Performance Evaluation of Texture Measures with Classification Based on Kullback Discrimination of Distributions

Timo Ojala[1], Matti Pietikäinen[1], and David Harwood[2]

[1]Department of Electrical Engineering
University of Oulu
FIN-90570 Oulu, Finland

[2]Center for Automation Research
University of Maryland
College Park, MD 20742, USA

Abstract

This paper evaluates the performance both of some texture measures which have been successfully used in various applications and of some new promising approaches proposed recently. For classification a method based on Kullback discrimination of sample and prototype distributions is used. The classification results for single features with one-dimensional feature value distributions and for pairs of complementary features with two-dimensional distributions are presented.

1. Introduction

Texture is an important characteristic for the analysis of many types of images. A wide variety of measures for discriminating textures have been proposed [1] [2]. Comparative studies to evaluate the performance of some texture measures have been carried out in [3][4][5], for example.

Most of the approaches to texture analysis quantify the texture measures by single values (means, variances etc.). These values are used as elements of feature vectors in performing classification. In this way much important information contained in the distributions of feature values might be lost. Some earlier studies also suggest that distributions of joint occurrences of pairs of features give better results than distributions of single features on their own.

This paper evaluates the performance both of some texture measures which have been successfully used in various applications and of some new promising approaches proposed recently. For classification a method based on Kullback discrimination of sample and prototype distributions is used. The classification performances for single features with one-dimensional feature value distributions and for pairs of complementary features with two-dimensional distributions are evaluated. Two different types of data sets are used in experiments: the same Brodatz's images as used by Harwood et al. [6], and images used in a recent comparative study by Ohanian and Dubes [5].

2. Texture Measures Used in This Study

Gray Level Difference Method

A class of local properties based on differences between pairs of gray levels or of average gray levels has been sometimes used in texture analysis [3][7]. Our feature set contained four measures based on the gray level difference method: DIFFX and DIFFY are histograms of absolute gray level differences between neighboring pixels computed in horizontal and vertical directions, respectively, while DIFF2 accumulates absolute differences in horizontal and vertical directions and DIFF4 in all four principal directions, respectively, in a single histogram, providing rotation invariant texture measures.

Laws' Texture Measures

The "texture energy measures" developed by Laws [8] or related measures developed by others have been used in various applications. Four Laws' 3x3 operators were considered in this study. L3E3 and E3L3 perform edge detection in vertical and horizontal directions, respectively, and L3S3 and S3L3 are line detectors in these two orthogonal directions.

Center-symmetric Covariance Measures

Laws' and other related studies of texture analysis suggest that many natural and artificial textures are measurably "loaded" with distributions of various specific local patterns of texture having these abstract symmetrical forms. Moreover, to measure the local "loading" of gray level symmetric (positive) or antisymmetric (negative) texture, we have only to compute local auto-covariances or auto-correlations of center-symmetric pixel values of suitably sized neighborhoods. In a recent study of Harwood et al. [6], a set of related measures was introduced, including two local center-symmetric auto-correlation measures, with linear (SAC) and rank-order versions (SRAC), together with a related covariance measure (SCOV). We applied these three measures in the present work.

Local Binary Patterns

Recently, Wang and He [9] introduced a new model of texture analysis based on the so-called texture unit, where a texture image can be characterized by its texture spectrum. A texture unit (TU) is represented by eight elements, each of which has one of three possible values (0,1,2) obtained from a neighborhood of 3x3 pixels. In total, there are 3^8 = 6561 possible texture units describing spatial three-level patterns in a 3x3 neighborhood. The occurrence of distribution of texture units computed over a region is called the texture spectrum.

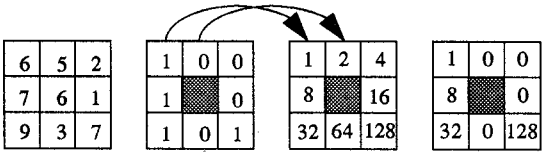

Fig. 1. Two-level version (LBP) of the texture unit.

In this study we propose to use a two-level version of the method of Wang and He. It provides a robust way for describing pure local binary patterns (LBP) in a texture. In the two-level version, there are only 2^8= 256 possible texture units instead of 6561. In binary case, the original 3x3 neighborhood (Fig. 1a) is thresholded by the value of the center pixel. The values of the pixels in the thresholded neighborhood (Fig. 1b) are multiplied by the weights given

1051-4651/94 $04.00 © 1994 IEEE

to the corresponding pixels (Fig. 1c). The result for this example is shown in Fig. 1d. Finally, the values of the eight pixels are summed to obtain the number (169) of this texture unit. LBP method is 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 (see Fig. 1b).

