International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 24, No. 6 (2010) 917–946
© World Scientific Publishing Company
DOI: 10.1142/S0218001410008263

World Scientific
www.worldscientific.com

EFFICIENT CLASSIFICATION OF SCANNED MEDIA USING SPATIAL STATISTICS*

GOZDE UNAL

Faculty of Engineering and Natural Sciences
Sabanci University, Istanbul, Turkey
gozdeunal@sabanciuniv.edu

GAURAV SHARMA

Electrical and Computer Engineering Department
University of Rochester, Rochester, NY, 14627 USA
gaurav.sharma@rochester.edu

REINER ESCHBACH

Xerox Research Center, Webster Xerox Corporation
Webster, NY, 14580 USA
reiner.eschbach@xerox.com

Photography, lithography, xerography, and inkjet printing are the dominant technologies for color printing. Images produced on these "different media" are often scanned either for the purpose of copying or creating an electronic representation. For an improved color calibration during scanning, a media identification from the scanned image data is desirable. In this paper, we propose an efficient algorithm for automated classification of input media into four major classes corresponding to photographic, lithographic, xerographic and inkjet. Our technique exploits the strong correlation between the type of input media and the spatial statistics of corresponding images, which are observed in the scanned images. We adopt ideas from spatial statistics literature, and design two spatial statistical measures of dispersion and periodicity, which are computed over spatial point patterns generated from blocks of the scanned image, and whose distributions provide the features for making a decision. We utilize extensive training data and determined well separated decision regions to classify the input media. We validate and tested our classification technique results over an independent extensive data set. The results demonstrate that the proposed method is able to distinguish between the different media with high reliability.

Keywords: Pattern recognition; spatial statistics; distance based methods; dispersion methods; classification of marking and printing processes; classification of inkjet; xerographic; lithographic media.

* This paper is based on work that was conducted while the authors were employed at Xerox Corporation.

1. Introduction

A large number of color hardcopy images are being produced daily using a wide variety of image production processes. Photography, lithography, xerography and inkjet printing are the dominant technologies for color printing. Images produced on these "different media" are often scanned either for the purpose of copying or creating an electronic representation for use in various applications. Identification of the media, i.e. reproduction process that is utilized in order to create the hardcopy image that was scanned, from the scanned data is beneficial. A specific example of such a benefit is improved color calibration of the scanned image.[28] Since scanner responses to the same perceived color on different media are typically different, a media-dependent color calibration of the scanner is required for accurately mapping the scanner responses to a standard color space.[15,28] Unfortunately, this requires identification of the input media type at the time of scanning. The identified media class can be utilized for automatically associating a media-specific calibration with the image data or for identifying a smaller subset of calibration profiles for further manual selection. At present, this is either absent or a cumbersome and error-prone operator selectable option. Automated identification of the scanned media type is a preferable alternative. This paper describes an automated media classification system based on the scanned image data itself with no additional sensors. The identified media may be used to appropriately tag the scanned file, and if desired, the suitable color calibration profile may be embedded in the file.

Our goal of classification of the marking process on a scanned piece of paper is similar in spirit, but different in scale, to the document classification problem. Document classification is an important step in information and content retrieval and analysis. Similar to the document classification problem, scanned media classification can be thought of as a problem of mapping the space between an input document and output classes of possible reproduction media. The scale of the latter though is at a microscopic level because reproduction technology of the document is to be determined by classifying the pattern of the dots that make up the document. Pattern recognition and machine learning techniques have been applied to document classification.[10,11] Different statistical classification techniques such as naive Bayes classifier, the nearest neighbor classifier, and decision trees have been evaluated in Refs. 19 and 34. Similarly, neural networks have been popular in implementing document classification techniques.[4,26] An earlier document classification method by Pavlidis and Zhou[22] utilized the white space to segment a page image followed by rule-based classification. To discriminate between text, diagrams and halftones, the authors used cross-correlation between streams of binary pixels along scan lines. In Ref. 8, JPEG-compressed documents have been classified using morphological operations. Offline document image processing methods based on artificial neural networks were reviewed in Ref. 20. For classification of visual scenes, text modeling and local invariant features were combined in Ref. 24. A clustering-based technique is designed for estimating globally matched wavelet filters through a training set, and multiple two-class Fisher classifiers have been used for document segmentation into text, background and picture components.[16] A knowledge-based hybrid top-down and bottom-up approach was utilized for layout segmentation of documents into images, drawings, and tables, as well as text objects.[17] For the automatic transfer of paper documents into electronic documents a geometric layout analysis to segment a document into text, image, tables, and ruling lines is proposed in Ref. 18 which uses a pyramidal quadtree structure for multiscale analysis and a periodicity measure. A survey of document image analysis techniques is given in Ref. 21.

Texture analysis is an important area of machine vision, within which our work can be categorized. An extensive overview of texture analysis is given in Ref. 29. A general definition of texture is not straightforward, however, in simple and intuitive terms texture is described as repetitive local order over a region, or as a spatial variation in pixel intensities. Statistical methods such as co-occurrence matrices,[14] autocorrelation features, Voronoi tessellation[29] have been popular in texture analysis. Model-based approaches like Markov Random Fields, and signal processing approaches such as the filter-based techniques that use Gabor, Fourier and wavelet domain analysis have also been successfully used in various texture problems such as texture segmentation, texture classification, and texture synthesis.[29] Different textural properties in the graphics part and the text (non-graphics) parts of a document were captured by wavelets at different bandpass channels and a signature formed from such feature maps were utilized for document segmentation.[1] Most of the techniques in the texture analysis field inherit ideas from statistical theory. On the other hand, the spatial statistics, which is a field of statistics that deals with spatial data, has been more widely utilized in geological, environmental, meteorological, geographical and related sciences in contrast to pattern analysis and computer vision field. In this work, we apply tools from the spatial statistics field to a machine vision problem that arises in printing and scanning industry.

Our classification technique relies on the strong correlation among the four main types of input media — photographic, lithographic, xerographic and inkjet and the spatial characteristics of the corresponding reproduction processes. Photographic reproduction is continuous tone (contone), whereas the other media classes employ halftone printing. Among the halftone systems, for technological reasons inkjet uses primarily dispersed dot aperiodic halftoning, whereas lithographic and xerographic reproduction use primarily periodic rotated clustered dot screens.[13] Lithographic reproduction typically uses a higher halftone frequency and has different noise characteristics from xerography. Analysis of the underlying halftone/contone spatial characteristics of scanned image data can therefore be used to identify the input media. Blow-ups of scanned image blocks from photographic, inkjet, xerographic and lithographic media are shown in Fig. 1. Photographic, i.e. contone, media exhibits very low (ideally no) spatial variance. Clustered halftone dots produced by xerographic or lithographic printing display more regular and periodic spatial arrangements whereas dispersed halftone dots produced by inkjet printing

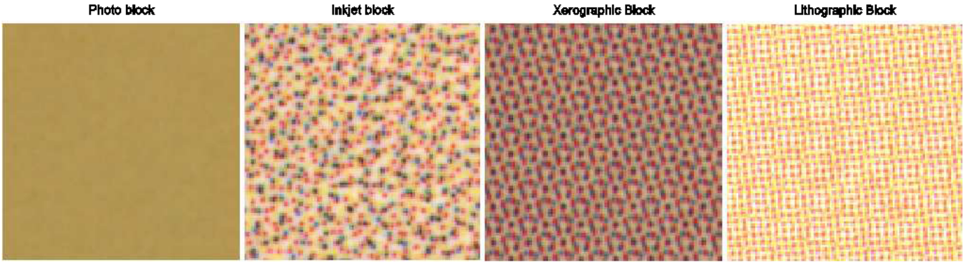


Fig. 1. Subregions from images on different media.

display high dispersion and no periodicity. Although the inherent spatial structure can be visually observed and hence be readily identified by someone familiar with different reproduction processes, automatic identification in a computationally efficient fashion using an automatic image processing system poses several challenges. Though classification based on 2-D power spectra has been proposed earlier,[2,27] real-time implementation requires classifiers that are much more computationally efficient.

