Perceptual Texture Retrieval Using Spatial Distributions of Textons (SDoT)

Xinghui Dong[*†], Junyu Dong[††], Shengke Wang[††], and Mike J. Chantler[†]
[†]The Texture Lab, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK
E-mail: {xd25, m.j.chantler}@hw.ac.uk
[††]Department of Computer Science, Ocean University of China, Qingdao, China
E-mail: {dongjunyu, neverme}@ouc.edu.cn

***Abstract*—It has been shown that the spatial information of local image characteristics is important to human perception and computational features. Inspired by these studies, we propose a set of new computational texture features based on the spatial distributions of textons (SDoT). First, gradient magnitude and gradient direction spectra are computed from a texture image. Second, the multiple gradient spectra simultaneous autoregressive (MGSSAR) models are estimated for each image. Both model coefficients and the variance of the model estimation error jointly construct a local feature space. Third, *k*-means is used to learn textons from the local features. All textons learned from a texture database are combined into a dictionary. Fourth, vector quantization is utilized to map a texture from the local feature space into the texton space. Finally, an aura matrix is computed from the texton map of each texture in order to encode the spatial distributions of the textons. The results of a perceptual texture retrieval experiment show that the proposed feature set performs more consistently with human observers than 56 existing feature sets. We attribute this to the fact that the proposed feature set encodes the spatial information of textons.**

I. INTRODUCTION

Field et al. [1] showed that human visual systems are able to exploit the continuity of the patterns in an image even though a grid has been overlaid on top of it. Dong et al. [2, 3] also examined 51 computational feature sets using two perceptual texture similarity estimation tasks and found that none of these use the (aperiodic) long-range interactions exploited by humans. In addition, the utilization of the spatial information of visual words was addressed in the literature [4]. Motivated by these studies, the aura matrix [5] is used to capture the spatial relationship between textons in this study.

As an intuitive characteristic, the edge (contour) has been found to play an important role in object identification [6]. Unfortunately, they cannot always be extracted from images accurately. In contrast, the gradient also encodes the structure information and is normally used for edge detection [7]. It has been observed that human visual systems exploit gradients as shape cues [8]. Furthermore, gradient magnitudes and/or gradient directions were used for image representation [9-12]. The former indicates how quickly one image is changing while the latter suggests the direction in which an image is changing most rapidly. Since gradient magnitudes and gradient directions encode different characteristics, and the higher order statistics calculated over three or more pixels are believed to be able to capture complicated spatial image structure [13], the higher order statistics of the joint distributions of gradient magnitudes and gradient directions are likely to provide powerful discrimination information. To the authors' knowledge, however, the exploitation of this type of feature has not been addressed so far.

The simultaneous autoregressive (SAR) model has been used for gray level texture analysis [14]. Multispectral SAR (MSAR) was also applied to color texture synthesis [15]. The MSAR captures the higher-order spatial relationship between the central pixel of a local neighborhood in an image spectrum and its neighboring pixels in both the same image spectrum and different spectra. In this study, we therefore adopt a multiple gradient spectra SAR (MGSSAR) model. The MGSSAR not only captures the joint distributions of gradient magnitude and gradient direction spectra but also encodes the higher order statistics between the central pixel and its neighboring pixels.

Considering the advantages of texton-based features [16], the texton dictionary learning technique is utilized to learn a set of textons from local MGSSAR features. Vector quantization is used to map the MGSSAR features into the texton space. Since texton histograms do not consider the spatial relationship between local features (or textons), an aura matrix [5] is computed from the texton map of each texture image in order to encode the spatial relationship between textons. The proposed feature set is compared with 56 existing feature sets in a perceptual texture retrieval task. The contributions of this paper are: (1) the joint use of gradient magnitude and gradient direction spectra to compute local higher order statistics; (2) the learning of textons from local MGSSAR features; and (3) the application of the aura matrix to encode the spatial distributions of textons.

The rest of this paper is organized as follows. Section 2 briefly describes the proposed method. The experimental setup and results are reported in section 3 and 4 respectively. Finally, our conclusions are drawn in section 5.