Complementary Feature Pairs

In most cases a single texture measure cannot provide enough information about the amount and spatial structure of local texture. Better discrimination of textures should be obtained by considering joint occurrences of two or more features. As an example of this kind of approach, the spatial gray level dependence method (co-occurrence method) estimates the joint gray level distribution for two gray levels located at a specified distance and angle. Shen and Bie [10] considered the use of gradient magnitude and direction, gray level and gradient direction, and gray level and gradient magnitude, respectively, jointly in their feature frequency matrix (FFM) scheme.

Our primary goal was to find pairs of such features which provide complementary information about textures, using a two-dimensional Kullback classification scheme. Complementary features should provide more or less uncorrelated texture information.

The center-symmetric covariance measures are abstract, measuring covariances of any local center-symmetric patterns. They provide robust information about the amount of local texture, but very little about exact local spatial patterns. This immediately suggests that we should consider texture analysis of pure spatial patterns, which would complement the analysis. The local binary patterns (LBP) were chosen for this purpose. Two different features were combined with LBP. LBP/C is based on the contrast measure introduced earlier and the other pair is LBP/SCOV. The contrast and SCOV measures were quantized to have only eight values to make sure that the two-dimensional feature histograms would not become too sparse.

Laws' masks chosen for this study perform edge or line detections in horizontal or vertical directions. These kinds of patterns can occur, however, in arbitrary directions, which suggests that a joint use of edge or line detectors in the orthogonal directions should be considered. In a similar way, difference histograms are usually computed for displacements in horizontal or vertical directions, and a joint use of these orthogonal directions should provide useful information for texture discrimination.

The pair L3E3/E3L3 corresponds edge detection, L3S3/S3L3 line detection and DIFFX/DIFFY absolute gray scale differences in the two orthogonal directions, respectively. DIFFY/SCOV combines absolute gray scale differences in vertical direction with the center-symmetric covariance measure SCOV. GMAG/GDIR is using gradient magnitudes and directions obtained by a 3x3 Sobel edge operator. All of these feature combinations were quantized into 32x32 bins.

3. Nearest-neighbor Classification Using Kullback Discrimination

In experiments with Image Set I, the classification of a sample was based on comparing 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 optimized Kullback's minimum cross-entropy principle (1) [11]. Here s and m are the sample and model distributions, n is the number of bins and s_i, m_i are the respective sample and model probabilities at bin i. This (pseudo-) metric measures likelihoods that samples are from alternative texture classes, based on exact probabilities of feature values of pre-classified texture prototypes.

$$D(s{:}m) = \sum_{i=1}^{n} s_i \log \frac{s_i}{m_i} \qquad (1)$$

The model distribution for each class was obtained by scanning the gray-scale corrected 256x256 texture image with the local texture operator. The distributions of local statistics were divided into histograms having a fixed number of bins; hence, the Kullback's cross-entropy measure had the same number of degrees-of-freedom for every pairing of a sample and a model. The number of bins used in quantization of the feature space plays a crucial role. Histograms with too modest a number of bins fail to provide enough discriminative information about the distributions. However, since the distributions have a finite amount of entries, it does not make sense to go to the other extreme. If histograms have too many bins, and the average number of entries per bin is very small, histograms become sparse and unstable. In most of our experiments, histograms with 32 bins were used. This corresponds to an average number of 32 entries per bin for samples 32x32 in size and 8 entries per bin for samples 16x16 in size, respectively.

The feature space was quantized by adding together feature distributions for every single model image in a total distribution which was divided into 32 bins having an equal number of entries. Hence, the cut values of the bins of the histograms corresponded to 3.125 (100 / 32) percentile of the combined data. Deriving the cut values from the total distribution and allocating every bin the same amount of the combined data guarantees that the highest resolution of the quantization is used where the number of entries is largest and vice versa. It should be noted that the quantization of feature space is only required for texture operators with a continuous-valued output. Output of some discrete operators like LBP, where two successive values can have totally different meaning, does not require any further processing; operator outputs are just accumulated into a histogram. The empty bins were set to one.