Our contribution in this paper, is the development of a new efficient system for automatic classification of reproduction technologies based on spatial analysis of the scanned images. We generate point patterns from small blocks of scanned data that are representative of the underlying halftone processes and analyze them using measures of dispersion and periodicity from the field of spatial statistics, which are specifically modified and tuned for our application. The measures are finally used in a decision criterion to classify the input media type. Experimental studies show that the input media, or equivalently the marking process, can be identified correctly to a high degree of confidence using the proposed method.[a]

The organization of the paper is as follows. In Sec. 2, we summarize the required background knowledge on spatial statistics field. In Sec. 3, the application of two spatial statistical measures to the problem of media classification is presented. Experimental results are given in Sec. 4, and the conclusions are given in Sec. 5.

2. Background on Spatial Statistics

Spatial statistics deal with data that have spatial inter-relations. The locations of data and attributes at each data point are entities of interest in analyzing spatial information. *Spatial point patterns* arise when the important variable to be analyzed is the location of events. An observed spatial point pattern can be thought of as a realization of a spatial stochastic process. Mathematically, the spatial point processes are expressed in terms of the number of events occurring in arbitrary subregions or areas Ω, of the whole study region R. Analysis of spatial point patterns implies

[a] A short and preliminary version of this work was presented at the 2004 IEEE Int. Conf. on Image Processing.[31]

identification of point pattern structure, inferring the parameters of the point process, i.e. the underlying model, from the observed pattern.

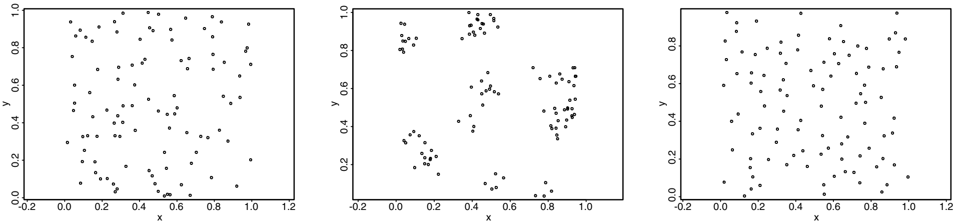
The simplest theoretical model for a spatial point pattern is that of Complete Spatial Randomness (CSR), in which the events are distributed independently according to a uniform probability distribution over the region R. Formally, the point process that gives rise to such an arrangement is called a Homogeneous Poisson Process, which is the only point process equivalent to CSR. The Poisson process serves as a reference in point pattern analysis, and is usually used as the null hypothesis in statistical tests on spatial structure. A Poisson process is usually defined by

(1) Counts of events in a finite region Ω with a Poisson distribution of mean $\mu(\Omega)$, that is: $\Pr(N(\Omega) = n) = \frac{1}{n}(\mu(\Omega))^n e^{-\mu(\Omega)}$ where $N(\Omega)$ is the number of events in set Ω.
(2) Counts in disjoint sets are independent.

CSR is the "white noise" of spatial point processes,[6] characterizing absence of structure in the data. An important question that arises in exploring spatial data is whether the observed events display any systematic spatial pattern or *departure from randomness*. Testing deviations from CSR gives an essential insight in terms of the structure of a point pattern. In CSR, events do not interact with each other, either repulsively (regularity of events), or attractively (aggregation) of events. There are two types of deviations from a random pattern (CSR):

(1) If events at short distances occur more frequently than are expected under CSR, and the pattern has a more uneven intensity of points with local peaks at aggregations,[33] the pattern is called *aggregated.*
(2) Patterns that have an evenness in distribution are called *regular* patterns. They exhibit more large inter-event distances than a CSR process.

A completely random spatial process, e.g. a realization of Homogeneous Point Process on the unit square, conditioned on $n = 100$ events is shown in Fig. 2(a). Figure 2(b) demonstrates an aggregated point process whereas Fig. 2(c) shows a realization of a regular point process. Starting with an observed point set, attempting



(a) (b) (c)

Fig. 2. Realizations of spatial point processes: (a) homogeneous Poisson process, (b) aggregated process, (c) regular process.

to analyze or identify the spatial pattern, a reasonable approach is to test for CSR hypothesis against regular or aggregated alternatives.

2.1. *Analysis of spatial point patterns*

Although full characterization of a stochastic process is not easy, useful aspects can be characterized in terms of first-order and second-order properties of a stochastic process. First-order properties describe the way in which the expected value of process varies across space, while second-order properties describe the covariance between values of the process at different regions in space.[12] The first-order properties, which are described in terms of the intensity of the process, are widely used in point pattern analysis.

The methods to analyze first-order properties of spatial point patterns can be classified into broadly two groups. The first group of techniques is *area-based*, so called quadrat count methods, relying on frequency distribution of the observed numbers of events in regularly defined subregions of the study area (quadrats). The second group of techniques are *distance-based*, using information on the distances between events to characterize the pattern.

2.1.1. *Quadrat methods*

Quadrat sampling refers to collecting counts of events in small subregions of a study region Ω. Subregions are generally rectangular in shape, and their placement can be either at random or contiguously over Ω. Under a hypothesis of complete spatial randomness, the distribution of number of points per quadrat is Poisson with mean λ, which is the intensity of the process given by $\lambda = n/|\Omega|$, where $|\cdot|$ denotes the cardinality of set Ω. A χ^2 test can be carried out by calculating $X^2 = \sum(\text{Obs.} - \text{Exp.})^2/\text{Exp.}$, where "Obs." and "Exp." correspond to observed and expected number of quadrats with a given number of events. A great number of variations of this statistic, X^2, motivated by the equality of mean and variance of Poisson distribution have appeared in literature.[6] A popular one is by David and Moore[7] who suggested using an index

$$D = \frac{s^2}{\bar{x}} - 1, \tag{1}$$

where s^2 is the sample variance, and $\bar{x}$ is the sample mean. The expected value of the index is 0 for a random pattern or equivalently a homogeneous Poisson process. We are interested in using this index as a *dispersion measure* to distinguish two types of departures from a Homogeneous Poisson Process (HPP): *Regular processes show less dispersion than a HPP*, *and aggregated processes show more dispersion than a HPP*.