II. OUR APPROACH

We introduce a feature set based on the spatial distributions of the MGSSAR textons. We term it with the title of "SDoT" (Spatial Distributions of Textons). The MGSSAR models are estimated from gradient magnitude and gradient direction

spectra. Model coefficients and the variances of model estimation errors are quantized into a texton space and each texture is mapped into a texton map in this space. Finally, an aura matrix is computed from the texton map of each texture in order to encode the spatial distributions of the textons.

A. Computing Gradient Magnitudes and Directions

Given a texture image $f(x, y)$, the derivative maps: $f_x(x, y)$ and $f_y(x, y)$ in x and y directions are computed using the method utilized by Canny [7]. The gradient magnitude and gradient direction spectra (see Fig. 1) are calculated from $f_x(x, y)$ and $f_y(x, y)$.

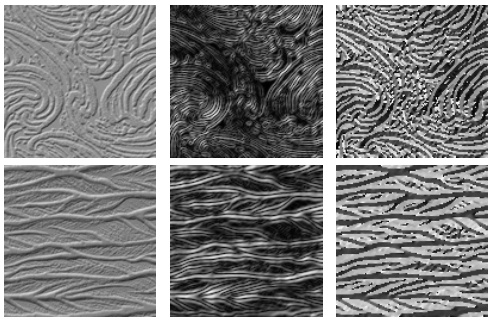


Fig. 1: The middle and right columns display the gradient magnitude and gradient direction spectrum maps of the texture images (top: "026", bottom: "067" in *Pertex* [17]) in the left column, respectively.

B. Estimating MGSSAR Models

The MSAR model [15] encodes the spatial interactions between the central pixel of one local neighborhood in an image spectrum and its neighboring pixels in the same spectrum as well as different image spectra. However, the estimation of the MSAR model based on small neighborhoods is probably under-determined. Thus, the estimation was conducted in a larger $M \times M$ moving-window in this research.

For a position (x, y) in an $M \times M$ (19×19 in this study) window in an image spectrum f_i, given that $f_i(x, y)$ has a zero mean inside the window, the MSAR is described as:

$$f_i(x, y) = \sum_{j=1}^{S} \sum_{k=1}^{N} c_{ijk} * f_j(x + \Delta x_k, y + \Delta y_k) + \sqrt{\rho_i}\omega_i(x, y), \quad (1)$$

where S is the number of image spectra, $i = 1, \ldots, S$, $c_{ij1} \ldots c_{ijN}$ are N MSAR coefficients which encode the relationship of $f_i(x, y)$ and its N neighboring pixels in the spectrum f_j, $(\Delta x_k, \Delta y_k)$ is the offset of the k-th neighboring pixel, ρ_i is the noise variance of the spectrum f_i, and $\omega_i(x, y)$ is the i.i.d. random variable with zero mean and unit variance.

When the symmetric neighborhood is used, i.e., c_{ij} has the same value for $f_j(x + \Delta x_k, y + \Delta y_k)$ and $f_j(x - \Delta x_k, y - \Delta y_k)$, Equation (1) is written as:

$$f_i(x, y) = \sum_{j=1}^{S} \sum_{k=1}^{4} c_{ijk} * (f_j(x + \Delta x_k, y + \Delta y_k) + f_j(x - \Delta x_k, y - \Delta y_k)) + \sqrt{\rho_i}\omega_i(x, y), \quad (2)$$

where $(\Delta x_k, \Delta y_k)$ is one of $\{(-l, 0), (0, l), (-l, l), (l, l)\}$, l ($l = 1, 2, 3$) is the level of the neighborhood.

