than the histograms become sparse and unstable.

In this paper we compare the L_1 and L_2 metrics and also compare the misdetection rates when using the Kullback discriminant, which is not a metric, but is a distortion measure. The Kullback discriminant [6] measures likelihoods that samples are from an alternative texture class, based on exact probabilities of feature values of pre-classified texture prototypes:

$$d_{Kullback}(s,m) = \sum_{i=1}^{n} (s_i \log \frac{s_i}{m_i})$$

Furthermore, we used the real distribution to create a distortion measure using the maximum likelihood paradigm and gave complete comparative results involving all of the distortion measures introduced.

2 MAXIMUM LIKELIHOOD ESTIMATE

Suppose we extract N sample blocks from a texture M. We define x_i as the feature vector corresponding to the sample block i, and m as the feature vector corresponding to M. The distortion between x_i and m is defined as n_i, the "noise" vector, which gives:

$$x_i = m + n_i \qquad i = 1, ..., N \tag{1}$$

The similarity probability can be defined:

$$P(x_i, m) = \prod_{j=1}^{B} \{\exp[-\rho(x_i[j], m[j])]\} \tag{2}$$

where ρ is the negative logarithm of the probability density of the noise and m and x_i are vectors of size B with index j, which we denoted as $m[j]$ and $x_i[j]$.

According to (2) we have to find the probability density function of the noise that maximizes the similarity probability: *maximum likelihood* estimate for the noise distribution [4].

Taking the logarithm of (2) we find that we have to minimize the expression:

$$\sum_{j=1}^{B} \rho(x_i[j], m[j]) \qquad i = 1, ..., N \tag{3}$$

According to (1), ρ depends only on the difference between its two arguments. We can replace (3) with:

$$\sum_{j=1}^{B} \rho(n[j]) \qquad n[j] = x_i[j] - m[j] \tag{4}$$

To analyze the behavior of the estimate we take the approach described in [11] based on *influence function*. The influence function characterizes the bias

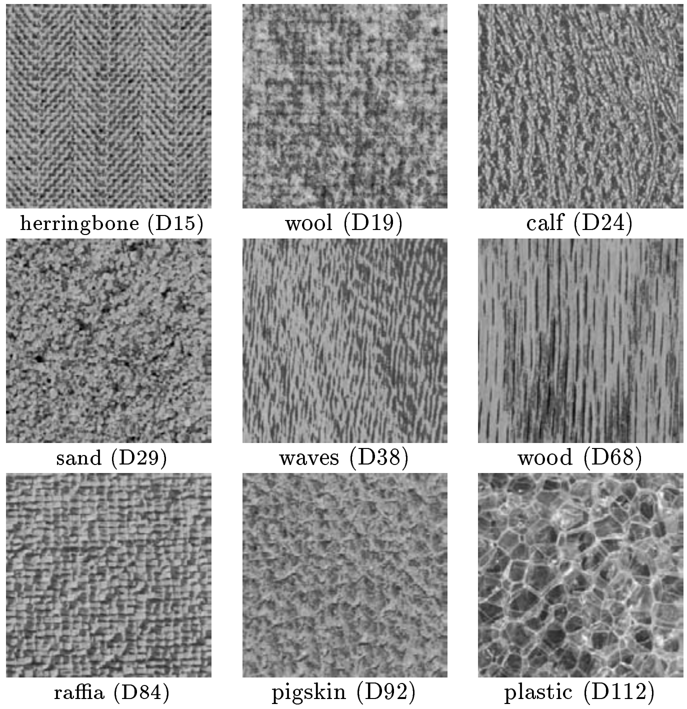

Figure 1: Brodatz textures

that a particular measurement has on the solution and is proportional to the derivative, ψ, of the estimate [1]:

$$\psi(n) \equiv \frac{d\rho(n)}{dn} \tag{5}$$

In case the noise is Gaussian distributed:

$$Prob\{n\} \sim \exp(-[n]^2) \tag{6}$$

then

$$\rho(n) = n^2 \qquad \psi(n) = n \tag{7}$$

If the errors are distributed as a *double* or *two-sized exponential*, namely

$$Prob\{n\} \sim \exp(-|n|) \tag{8}$$

then, by contrast,

$$\rho(n) = |n| \qquad \psi(n) = sgn(n) \tag{9}$$

In this case, the best match is obtained by minimizing the *mean absolute deviation*, rather than the *mean square deviation*. Here the tails of the distribution, although exponentially decreasing, are asymptotically much larger than any corresponding Gaussian.