2.1.2. *Distance methods*

Second popular group of tests for analysis of spatial point patterns is based on distances between events. Most of these techniques also use the Poisson distribution as the underlying model for inferences about the pattern. Widely used distances are nearest-neighbor (nearest-event) and event-to-event distances. A *nearest-event distance* is the distance from any given event to the closest event in the study region Ω. An earlier test is by *Clark and Evans*[5] which uses mean nearest-event distance $\bar{y}$. Intuitively, with respect to the range of the histogram of the data, small values of $\bar{y}$ indicate aggregation, whereas large values of $\bar{y}$ indicates regularity. Diggle[9] suggested that a test based on the entire empirical distribution function of nearest-event distances, are more powerful. He referred to such techniques as *refined nearest-neighbor analysis*. Once, the empirical distribution function $\hat{G}(y)$ of the distance measure is calculated, it will be compared against the theoretical distribution function $G(y)$ under Homogeneous Poisson Process. The significance of the test can be evaluated using Monte Carlo (MC) simulations.

Calculating all event-to-event distances in a large observation region will have a high computational load. Doing MC tests in addition would increase the load. Therefore, tests based on nearest neighbors, which involve only n nearest-event distances, are more popular and are quite powerful. Also tests that do not require Monte Carlo simulations are more appealing due to computational and practical considerations.

For periodic halftones obtained by clustered dot dithered screens (such as those in Lithography and Xerography observed in Fig. 1), the noise in the halftones also impacts observable spatial statistics, as briefly discussed in Appendix A.

2.2. *Statistic selection*

We are interested in classification of spatial point patterns by testing departures from CSR. For this purpose, the dispersion measure based on quadrat counts is well suited since this is a strongly established statistic, which is asymptotically known to follow a χ^2 distribution. However, the choice of quadrat size is important. If a scale at which the structure of the pattern is detectable can be found, then index of dispersion is an effective measure for testing two-way departures from CSR. Index of dispersion provides a *global* test to detect heterogeneity, which usually manifests itself in terms of aggregation. Thus, quadrat-based analysis is powerful for aggregation, however, weak for regularity.

For distance-based methods requiring MC simulations, a choice of which statistic to use is not that straightforward. These methods emphasize *local* characteristics (as opposed to the global dispersion index), thus are more sensitive to aggregation and regularity, and they are sensitive to the choice of scale. A combination of these two techniques modified and tuned for our application to the classification of the marking processes will be utilized as outlined in the following section.

3. Halftone/Contone Patterns

The decision tree which is depicted in Fig. 3, explains the sequential decisions that are made for the media identification problem.

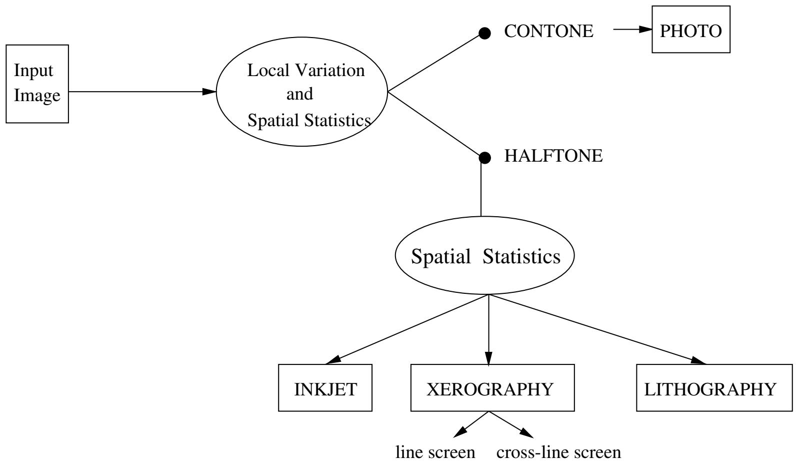


Fig. 3. Decision tree for the media classification problem.

3.1. *Contone media*

The first step in decision making is to differentiate between the two major types of processes, namely the contone process and the halftone process. Contone, i.e. photographic images have a much smaller variation than the halftone images. Hence, a front end processing is to look at local variation of the scanned data in order to detect photographic media. For locality, the scanned image is processed in a block-by-block fashion, and initially, mean intensity and standard deviation of the intensity over each block is calculated. If the standard deviation of the block is less than some predetermined threshold value, which is to be selected through experiments, then the block is a candidate *constant* block with very small spatial variation. The remaining blocks are labeled as *varying*. Irrespective of the media, there exist scanned image blocks that contain almost no information about the underlying process. For example, no halftone structure is present in white regions or solid regions where each of the colorants is either absent or fully covers the print. Thus saturated blocks of an image are one of the examples of such unusable regions. Given a scanned image $\mathbf{I} = (I_R, I_G, I_B)$, $\mathbf{I} : \Omega \to \mathbb{R}^3$, with three channels corresponding to red, green, and blue, we define a measure of *saturation* as

$$S = 1 - \frac{\min\{\bar{I}_R, \bar{I}_G, \bar{I}_B\}}{\max\{\bar{I}_R, \bar{I}_G, \bar{I}_B\}}, \tag{2}$$

where $\bar{I}$ refers to the mean intensity of I. Hence, a block with an S value close to 1 will be highly saturated, and this implies that there is almost no information to be gained from processing this block. Similarly, if the luminance of the block is close to darker end, i.e. close to black, then again, no information is present in such a block hence it is not usable. The *blackness* and *whiteness* can be defined as the closeness of the quantities $\max\{I_R, I_G, I_B\}$ and $\min\{I_R, I_G, I_B\}$ to zero and the maximum value, respectively.

As a result, the saturated and black blocks are labeled as *non-informative*. Using the remaining informative blocks a decision is then made whether the media is photographic by evaluating the fraction of the informative blocks that are varying and declaring the media to be photographic if this fraction is below a preset threshold.

Next, the spatial statistics of varying blocks are utilized to distinguish between the different halftone media using the spatial point pattern analysis methods which are adapted specifically to the problem of media classification.