In the case that gradient magnitude and gradient direction spectra are used, two independent equations are obtained as:

$$f_i(x, y) = \varepsilon\{f_i(x, y) | c_i\} = q_i(x, y)^T c_i, i \in \{GMag, GDir\}, \quad (3)$$

where (x, y) is a position inside a window in one spectrum, $q_i(x, y) = [f_{GMag}(x + \Delta x_k, y + \Delta y_k), f_{GDir}(x + \Delta x_k, y + \Delta y_k)]$, $c_i = [c_{iGMag}, c_{iGDir}]$. $\hat{c}_i$ and $\hat{\rho}_i$ are solved using the least squares (LS) estimation as:

$$\hat{c}_i = [\sum_{x=1}^{M} \sum_{y=1}^{M} q_i(x, y) q_i^T(x, y)]^{-1} [\sum_{x=1}^{M} \sum_{y=1}^{M} q_i(x, y) f_i(x, y)], \quad (4)$$

$$\hat{\rho}_i = \frac{1}{M^2} \sum_{x=1}^{M} \sum_{y=1}^{M} \left(f_i(x, y) - \hat{c}_i^T q_i(x, y)\right)^2. \quad (5)$$

In this study, the estimation was conducted using a moving-window on the three levels separately rather than only using a single level in order to provide a multi-scale representation. The moving-window with a shift of four pixels was used to enhance computational speed. The $\hat{c}_i$ and $\hat{\rho}_i$ obtained from Equations (4) and (5) on the three levels were combined into a 54-D MGSSAR feature vector for each window.

C. Obtaining MGSSAR Textons

Since the optimal number of textons varies for different textures, we first learned 10 textons (as reported in [16]) from the MGSSAR features using *k*-means for each texture. In total, 3340 textons were derived and were concatenated into a texton dictionary. It should be noted that the dimensionality (54-D) of the MGSSAR feature vector is less than the image patch feature vector (361-D) extracted from the same size (19×19 pixels) of windows. This accelerates the computational speed of *k*-means. In order to enhance the speed of the following vector quantization operation and reduce the dimensionality of the aura matrix computed later, *k*-means was further applied to the textons and a smaller number of textons were obtained.

D. Quantizing MGSSAR Features

Vector quantization was applied to the MGSSAR features in order to obtain more compact features. Each pixel (only those used for extracting the MGSSAR features were considered) in one texture was assigned the label of the texton which lay closest to this pixel in the MGSSAR feature space.

E. Encoding the Spatial Distributions of Textons

The aura matrix [5] encodes the spatial distributions of one gray level (or other labels) in the neighborhood of each other gray level. Thus, it captures the spatial interactions between different gray levels. An aura matrix [5] was calculated from the texton map of a texture image in order to encode the spatial distributions of the textons. This matrix is used for the representation of the texture image. Since the moving-window with a shift of four pixels was used for estimating MGSSAR models, an $N' \times N'$ ($N' = \left\lfloor \frac{N}{4} - 4 \right\rfloor$) texton map was obtained corresponding to the central pixels of these windows in an $N \times N$ texture image. Hence, the speed of the computation of aura matrices is further accelerated.

III. PERCEPTUAL TEXTURE RETRIEVAL EXPERIMENT

We tested the SDoT feature set using a perceptual texture retrieval task [3] which benchmarks computational texture rankings against human perceptual rankings. The perceptual rankings were obtained using the perceptual similarity matrix derived from the 334 *Pertex* textures [17, 18] using free-grouping. To be specific, all other 333 textures were sorted in descending order of perceptual similarity to provide a ranked

list for each texture. The computational rankings were obtained using the SDoT feature set under the framework proposed by Dong et al. [3]. The *Chi-square* statistic (see Equation (6)) was used to calculate pair-wise distances for the SDoT feature set. For each texture, all other 333 textures were sorted in descending order of computational similarity to generate a ranked list. Note that we only considered the multi-resolution feature vector in this study.

$$\chi^2(x,y) = \frac{1}{2}\sum_i \frac{(x_i - y_i)^2}{x_i + y_i}. \quad (6)$$

The *G* [19] and *M* [20] measures ($G, M \in [0,1]$) used in [3] were used to measure the consistency between computational and human perceptual texture rankings. These measures are defined as:

$$G = 1 - \frac{\sum_{i=1}^{R}(|r_i - r_i'|) + \sum_{i=1}^{N-R}[(N+1) - r_i] + \sum_{i=1}^{N-R}[(N+1) - r_i']}{N(N+1)}, \quad (7)$$