In experiments with Image Set II, a single model distribution for every class was not used as with the Brodatz's images. Every sample was in its turn classified using the other samples as models, hence the leave-one-out approach was applied. The sample was assigned the label of the model that minimized two-way test-of-independence (2) that is a modification from Kullback's criterion [12]:

$$G = 2\left(\left[\sum_{s,m}\sum_{i=1}^{n} f_i \log f_i\right] - \left[\sum_{s,m}\left(\sum_{i=1}^{n} f_i\right)\log\left(\sum_{i=1}^{n} f_i\right)\right] - \left[\sum_{i=1}^{n}\left(\sum_{s,m} f_i\right)\log\left(\sum_{s,m} f_i\right)\right] + \left[\left(\sum_{s,m}\sum_{i=1}^{n} f_i\right)\log\left(\sum_{s,m}\sum_{i=1}^{n} f_i\right)\right]\right) \qquad (2)$$

where s, m are the two texture samples (test sample and model), n is the number of bins and f_i is the frequency at bin i.

To compare distributions of complementary feature pairs, metrics D and G were extended in a straightforward manner to scan through the two-dimensional histograms. If quantization of the feature space was required, it was done separately for both features using the same approach as with single features. Therefore the two-dimensional distribution was likely to have bins with zero entries. Regarding the stability of the classification process it was important to handle these empty bins correctly. Setting every empty

bin to one turned out to be a good solution.

4. Experiments with Image Set I

In these experiments, nine classes of textures - grass, paper, waves, raffia, sand, wood, calf, herringbone and wool - taken from Brodatz's album [13] were used. The same set of images was used by Harwood et al. [6]. The texture images were corrected by mean and standard deviation in order to minimize discrimination by overall gray-level variation which is unrelated to local image texture. The correction was applied to the whole 256x256 images instead of correcting every sample window separately. The mean gray value of each corrected image was set to 256 and the standard deviation to 40.

The test samples were obtained by randomly subsampling the original texture images. 1000 subsamples of 32x32 or 16x16 pixels in size were extracted from every texture class, resulting in a classification of 9000 random samples in total. When classifying a particular sample, the sample distribution was subtracted from the model distribution of the true class of this sample so that an unbiased error estimate was obtained.

Table 1 shows the classification error rates for 32x32 and 16x16 samples from nine 256x256 images representing nine texture classes.

Table 1. Error rates for single features with Image Set I.

Feature	32x32	16x16	Feature	32x32	16x16
LBP	2.30	12.52	E3L3	17.62	39.58
DIFFX	3.04	14.31	L3S3	14.28	33.78
DIFFY	3.30	12.84	S3L3	7.58	23.68
DIFF2	8.43	13.50	SCOV	8.07	29.62
DIFF4	8.73	14.32	SAC	11.83	36.92
L3E3	19.82	42.46	SRAC	8.46	32.77

Best performance is obtained for the local binary pattern (LBP) feature. The difference histogram features also performed very well. The covariance measures perform better than Laws' measures, but the error rates for the 16x16 samples are quite poor for these two related approaches. This indicates that for good results the covariance and Laws' measures require larger sample sizes than the other approaches considered here.

Table 2 presents the results for pairs of complementary measures. LBP/C achieves very low error rates. The results for LBP/SCOV are nearly as good and very good results are also obtained with DIFFX/DIFFY and L3S3/S3L3 features. The poorest performance was obtained for GMAG/GDIR. After comparing Tables 1 and 2 we can conclude that pairs of complementary features give significantly better results than single features. The small error rates obtained for small 16x16 samples demonstrate the power of our classification scheme.

Table 2. Error rates for pairs of features with Image Set I.

Features	# bins	32x32	16x16
LBP/C	256 x 8	0.19	1.90
LBP/SCOV	256 x 8	0.23	2.60
L3E3/E3L3	32 x 32	0.89	12.18
L3S3/S3L3	32 x 32	0.22	4.00
DIFFX/DIFFY	32 x 32	1.51	3.89
DIFFY/SCOV	32 x 32	0.86	6.57
GMAG/GDIR	32 x 32	1.81	16.13

5. Experiments with Image Set II

Recently, Ohanian and Dubes [5] studied the performance of four types of features: Markov Random Field parameters, Gabor multi-channel features, fractal-based features, and co-occurrence features. They used four classes of images in experiments: fractal images, Gaussian Markov Random Field (GMRF) images, leather images, and painted images. Each image class contained four types of images: the synthetic fractal and GRMF images were generated by using different parameter values and the natural images represented different types of leather or painted surface, respectively. With these images four 4-class problems and a 16-class problem were established. Whitney's forward selection method was used for feature selection and a kNN (k=9) decision rule for classification. The co-occurrence features generally outperformed other features followed by fractal features.