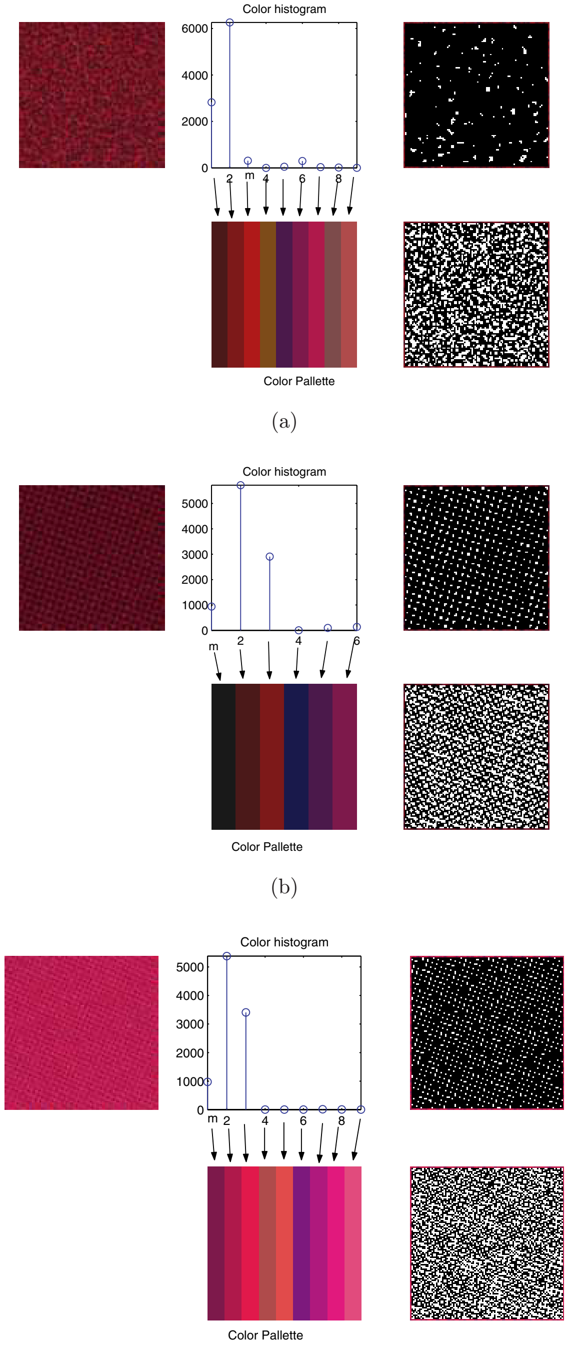
3.2. *Point pattern generation*

Each scanned image block is processed to extract a point pattern. Close investigation of halftone dot patterns in Fig. 1 reveals the fact that a group of dots with a certain color value when viewed separately from dots with other colors, represents the underlying spatial halftone pattern. This is equivalent to viewing a color level set of the color block. Suppose our study region $\Omega \subset \mathbb{R}^2$ corresponds to the spatial domain of a small block image $\mathbf{I}$ defined as: $\mathbf{I} : \Omega \to \mathbb{R}^3$. If we denote a color vector in $\mathbb{R}^3$ as $\mathbf{c_i}$, a color level set , say $LS(x, y)$, is given by

$$LS(x,y) = \begin{cases} 1, & \{(x,y) : \mathbf{I}(x,y) \in \mathbf{c_i}\} \\ 0, & \{(x,y) : \mathbf{I}(x,y) \notin \mathbf{c_i}\}, \end{cases} \tag{3}$$

Selection of a representative color value can be carried out by examining the color histogram $H_c : \mathbb{R}^3 \to \mathbb{N}$, ($\mathbb{N}$: natural numbers), of the block. The scanned output produced by the scanner has three separate channels, of which red channel corresponds to cyan dots, green channel corresponds to magenta dots, and blue channel corresponds to yellow dots of the input media placed on the platen. Each channel's gray values in $[0, 255]$ of the color space $[0, 255]^3 \equiv [0, 255] \times [0, 255] \times [0, 255]$ is binned into N intervals, then an N^3 number of color cubes whose sides are of length $\Delta = 255/N$ can be used to find the color histogram H_c of the image block.

For example, in Fig. 4(a) at the top left, a color block from an inkjet media, and its color histogram in the middle are depicted. The horizontal axis on this plot serves as an index into the color palette (indicated by arrows). Suppose colors in the block image fall into n_c number of color cubes that are in $\mathbb{R}^3$. They can be depicted by n_c vertical color strips on a color palette, and each color strip represents a color cube $\mathbf{c_i}, i = 1, \ldots, n_c$. On the right, two sample level sets which are extracted out of the block using a specific color from the histogram are given. If we select the color cube $\mathbf{c_3}$, whose frequency is closest to the mean frequency of present colors (marked by letter m on the color histograms), the level set at top right is obtained. The level set at the bottom right corresponds to the color cube $\mathbf{c_2}$ with maximum frequency. Similar plots are given in Figs. 4(b) and 4(c) for sample xerographic and lithographic blocks. In these figures, the level sets shown on the right serve as the point patterns which clearly display the regular and periodic structure of spatial dot arrangements resulting from xerographic and lithographic printing process and the dispersed and aperiodic structure of dots from inkjet printing process.

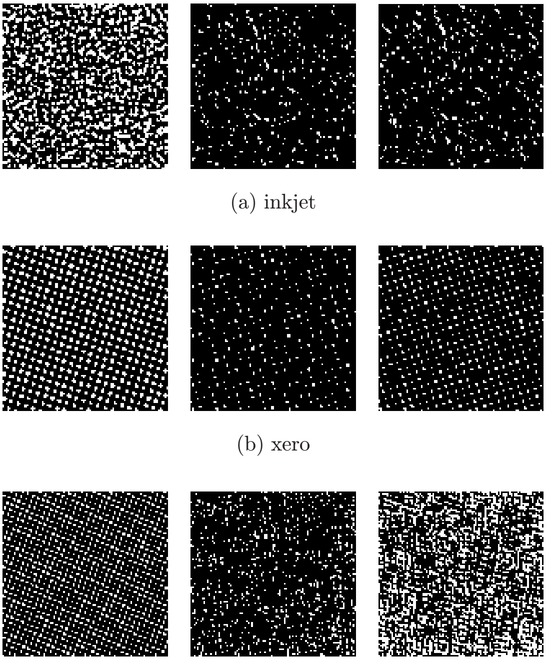


(a)

(b)

(c)

Fig. 4. Color point pattern generation. (a) Point pattern generation from color histogram of an Inkjet media (binsize $\Delta = 50$). (b) Point pattern generation from color histogram of a xerographic media (binsize $\Delta = 50$). (c) Point pattern generation from color histogram of a lithographic media (binsize $\Delta = 50$).



(a) inkjet

(b) xero

(c) litho

Fig. 5. Level sets obtained from image channels of different media. First column: red channel, second column: green channel, third column: blue channel.

Another way to obtain point sets from a color image block is to work with image channels, and generate level sets from each channel separately as shown in Fig. 5. In (a), the point patterns clearly show a dispersed structure in which there are some regions which do not contain events (events are denoted by light pixels), and there are some regions with an aggregation of events. Evenness in distribution of events or regularity with clear periodicity of dot patterns can be observed in (b) and (c) which correspond to xerographic and lithographic media, respectively.

3.3. *Dispersion measure*

In order to quantify the extent of dispersion, an appropriate measure, D, was given in Eq. (1), which can be calculated from the sample variance s^2, and sample mean $\bar{x}$ of event counts in randomly or contiguously placed quadrats over the study region. For a completely random pattern, the observed point set would be a realization from HPP (homogeneous Poisson process), hence the sample variance s^2 and the sample mean $\bar{x}$ of the event counts in quadrats would be equal. This implies that the dispersion measure D calculated for each block would be close to 0 (with high probability). Complete spatial randomness is generally unattainable in practice. The point patterns generated from inkjet media closely resemble aggregated patterns.

This is an expected result because the inkjet printing produce dot patterns that have uneven distribution of events over space, and are dispersed in nature. As a result, the dispersion measure D for a point pattern that belongs to inkjet media will be positive, i.e. $D > 0$. On the other hand, the point patterns generated from either xerographic or lithographic media fall into the second class of departures from completely spatial randomness, i.e. the regular patterns.

For real data, it is not feasible to expect that all point patterns extracted from the image produce a single D value. Hence, a symmetric density function of D which is peaked at 0 with small tails at both sides would result for a random pattern. With the same reasoning, for a scanned image on inkjet media, the density of D over blocks of the image will be positively skewed with a peak on the positive side. Xerographic and lithographic media, on the other hand will result in a positively skewed D distribution with a peak at the negative side.