For Gaussian distributed errors, equation (7) assigns greater weights to greater deviations. By contrast, when the tails are more prominent as in (8), then (9) assigns the same relative weight to all deviant points, with only the sign information used.

In summation, one can note that equation (7) resembles the L_2 metric while equation (9) resembles the L_1 metric. Thus, *maximum likelihood* gives a direct connection between the noise distribution and the comparison metrics. Considering ρ as the negative logarithm of the probability density of the noise then the corresponding metric is given by equation (3).

3 EXPERIMENTS

In our experiments, nine classes of textures - herringbone, wool, calf, sand, waves, wood, raffia, pigskin and plastic - taken from Brodatz's album [2] were used (Figure 1). The texture images were normalized to have the same mean and standard deviation in order to avoid gray-level bias which is unrelated to the image texture. The test samples were obtained by randomly subsampling the original texture images. 1000 subsamples of 32×32 or 16×16 pixels in size were extracted from every texture class, resulting in a classification of 9000 random samples in total. Regarding implementation, we used the same number of bins for the texture classification methods as in the survey by Ojala [9]. In the case of Trigrams, 512 bins were used.

3.1 Distribution Analysis

From the maximum likelihood paradigm, the first critical step is to determine the real noise distribution according to Eq. (1). Considering m the feature vector corresponding to a texture class M and x_i the feature vector corresponding to the sample block i extracted from M then, the real noise distribution is seen as the normalized histogram of differences between the elements of the two vectors x_i and m. We extract 10000 samples blocks from each texture class and accumulate the differences in a 1000 bins histogram.

The next step is to determine the distortion between the real noise distribution and the distributions associated with the L_1 and L_2 distance measures, namely, the double exponential distribution and the Gaussian distribution. We present the quantitative results in Table 1. From this table is clear that L_1 has a lower modeling error than L_2 and therefore, L_1 is a more appropriate distance measure than L_2 regarding modeling the noise distribution. Furthermore, we visually display in Figure 2 (a) and (b) the average similarity noise distribution for single feature matched by the best fit exponential and best fit Gaussian, respectively. We can further conclude that the noise distribution is not Gaussian as assumed with regard to the L_2 measure.

For the feature pairs, we show the numerical modeling errors in Table 2 and the average noise distribution along with the best fit exponential and Gaussian in Figure 2 (c) and (d), respectively. The results are consistent with the single feature tests.

	LBP	DIFFX	DIFFY	DIFF2	DIFF4	L3E3	E3L3	L3S3	S3L3	SCOV	SRAC	Trig
L_1	0.052	0.025	0.029	0.05	0.078	0.023	0.016	0.019	0.021	0.011	0.031	0.05
L_2	0.065	0.041	0.042	0.068	0.09	0.03	0.035	0.039	0.042	0.03	0.037	0.06

Table 1: The approximation error for the corresponding noise distribution using single features

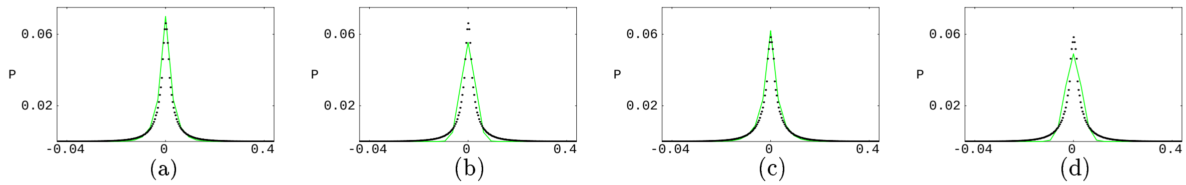

Figure 2: Average similarity noise distribution for single feature (a),(b) and for pairs of features (c),(d): (a),(c) best fit exponential, (b),(d) best fit Gaussian.

3.2 Misdetection Rates

The next step is to determine the real misdetection rates from Brodatz's test database. The misdetection rate is the percentage of misclassified texture blocks. For this test, we chose L_1 and L_2 because of their ease in implementation and overall ubiquity in current research. As a benchmark, we also compared them to the Kullback discriminant because the Kullback discriminant is frequently used in the pattern recognition literature and is a generally well known distortion measure.

In Figure 3, we display the misdetection rates of the L_1, L_2, and Kullback distortion measures versus the sample size for each of the texture distribution features. Note that the L_1 measure consistently yields lower misdetection rates, which agrees with the maximum likelihood paradigm.

Since the Trigrams require a threshold, this parameter affects the performance of the method. For the tests, we used the optimal threshold as selected from Figure 4.