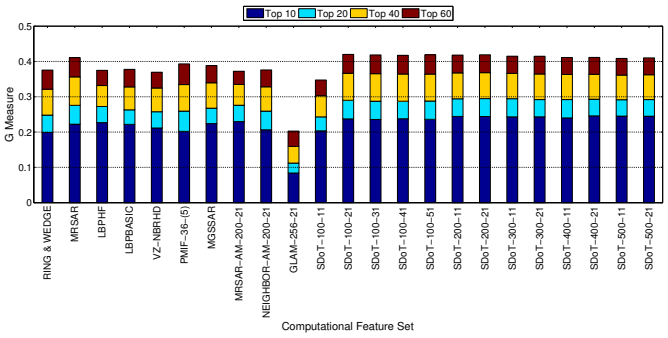
$$M = 1 - \frac{\sum_{i=1}^{R}(\left|\frac{1}{r_i} - \frac{1}{r_i'}\right|) + \sum_{i=1}^{N-R}(\frac{1}{r_i} - \frac{1}{N+1}) + \sum_{i=1}^{N-R}(\frac{1}{r_i'} - \frac{1}{N+1})}{2\sum_{i=1}^{N}(\frac{1}{i} - \frac{1}{N+1})}, \quad (8)$$

where R is the number of all relevant textures in N retrieved textures, r_i is the rank order of the i-th texture retrieved by a feature set, and r_i' is the rank order of the i-th texture image ranked by human observers of the i-th texture retrieved. The N was set as 10, 20, 40 and 60 respectively (only 10, 20 and 40 were used in [3]). The evaluation process was the same as that used by Dong et al. [3].

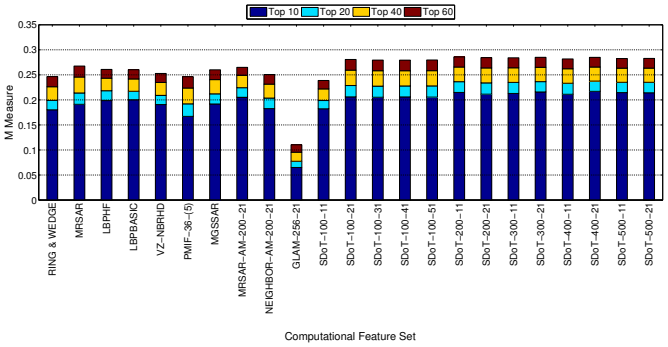
IV. Experimental Results

The SDoT feature set was tested with the aura matrices computed using different sizes of neighborhoods and different numbers of textons. Among the 51 feature sets tested by Dong et al. [3], five feature sets: LBPBASIC [21], LBPHF [22], MRSAR [14], RING & WEDGE [23] and VZ-NBRHD [16] were used as baselines because they performed retrieval better than the 46 other features. In addition, five more feature sets were also adopted as baselines: (1) MGSSAR (see Section II.B); (2) GLAM (Gray Level Aura Matrices) [5]; (3) and (4): two other "the aura matrix of textons" feature sets: MRSAR-AM and NBRHD-AM in terms of MRSAR textons (similar to MGSSAR textons) and VZ-NBRHD textons; and (5) PMIF (Perceptually Motivated Image Features) [24].

Fig. 2 shows the performance obtained using the feature sets introduced above. It is observed that the performance of the SDoT feature set was inferior when 100 textons and the 11×11 neighborhoods were used. However, in most cases, it outperformed its counterparts (including the other 46 feature sets examined in [3]) when more textons and/or larger neighborhoods were used, especially, when the top 10 textures were retrieved. The SDoT feature set also outperformed MRSAR-AM and NEIGHBOR-AM when aura matrices were computed at the same conditions. In addition, when the same neighborhood size was used, it outperformed the GLAM feature set [5]. Table 1 lists the values of the *G* and *M* measures obtained using SDoT and the best feature set reported in [3]. Fig. 3 further displays the top 10 "best" (a) and the 10 "worst" (b) query textures for the "SDoT-200-21" (200 textons and 21×21 neighborhoods) feature set when the *M* measure was considered. It can be seen that the proposed feature set retrieved some textures with long-range structures well while it failed when the margins between local elements were large, the continuous, global structures were large, or there was no obvious structure in images.