Histograms of the dispersion measure D, $(H_D : [-1, \infty) \subset \mathbb{R} \to \mathbb{N})$, calculated for three different reproductions of a sample composition image, which is reproduced in all four different media and then scanned, are given in Fig. 6. The plots on the left column show D histograms computed from the point patterns generated via color level set extraction whereas the plots on the right column show D histograms for red, green and blue channels computed from the point patterns extracted via gray level

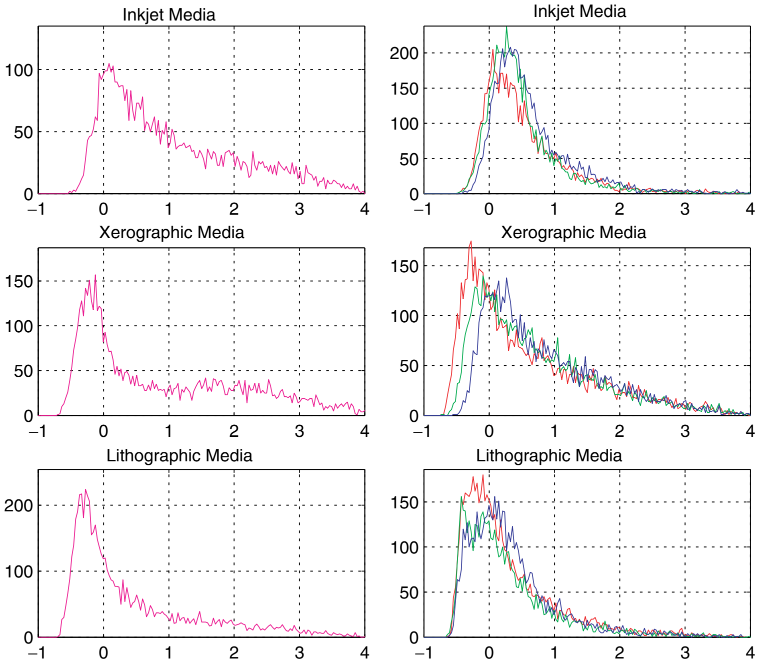


Fig. 6. Histograms of dispersion measure D calculated over the image on three different media shown on three different rows. Column 1: from a point pattern via a color level set; Column 2: from point patterns via gray level sets, hence plotted as red, green and blue plots.

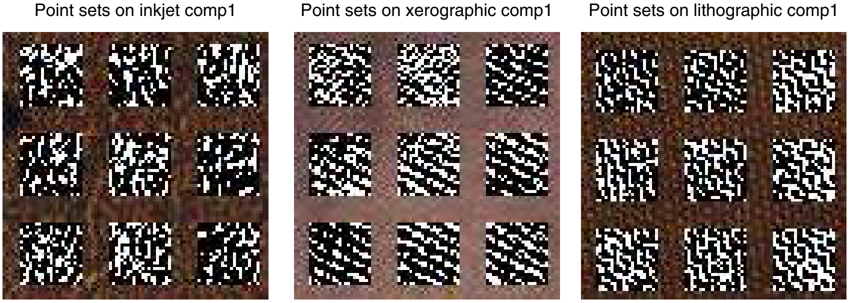

Point sets on inkjet comp1
Point sets on xerographic comp1
Point sets on lithographic comp1

Fig. 7. Sample point sets for the image.

sets. A few sample blocks from the image on inkjet, xerographic and lithographic media along with the point patterns generated through color level sets are shown in Fig. 7 for visualization of dispersion and regularity. The blocks here are from a sample composite/collage image that is a combination of several natural images with both busy and homogeneous regions, hence it constitutes a representative example of scanned documents. It can be observed on plots of the first row in Fig. 6 that the inkjet media produces a dispersion measure density which is skewed away from zero to the right, i.e. in the positive direction. Plots for the density of the dispersion measure for xerographic input media (on the second row) show that point patterns generated from this technology have dispersion density with a peak to the left of zero, i.e. toward negative values. Similarly, density of the dispersion measure D for lithographic media (on the third row), exhibits a negatively-skewed behavior. It can be observed that tail of the density function over negative values is small whereas the positive side of the density is heavy-tailed for all types of media. The heavy tail on the positive side of the density plots is due to noise inevitably present in real life image data and the point sets generated from such data. Another reason is the fact that dispersion measure D takes values in the real interval $[-1, \infty)$, therefore, there is an inherent bias towards positive side — even for a perfectly regular pattern the lowest value that D can attain is -1 whereas the maximum value is unbounded.

Recall that a front-end processing separates the blocks of an image into two classes, namely constant blocks, and varying blocks. For photographic scanned images, ideally, the intensity variation should be close to zero. However, the real images scanned may have variations that can specify a considerable number of blocks of a photo image in the class of varying blocks. Moreover, the blocks are taken in a raster scan fashion, and some blocks will be naturally on boundaries of different homogeneous regions over an image. Point patterns that are generated from these two types of photographic blocks will have large solid white regions, and large solid black regions. The point patterns of sample blocks from two images on photographic media are shown in Fig. 8.

Dispersion analysis on such blocks results in very high dispersion values. Density of dispersion D over these two photographic images are given in Fig. 9. The density

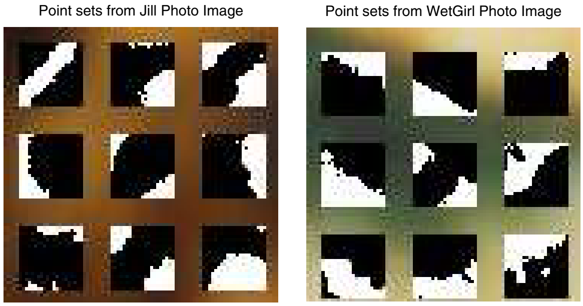

Point sets from Jill Photo Image
Point sets from WetGirl Photo Image

Fig. 8. Point patterns generated via color level set extraction from two photographic images.

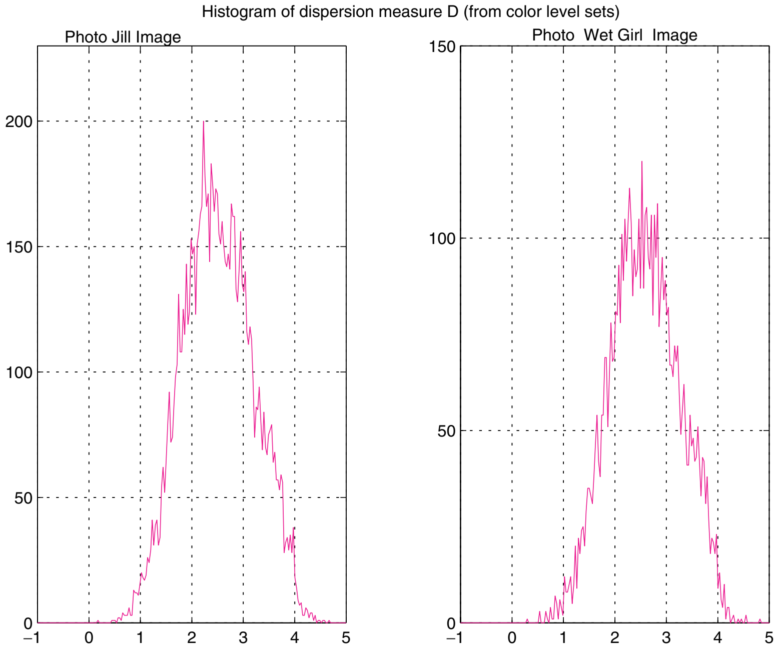


Fig. 9. Histograms of dispersion measure D calculated over natural images on two photographic media. Both plots are produced from point patterns extracted via a color level set.

for both images are skewed towards positive values much higher than 0. This result is expected because the randomly sampled quadrat counts produce a very high variance when compared to their mean, hence the dispersion will attain a large positive value.