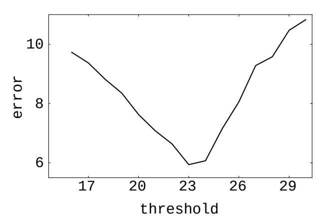

Figure 4: Trigrams error rate for different threshold values using L_1 and 32 x 32 sample

Regarding the feature pairs, the misdetection rates are shown in Figure 5. These results are consistent with the single feature tests.

3.3 Ideal Distribution

The main advantage of L_1 and L_2 is the ease in implementation and analytic manipulation. However, neither distance measure models the real noise distribution accurately, so we expect that we can lower the misdetection rates further. Using equations (2)-(4), we derived from the real noise distribution a distortion measure within the maximum likelihood paradigm, which we denote as the ML distortion measure. The ML distortion measure is directly related to the real noise distribution which is a discrete distribution with known points. Consider that we have to compare two vectors then, for each difference value between corresponding elements we have to calculate according to Eq. (3) the negative logarithm of the probability density of the real noise in that point. Since the distribution is discrete, the value of the probability in any arbitrary point is calculated by using interpolation between the two known adjacent probability values. The sum of all values calculated in this way resembles the ML distortion measure.

In Figures 3 and 5 we displayed the misdetection rates between the ML distortion measure and the other distance measures. Note that the ML distortion measure consistently has lower misdetection rates for all sample sizes.

	L_1	L_2	K	ML	RG
LBP	1.98	2.42	3.05	1.51	24
DIFFX	7.14	12.71	8.37	4.04	44
DIFFY	8.47	12.40	10.05	4.67	45
DIFF2	6.02	10.73	7.56	4.91	19
DIFF4	6.57	10.87	7.12	3.62	45
L3E3	9.95	13.73	10.62	8.75	13
E3L3	18.47	22.12	19.33	15.24	18
L3S3	9.65	13.93	10.66	8.48	13
S3L3	15.49	17.77	15.64	13.12	16
SCOV	8.71	11.26	10.38	7.29	17
SRAC	11.18	13.38	11.85	9.83	13
Trig	5.94	13.02	7.30	3.27	45

Table 3: Error rates for single features considering 32 x 32 samples. The last column represent the *relative gain* (RG) in % obtained using the ML distortion measure in comparison with the best of the other measures (L_1)

	L_1	L_2	K	ML	RG
LBP/C	0.04	0.76	0.28	0.02	50
LBP/SCOV	0.35	1.27	0.88	0.12	66
DIFFX/DIFFY	1.91	4.91	2.32	1.05	46
DIFFY/SCOV	1.12	2.24	1.48	0.83	26
L3E3/E3L3	0.92	1.23	1.12	0.58	37
L3S3/S3L3	0.42	1.56	1.03	0.31	27

Table 4: Error rates for pairs of features considering 32 x 32 samples. The last column represent the *relative gain* (RG) in % obtained using the ML distortion measure in comparison with the best of the other measures (L_1)

In order to give a quantitative value for the improvement in accuracy introduced by the ML distortion measure we define the *relative gain* as being:

$$RG = 1 - \frac{err_{ML}}{err_{MIN}}$$

	LBP/C	LBP/SCOV	DIFFX/DIFFY	DIFFY/SCOV	L3E3/E3L3	L3S3/S3L3
L_1	0.045	0.044	0.026	0.019	0.015	0.018
L_2	0.05	0.064	0.041	0.037	0.031	0.039

Table 2: The approximation error for the corresponding noise distribution using pairs of features