In our study, the same set of images was used. The images were again corrected by mean and standard deviation as described in section 4. Like Ohanian and Dubes, we treated each of the four classes of images as a separate 4-class problem. 200 non-overlapping samples of size 32x32 were extracted from each image (texture type), resulting in a classification problem of 800 samples. Every sample was in its turn classified using the other 799 samples as models (3199 models in the 16-class problem) by applying a two-way-of-independence G test with the leave-one-out approach. A 9-NN classification resembling the principle of Ohanian and Dubes was used. Additionally, the 16-class problem, including all 3200 non-overlapping samples extracted from all 16 images was considered.

The error rates for the four 4-class problems and the 16-class problem using distributions of single features are summarized in Table 3.

Table 3. Error rates for single features with Image Set II.

Feature	Fractal	GMRF	Leather	Painted	16-class
LBP	37.75	0.00	24.12	10.75	18.16
DIFFX	0.00	10.62	2.25	6.25	5.75
DIFFY	0.00	10.50	0.00	2.88	3.72
DIFF2	0.00	7.88	0.12	1.75	2.81
DIFF4	0.00	14.50	0.00	0.88	3.69
L3E3	1.00	6.88	23.12	40.62	33.06
E3L3	2.00	26.50	27.25	49.25	36.09
L3S3	0.00	7.38	39.25	50.12	31.53
S3L3	0.00	13.25	36.38	22.00	27.22
SCOV	0.00	3.38	28.38	31.37	16.56
SAC	43.25	10.88	46.12	34.00	46.28
SRAC	12.75	2.12	34.75	12.00	18.25

It can be seen that fractal images were quite easy to discriminate with most of the features. However, gray scale invariant features LBP, SAC and SRAC did not perform well. This indicates that gray scale contrast is important for discriminating these images.

LBP, SRAC and SCOV performed best for GMRF images. The worst results for the difference histogram features were obtained with these images. The classification of leather images is somewhat more difficult. The excellent results obtained with DIFFY, DIFF4 and DIFF2 are surprisingly good. The results with the other features are much poorer. The painted surfaces were also difficult for some features to discriminate. DIFF4, DIFF2 and DIFFY, again, achieved very low error rates. DIFFX, LBP and SRAC also performed quite well, but the results for the other features are much poorer.

The difference histogram features performed best in the 16-class problem, too. SCOV, LBP and SRAC also performed quite well. The poor performance of Laws' fea-

tures is mainly caused by the difficulty of discriminating leather and paint images with these measures. SCOV and SRAC performed better than Laws' measures, but, as expected, they also had problems with leather and painted images.

The classification results for pairs of features are presented in Table 4. The best overall performance is for DIFFY/SCOV, and nearly as good results were obtained by DIFFX/DIFFY. The performance of GMAG/GDIR was poorest. In general, the improvement achieved with pairs of features compared to the best results for single features is not as significant for this image set as for Brodatz's textures.

Table 4. Error rates for pairs of features with Image Set II.

Features	# bins	Fractal	GMRF	Leather	Paint	16-class
LBP/C	256 x 8	0.00	1.75	29.62	23.12	12.38
LBP/SCOV	256 x 8	0.00	1.25	30.38	32.62	14.97
L3E3/E3L3	32 x 32	0.88	7.00	16.38	43.62	29.66
L3S3/S3L3	32 x 32	0.00	0.75	11.50	15.00	11.81
DIFFX/DIFFY	32 x 32	0.12	5.12	0.25	1.12	3.44
DIFFY/SCOV	32 x 32	0.12	4.75	1.75	1.62	2.88
GMAG/GDIR	32 x 32	1.88	9.38	35.25	44.12	35.79

6. Discussion and Conclusions

Most of the earlier approaches to texture classification quantify the texture measures by single values. The very good results that we obtained by using distributions of simple texture measures suggest that the distributions of feature values should be used instead of single values.

The gray level difference method achieved the best overall performance discriminating most of the textures very well. It is very easy to compute, and it performed well even with small 16x16 samples, which make this approach very attractive for many applications, including texture classification and segmentation.

The texture measures based on local binary patterns are also computationally extremely simple. These measures performed very well, especially with the Brodatz textures. LBP is gray scale invariant and can be combined with a simple contrast measure to make it even more powerful. The method is rotation variant which is undesirable in certain appications, but it should be possible to derive rotation invariant versions of LBP.