To account for the noise effects of real data and the positive-biasedness of the dispersion measure, the decision to distinguish regular patterns from Poisson or aggregated patterns is based on the areas under the histogram $A_{H_{-\beta}} = \int_{-\beta}^{0} H_D(x)dx$, and $A_{H_{+\beta}} = \int_{0}^{\beta} H_D(x)dx$, where $\beta \in (0,1)$ is a parameter which is chosen as 0.5 in this work. Point patterns obtained from photographic images in our training set result in a dispersion density that is peaked in $D \in (2, 3)$ interval. Hence, the area under dispersion density, $A_{H_{2.5}} = \int_{2.5-\beta}^{2.5+\beta} H_D(x)dx$, can also be calculated as a measure in detecting photographic media (in case it has not already been classified as photo by initial variance calculations). Hence, we have the following rules according to $A_{\max} = \max(A_{H_{-\beta}}, A_{H_{+\beta}}, A_{H_{2.5}})$:

$$\text{Decision} = \begin{cases} \text{Regular (Xero/Litho)} & \text{if } A_{\max} = A_{H_{-\beta}} \\ \text{Aggregated (Inkjet)} & \text{if } A_{\max} = A_{H_{+\beta}} \\ \text{Photo} & \text{if } A_{\max} = A_{H_{2.5}} \end{cases} \tag{4}$$

This decision provides a first broad classification of the underlying halftone/contone process into three classes, namely photo media, inkjet media and xerographic/lithographic media.

3.4. *Periodicity measure*

To classify between the two types of media: xerography and lithography, a prominent characteristic of both, which is the *periodicity*, can be easily traced in the point patterns. A critical distinguishing factor between xerographic and lithographic media is the low frequency exhibited by xerographic printing process against the high frequency exhibited by lithographic printing process. Hence, utilizing a distance-based spatial statistical measure can help distinguish between xerography and lithography by detecting stronger large-distance neighbor events for a regular point process which is expected to come from a lithographic media. Xerographic printing process produces a noisier signal, and halftone screen frequency in xerography is lower. This causes weak periodicities at larger distances when compared to those of lithography.

One of the distance methods, i.e. refined nearest-neighbor analysis, explained in Sec. 2.1.2, considers empirical distribution function of nearest-event distances. For a completely random pattern, theoretically the distribution function is $G(y) = 1 - e^{-\lambda\pi y^2}$, which gives the probability of having at least one additional event at distance y. The empirical distribution function of nearest-event distances, $\hat{G}(y)$, can be computed from a point pattern and compared against the theoretical distribution.

In this study, due to computational considerations, for example, the high computational load involved in searching for a nearest event in the two-dimensional plane, we utilize a modified one-dimensional (1-D) version of the nearest-event distance. Each block is scanned in two 1-D directions, the distance between consecutive events are computed, and histograms of the nearest-event distances in two 1-D directions, x and y, are calculated. In 1-D case, we use the term *nearest-event distance* interchangeably with *inter-event distance*, which gives the distance from one event to the next neighbor event in the direction of scanning.

Histograms of inter-event distances in 1-D are calculated and accumulated over blocks of the image. Hence an estimate of the average probability density function for inter-event distances in two directions of the overall image is obtained. These density functions are used to test against the null hypothesis of complete spatial randomness or even for further classification. As mentioned in Sec. 2.1.2, Monte Carlo (MC) simulations are carried out for distance methods to assess the significance of the test. Although we utilize a large number of blocks whose point patterns can be viewed as realizations from the underlying point process, the number of events in each realization, i.e. in point patterns of each block, is not the same due to the variation in image content. Therefore, MC simulations, which require same statistical parameters for all simulated realizations used in testing are not suitable here. Instead, over the whole image, we find a sample average of the empirical distribution function of 1-D inter-event distances. This provides a global estimate of the distribution of halftone dot patterns generated by the specific reproduction technology.

Sample average of the histograms of inter-event distances for the same sample image on inkjet media in Fig. 10 show that inkjet printing produces dot patterns with no periodicity, and inter-event distances decrease roughly exponentially as distances get larger. A single peak is detected at a short inter-event distance which is marked by a vertical (cyan) line on each plot. Aggregated processes have an excess of short-distance neighbor events. This is compatible with underlying aggregated character. Thus the distribution of inter-event distances is peaked at a short distance y, and has decreasing behavior with no further peaks at larger distances. This agrees with what

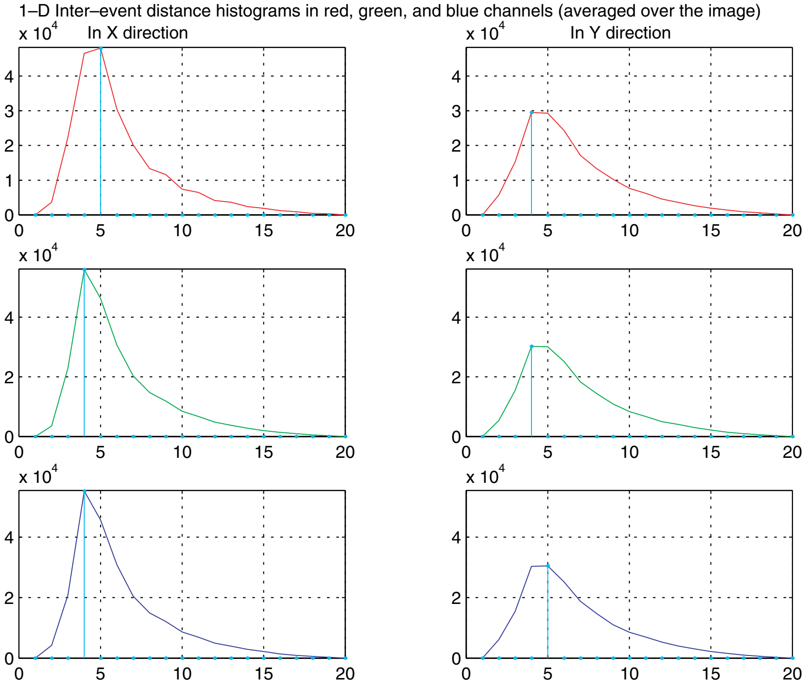


Fig. 10. Inter-event distance histograms for the image on inkjet media, where left (right) column depicts densities in x (y) direction for red, green and blue channels, respectively.