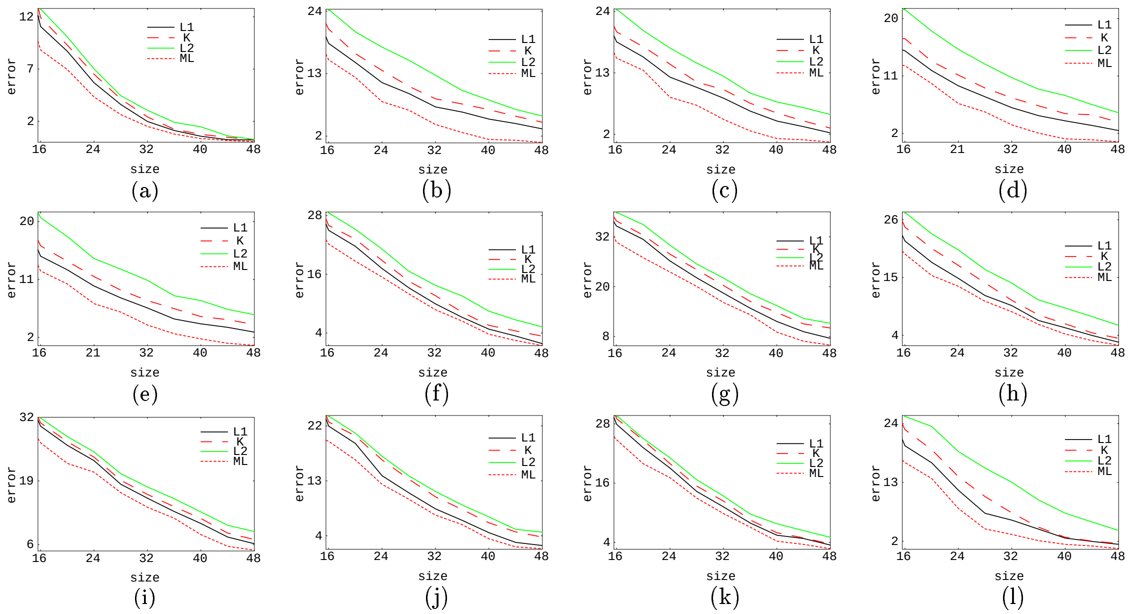

Figure 3: Error rates for single features as function of sample size using L_1 (L1), L_2 (L2), the Kullback discriminant (K) and the optimal maximum likelihood distortion measure (ML): (a) LBP, (b) DIFFX, (c) DIFFY, (d) DIFF2, (e) DIFF4, (f) L3E3, (g) E3L3, (h) L3S3, (i) S3L3, (j) SCOV, (k) SRAC, (l) Trigram.

where err_{ML} denote the error rate obtained using the ML distortion measure and err_{MIN} is the minimum error obtained using all the other distortion measures.

Tables 3 and 4 summarize the results for single and feature pairs across the L_1, L_2, K and ML distortion measures. Note that combining complementary features results in significantly lower error rates.

Overall, for the single features LBP had the least error rate and for the feature pairs LBP/C and LBP/SCOV provide the best results. The *relative gain* obtained by using the ML distortion measure has significant values.

4 DISCUSSION AND CONCLUSIONS

This research is differentiated from the previous surveys in texture classification in that we have investigated the role of the underlying noise distribution and corresponding metric in the paradigm of maximum likelihood. Our experiments on both the noise distribution and the misdetection rates from using a particular distortion measure provide strong evidence of the maximum likelihood theory.

Most of the pattern recognition literature uses the Kullback discriminant or the sum of squared distance (L_2). In the maximum likelihood paradigm, it is provable that the Gaussian distribution results in the L_2 metric, and the exponential distribution results in the L_1 metric. By linking the distributions with the metrics, we can directly show why a particular metric would outperform another metric. Specifically, the metric which will have the least misdetection rate should be the metric whose distribution best matches the real noise distribution from the test set.

Here in this paper, we have found that the noise distribution is modeled better by the exponential than the Gaussian distribution. Consequently, among the analytic distortion measures, L_1 consistently had a lower misdetection rate than L_2. Furthermore, L_1 had a lower misdetection rate that the Kullback discriminant. Note that the exponential distribution always had a lower modeling error than the Gaussian distribution, and that results in lower misdetection for L_1 rates than the L_2 and Kullback. This is interesting because we had expected that on some of the textures, L_2 would have had a lower error rate. One interesting hypothesis is that the distortion in textures is exponential, not Gaussian.

Given that the modeling of the real noise distribution is linked with the misdetection rate, the next logical question is, "What is the misdetection rate when we directly model the real noise distribution?" It is also validated that the lowest misdetection rate occurs when we use an approximate, quantized model for the real noise distribution. The corresponding distortion measure clearly outperforms the rest of the distortion measures as shown in Figures 3 and 5.

Regarding completeness we have given the absolute error rates. We have also provided one of the possible measures of improvement denoted as *relative gain*. This measure reflects the significance of the ML distortion measure in comparison with the best of the other measures. It should be noted that the real significance of a change in error rate can only be made with regard to a particular application - whether the acceptable error rate is 1 in a hundred or thousand.

In the summary Figure 6 we show the average comparative results for L_1, L_2, K and ML for single and complementary feature pairs.