Center-symmetric covariance features performed very well for some of the textures being more powerful than Laws' measures. Both of these approaches require larger sample sizes than the gray level difference and LBP methods. A reason for this is that most of the discriminative information for these kinds of measures is contained in the match maxima, as shown by Pietikäinen et al. [14]. To get statistically reliable information on the distributions of local extrema, quite large image windows may be needed. Covariance measures are computationally more complex than Laws' measures, but they are all rotation invariant while Laws' measures are not. In addition, the SRAC measure, which performed quite well, is invariant under any monotonic transformation including correction by mean and standard deviation and histogram equalization.

There are many applications, for example in industrial inspection and remote sensing, in which the gray-scale invariance of a texture measure is of great importance due to uneven illumination or great within-class variability. Recent results which we have obtained with applying texture classification to a difficult metal sheet inspection problem have demonstrated that the gray-scale invariant LBP and SRAC measures can be more powerful in such applications than the other approaches considered in this paper [15].

The quite poor performance of Laws' approach indicates that the discriminative power of these measures is mostly contained in the variances of the feature distributions which have been used in most of the earlier studies. The whole distribution does not seem to provide much additional information.

The use of pairs of complementary measures generally improves the classification accuracy. This was particularly evident in the case of Brodatz's textures. The computationally simple DIFFX/DIFFY and LBP/C pairs were among the best feature pairs. The performance of Laws' approach can also be significantly improved by using pairs of orthogonal masks instead of single features. The gradient magnitude/direction pair did not perform as well as the other pairs in our experiments.

7. Acknowledgements

The financial support provided by the Technology Development Center of Finland and the Academy of Finland is gratefully acknowledged. We also wish to thank Prof. Richard C. Dubes and John Lees from the Michigan State University for providing a set of test images used in this study.

8. References

[1] Van Gool, L., P. Dewaele and A. Oosterlinck. Texture analysis anno 1983. Computer Vision, Graphics and Image Processing, Vol. 29, No. 3, pp. 336-357, 1985.

[2] Haralick, R.M. and L. Shapiro. Computer and Robot Vision, Vol. 1, Addison-Wesley, 1992.

[3] Weszka, J., C. Dyer and A. Rosenfeld. A comparative study of texture measures for terrain classification. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-6, pp. 269-285, 1976.

[4] Du Buf, J. M.H., M. Kardan and M. Spann. Texture feature performance for image segmentation. Pattern Recognition, Vol. 23, No. 3/4, pp. 291-309, 1990.

[5] Ohanian, P.P. and R.C. Dubes. Performance evaluation for four classes of textural features. Pattern Recognition, Vol. 25, pp. 819-833, 1992.

[6] Harwood, D., T. Ojala, M. Pietikäinen, S. Kelman and L.S. Davis. Texture classification by center-symmetric auto-correlation, using Kullback discrimination of distributions. University of Maryland, Center for Automation Research, Technical Report CAR-TR-678, 1993.

[7] Unser, M. Sum and difference histograms for texture classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 1, pp. 118-125, 1986.

[8] Laws, K.I. Textured image segmentation. Report 940, Image Processing Institute, Univ. of Southern California, 1980.

[9] Wang, L. and D.C. He. Texture classification using texture spectrum. Pattern Recognition, Vol. 23, pp. 905-910, 1990.

[10] Shen, H.C. and C.Y.C. Bie. Feature frequency matrices as texture image representation. Pattern Recognition Letters, Vol. 13, No. 3, pp. 195-205, 1992.

[11] Kullback, S. Information Theory and Statistics, Dover Publications, New York, 1968.

[12] Sokal, R.R. and F.J. Rohlf. Biometry. W.H. Freeman and Co, 1969.

[13] Brodatz, P. Textures: A Photographic Album for Artists and Designers. Dover Publications, New York, 1966.

[14] Pietikäinen, M., A. Rosenfeld and L.S. Davis. Experiments with texture classification using averages of local pattern matches. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-13, No. 3, pp. 421-426, 1983.

[15] Pietikäinen, M., T. Ojala, J. Nisula, J. Heikkinen. Experiments with two industrial problems using texture classification based on feature distributions. SPIE Vol. 2354 Intelligent Robots and Computer Vision XIII: 3D Vision, Product Inspection and Active Vision, Boston, Mass., 31.10.-4.11.1994, 8 p.