(a)



(b)

Fig. 2: The *G* and *M* measures obtained using 10 baseline feature sets and the SDoT feature sets with different parameters. Here, "A-*m*-*n*" means *m* textons and *n*×*n* neighborhoods are used by the feature set "A".

	Top 10		Top 20		Top 40		Top 60	
	SDoT	**Best [3]**	**SDoT**	**Best [3]**	**SDoT**	**Best [3]**	**SDoT**	**Best [3]**
G	**0.24**	0.23	**0.30**	0.28	**0.37**	0.36	**0.43**	0.41
M	**0.21**	0.20	**0.23**	0.22	**0.26**	0.25	**0.28**	0.27

Table 1: Comparison of the *G* and *M* measures obtained using SDoT-200-21 (200 textons and 21×21 neighborhoods) and the best feature set reported in [3].

V. Conclusions and Future Work

We proposed a new feature set, namely, SDoT, based on the spatial distributions of the textons extracted from the multiple gradient spectra simultaneous autoregressive (MGSSAR) model features. The proposed feature set encodes the higher order statistics computed between one pixel and its neighboring pixels in the same gradient spectrum as well as different gradient spectra. The spatial distributions of the textons learned from the local higher order statistics are further captured using the aura matrix [5]. The SDoT feature set was compared with the MGSSAR feature set and other 55 feature sets under a perceptual texture retrieval framework. Experimental results show that it is superior to all its

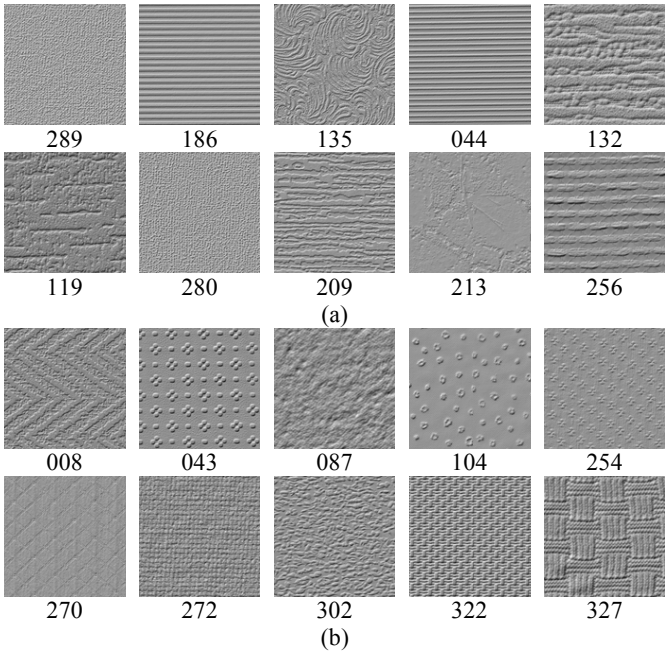


Fig. 3. Results for SDoT-200-21 when 10 textures were retrieved: (a) the top 10 "best" query textures sorted in a descending order of *M* measures; and (b) 10 "worst" query textures with *M* measures of 0.

counterparts according to two different performance measures. It is reflected that SDoT ranks textures more consistently with humans than the other feature sets tested in this study. We attribute this performance to the fact that the proposed feature set exploits the spatial relationship between textons.

However, the performance of the proposed feature set still lags behind that obtained by human observers [3]. It should be noted that aura matrices are computed over relatively small local neighborhoods. As a result, they cannot encode the spatial distributions over a larger spatial region than these neighborhoods. Although larger neighborhoods could be used, the computational cost will increase. In addition, it might yield an "averaging effect" and impair the discriminatory power of the feature set [14]. Since contours contain longer-range image structure information, the performance of the proposed feature set could be boosted by merging this type of information into it. We will investigate this in our future work.

ACKNOWLEDGMENTS

This work was supported by the Life Sciences Interface theme of Heriot-Watt University and grants from National Natural Science Foundation of China (Project No. 61271405 and 61301241), the PhD Program Foundation of Ministry of Education of China (20120132110018), Qingdao Science and Technology Plan Projects (12-1-4-1-(8)-jch).