The maximum likelihood paradigm also indicates

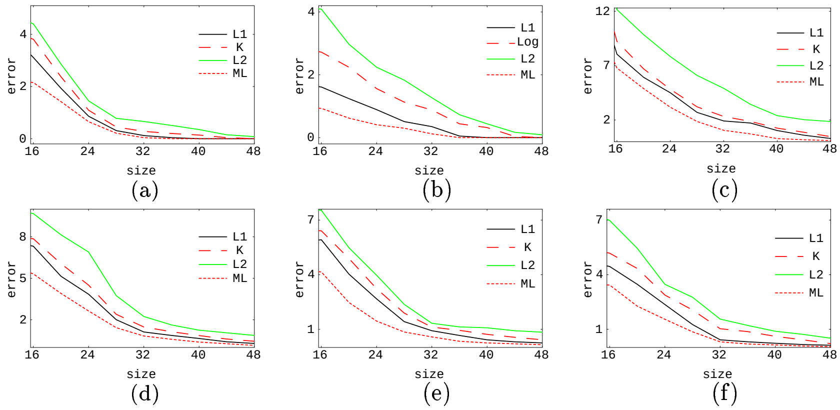

Figure 5: Error rates for pairs of features as function of sample size using L_1 (L1), L_2 (L2), the Kullback discriminant (K) and the optimal maximum likelihood distortion measure (ML): (a) LBP/C, (b) LBP/SCOV, (c) DIFFX/DIFFY, (d) DIFFY/SCOV, (e) L3E3/E3L3, (f) L3S3/S3L3.

the breaking points. In the case of the L_1 metric, the distribution should be exponential. For the L_2 metric, the distribution should be Gaussian. In our experiments, the real noise distribution was closer to exponential, but in other domains, the real noise distribution could be Gaussian or some other distribution. The critical step is to measure the real noise distribution before applying a metric.

This gives rise to several conclusions from our maximum likelihood investigation on texture classification.

- If it is necessary to perform analytic computations on the metric or no ground truth is available, then L_1 is the clear choice for analytic equation manipulations. Furthermore, L_1 has the most simplicity, and therefore is easiest to implement.
- If minimizing the misdetection rate is of primary importance and it is possible to create or find test sets, then the best direction is to measure the real distribution and convert it to a distortion measure using maximum likelihood theory.

Thus, the primary contribution of this paper is the theoretical link between the noise distribution and the corresponding distortion measure. Our research shows that any texture classification method can be optimized by including information about the real noise distribution.

In future work, we plan to examine different analytic models for representing the real distribution. We would like to find or derive a parameterized model for which it is also possible to derive the partial derivatives. This would be beneficial for many theoretical proofs and applications in robust estimation.

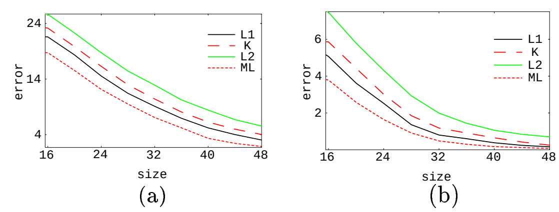

Figure 6: Average error rates for single feature (a) and for pairs of features (b) using: L_1 (L1), L_2 (L2), the Kullback discriminant (K) and the optimal maximum likelihood distortion measure (ML)

References

[1] M.J. Black. *Robust Incremental Optical Flow*. PhD thesis, Yale University, September 1992.

[2] P. Brodatz. *Textures: A Photographic Album for artists and designers*. Dover Publications, 1966.

[3] D. Harwood, T. Ojala, M. Pietikainen, S. Kelman, and L. Davis. Texture classification by center-symmetric auto-correlation using Kullback discrimination of distributions. *Pattern Recognition Letters*, 16:1–10, 1995.

[4] P.J. Huber. *Robust Statistic*. Wiley, 1981.

[5] D.P. Huijsmans, S. Poles, and M.S. Lew. 2D pixel trigrams for content-based image retrieval. In *Proc. 1st IDB-MMS A'dam*, pages 139–145, 1996.

[6] S. Kullback. *Information theory and statistics*. Dover Publications, 1968.

[7] K.I. Laws. Textured energy measures. In *Proc. of Image Understanding Workshop*, pages 47–51, 1980.

[8] P. Ohanian and R. Dubes. Performance evaluation for four classes of textural features. *Pattern Recognition*, 25:819–833, 1992.

[9] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distribution. *Pattern Recognition*, 29:51–59, 1996.

[10] T. Reed and J. Du Buf. A review of recent texture segmentation and feature extraction techniques. *CVGIP*, 57(3):359–373, 1993.

[11] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outliers Detection*. Wiley, 1987.

[12] M. Unser. Sum and difference histograms for texture classification. *IEEE Trans. Pattern. Anal Mach. Intell.*, 8(1):118–125, 1986.

[13] L. Van Gool, P. Dewaele, and A. Oosterlinck. Texture analysis. *CVGIP*, 29(3):336–357, 1985.