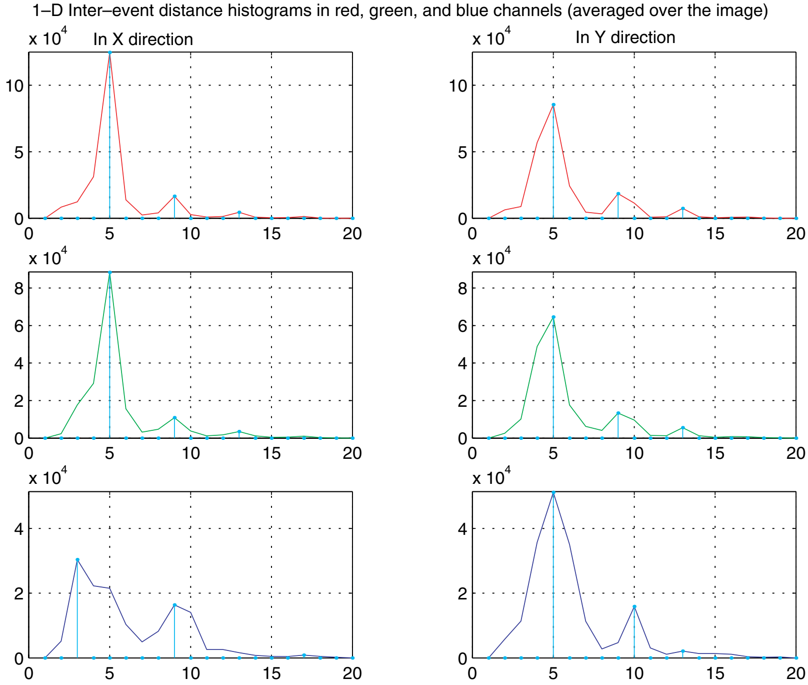


Fig. 11. Inter-event distance histograms for the image on xerographic media, where left (right) column depicts densities in x (y) direction for red, green and blue channels, respectively.

is observed in our experiments, i.e. a distinctively unimodal character for inter-event distance histograms of inkjet media. Therefore, the characteristic feature of the inter-event (or nearest-event) distributions classifies inkjet input media in the class of aggregated patterns.

Inter-event distance histograms for the same image reproduced on xerographic media are given in Fig. 11. Peaks detected on the histograms are also marked by vertical (cyan) lines. The presence of a second and even a third peak in these plots, although weak, is an indication of a global periodicity in the point patterns. A generally trimodal characteristic of the inter-event distance histograms is noted in this case.

Figure 12 shows the point patterns generated from a subregion of size 100 × 100 of the sample composition image on xerographic media whose distance histograms are given in Fig. 11. Regular arrangements of the events in these point patterns are transparent to the eye as well as to the dispersion measure and to the final criteria we will obtain from the distance measures.

For a scan of an image reproduced on lithographic media, inter-event distance histograms in x and y directions display a distinctively trimodal character as shown in Fig. 13. Existence of a strong second peak in addition to a relatively strong third peak when compared to those of xerographic media is an indication of stronger

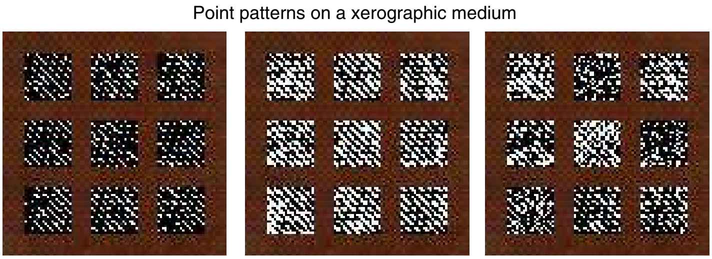


Fig. 12. Point sets extracted from a xerographic media of rotated screens.

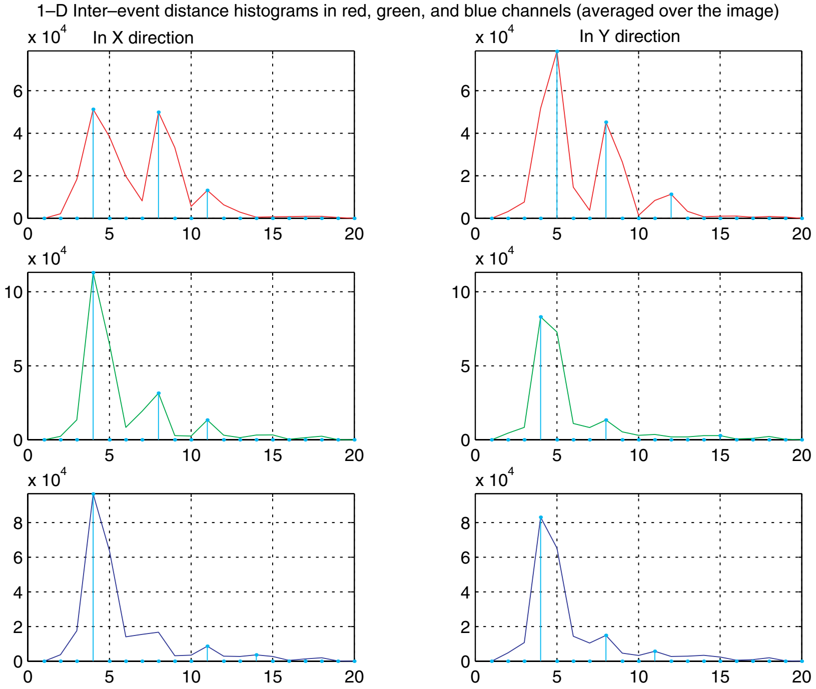


Fig. 13. Inter-event distance histograms for comp1 image on lithographic media, where left (right) column depicts densities in x (y) direction for red, green and blue channels, respectively.

periodicity characteristics. This is expected for lithographic printing process, which is less noisy and the screen patterns have higher frequency. We observe that the signal is the cleanest in red channel in nearly all experimentally tested images, hence strength of the larger-distance peaks can be better observed in this channel.

Another scan of a natural image, which is reproduced via a common xerographic line screen, results in the histogram plots as shown in Fig. 15. Distance histograms of consecutive events in one of the directions, here x direction, does not display periodicity whereas histograms in the other direction, here y direction, show the detection of weak second and third peaks at larger distances. This is in agreement with the halftone dot patterns generated by a line screen in xerography. Typical point patterns generated in a region of size 100 × 100 from a scanned image on such xerographic media are given in Fig. 14, where the regular feature of the patterns is easily discernible. As a result, the distance histogram we utilize can also detect line screen xerography as can be observed from the asymmetry of the densities in two orthogonal directions in Fig. 15.

3.4.1. *A final distance-method-based measure*

Observations over an extensive data set (on the order of hundreds of representative blocks from images) lead to the following rules. There was only one single peak for

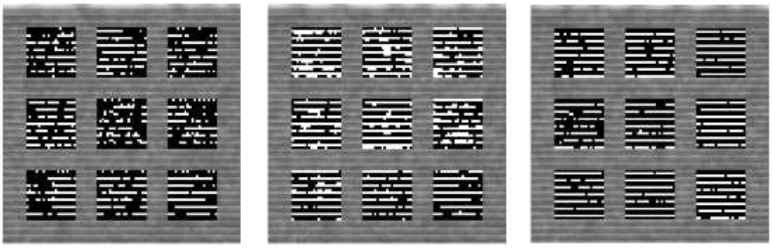


Fig. 14. Point sets extracted from a xerographic media of line screens.