REFERENCES

[1] D. J. Field, A. Hayes and R. F. Hess, "Contour integration by the human visual system: evidence for a local 'association field'," Vision Research, vol. 33, pp. 173-193, 1993.

[2] X. Dong and M. J. Chantler, "The Importance of Long-Range Interactions to Texture Similarity," Proc. of CAIP, 2013.

[3] X. Dong, T. Methven, and M. J. Chantler, "How Well Do Computational Features Perceptually Rank Textures? A Comparative Evaluation," Proc. of ICMR, 2014.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," Proc. of CVPR, 2006.

[5] I M., Elfadel and R.W., Picard, "Gibbs random fields, cooccurrences, and texture modeling," IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 1, pp. 24-37, 1994.

[6] J. De Winter and J. Wagemans, "The awakening of Attneave's sleeping cat: Identification of everyday objects on the basis of straight-line versions of outlines," Perception, vol. 37(2), pp. 245-270, 2008.

[7] J. Canny, "A Computational Approach to Edge Detection," IEEE Transactions on PAMI, vol. 8(6), pp. 679-698, 1986.

[8] M. Bloj, G. Harding, J. M. Harris, "Learning to use illumination gradients as shape cues," Session: 3D perception: Shape from shading and contours - Poster Presentation.

[9] O. Drbohlav and M. J. Chantler, "Illumination-invariant texture classification using single training images," Proc. of Texture 2005, pp. 31-36, 2005.

[10] T. Ojala, M. Pietikäinen, and D. Harwood, "A Comparative Study of Texture Measures with Classification Based on Feature Distributions," Pattern Recognition, vol. 29, pp. 51-59, 1996.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. of CVPR, vol. 1, pp. 886-893, 2005.

[12] T. Kobayashi, and N. Otsu, "Image Feature Extraction Using Gradient Local Auto-Correlations," Proc. of ECCV, part I, pp 346-358, 2008.

[13] B.A. Olshausen and D.J.Field, "Natural image statistics and efficient coding," Network: Computation in Neural Systems, vol. 7, no. 2, pp. 333-339, 1996.

[14] J. Mao, and A.K. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," Pattern Recognition, vol. 25(2), pp.173-188, 1992.

[15] J. Bennett, and A. Khotanzad, "Multispectral random field models for synthesis and analysis of color images," IEEE Transactions on PAMI, vol. 20(3), pp. 327-332, 1998.

[16] M. Varma, and A. Zisserman, "A Statistical Approach to Material Classification Using Image Patch Exemplars," IEEE Transactions on PAMI, vol. 31, pp. 2032-2047, 2009.

[17] A.D.F. Clarke, F. Halley, A. Newell, L. Griffin and M. J. Chantler, "Perceptual Similarity: A Texture Challenge," Proc. of BMVC, pp. 120.1-120.10, 2011.

[18] A.D.F. Clarke, X. Dong, and M. J. Chantler, "Does Free-sorting Provide a Good Estimate of Visual Similarity," Proceedings of Predicting Perceptions, pp. 17-20, 2012.

[19] R. Fagin, R. Kumar, D. Sivakumar, "Comparing Top K Lists," Proceedings of 14th ACM-SIAM Symposium on Discrete Algorithms, 28-36, 2003.

[20] J. Bar-Ilan,, M. Mat-Hassan,, M. Levene, "Methods for Comparing Rankings of Search Engine Results," Computer Networks, vol. 50(10), pp. 1448-1463, 2006.

[21] T. Ahonen, and M. Pietikäinen, "Image description using joint distribution of filter bank responses," Pattern Recognition Letters, vol. 30(4), pp. 368-376, 2009.

[22] T. Ahonen, J. Matas, C. He, M. Pietikainen, "Rotation Invariant Image Description with Local Binary Pattern Histogram Fourier Features," Proc. of SCIA, pp. 61-70, 2009.

[23] J.M. Coggins, and A.K. Jain, "A Spatial Filtering Approach to Texture Analysis," Pattern Recognition Letters, vol. 3, pp. 195-203, 1985.

[24] X. Dong and M. J. Chantler, "Texture Similarity Estimation Using Contours," Proc. of BMVC, 2014.