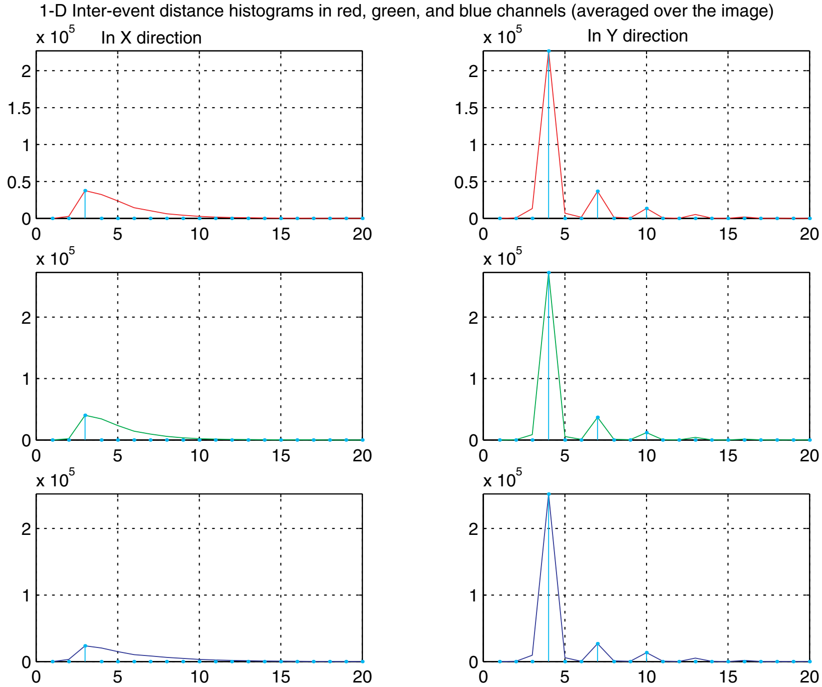


Fig. 15. Inter-event distance histograms for a natural image on a xerographic line media.

inkjet media, there was a weak second peak and a weaker third peak for xerographic media, and there was either a very strong second peak or a stronger third peak for lithographic media. These observations are all compatible with what is expected in theory from regular and aggregated process, and even further classification of regular processes into high and low frequency classes. This lead us to use 2-D decision planes which summarizes the information obtained from the inter-event distance histograms of an image in two directions, namely x and y. The final test for a three-way classification of the input media follows by plotting the value at Peak 3, the third detected peak, over the value at Peak 2, the second detected peak, versus the value at Peak 2 over the value at Peak 1. The strength of Peaks 3 and 2 produced by lithographic media provides good means of its identification. Thus, the quantities Peak2/Peak1 (strengths) and Peak3/Peak2 (strengths), evaluated along both x and y directions, are obtained as the distance measures to be used in the final decision phase.

4. Experimental Results

In this section, we present the experimental results and the implementation details including the decision phase. The training set included scanned pages, which were representative of color image documents encountered in typical office settings. Often a large percentage of the document scans are black and white where the current methodology is not required. However, as we indicate in the introduction, the problem of accurate media identification is critical in color scanning in order to obtain more accurate color reproduction.

4.1. *Implementation and parameters*

Several parameters that are used in implementation of both techniques are set after experimenting with the training set, which included images that were scanned at 600 dpi resolution and digitized with eight bits per sample corresponding to a pixel value range from 0 through 255. The quadrat size should be chosen so as to resolve the differences in structure between the different types of halftones and therefore depends on the scan resolutions. For our 600 dpi scans quadrats of size 2×2 and 3×3 were found to be reasonable choices to capture point pattern structure over small blocks of size 20×20. With a change in resolution, the quadrat sizes should be scaled appropriately and 4×4 or 5×5 quadrats may be needed for higher resolution scans in order to capture the larger apparent pattern of the halftone dots in the scans. The method is, however, not unduly sensitive to the choice of quadrat size since a wide range of quadrat sizes will capture the strong differences in the structure of the point processes corresponding to our different media. We use random quadrat sampling method in calculation of the dispersion measure D for each block.

We found that a standard deviation value of $\mathrm{STD} = 8$ to be an upper limit for labeling constant blocks. For generating point patterns in order to calculate dispersion measure over 20 × 20 blocks, in a color level set extraction, with binsize $\Delta = 10$, and the color chosen is the maximum frequency color from the color distribution.

After the dispersion measure is calculated over all varying blocks of an input image, the three distinct areas under the dispersion histograms are calculated as explained in Sec. 3.3. The parameter β was experimentally set to a value of 0.5 based on a training data set of representative patches derived from images that included multiple examples of each of the different media under consideration.

After we have found peaks in averages of 1D inter-event distance histograms, we plot Peak3/Peak2 versus Peak2/Peak1 for both x and y directions for all the images in the training set as shown in Fig. 16. As can be observed from these scatter plots, there is a distinct separation between different halftone media. Lithographic images either have high Peak2/Peak1 or Peak3/Peak2 ratios whereas Xerographic images

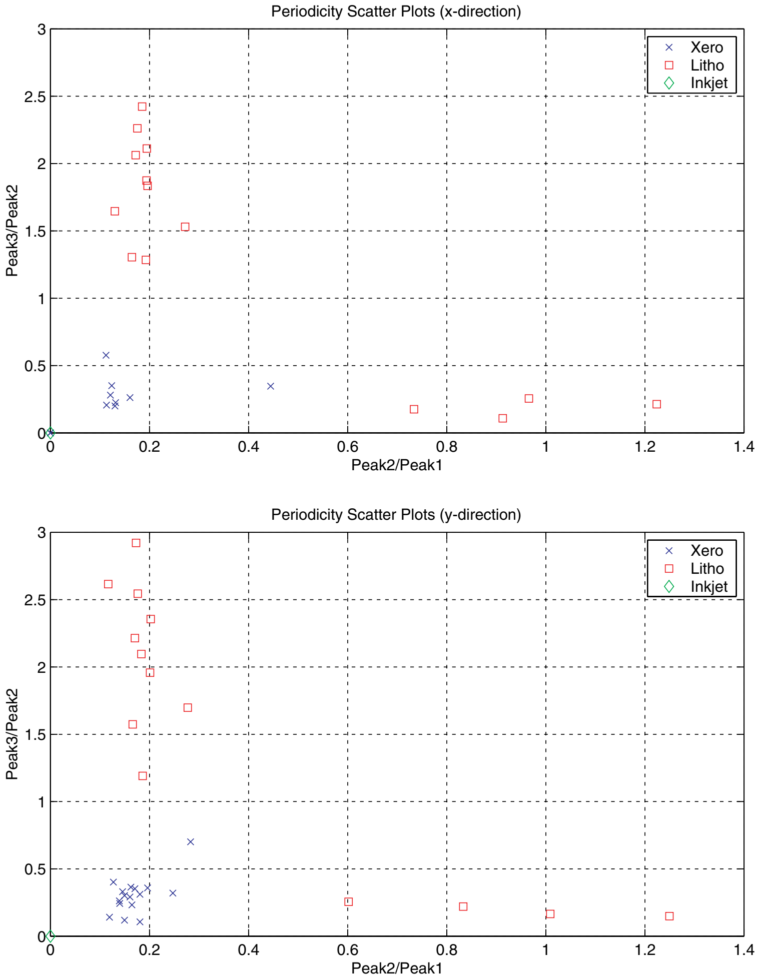


Fig. 16. Scatter plots for Peak3/Peak2 versus Peak2/Peak1 for images in the training set.

cluster together at smaller but postive values for both these peak ratios. Inkjet images on the other hand are all tightly clustered close to the origin (since only the first peak could be detected for this media, the peak ratios are taken to be zero). Scatter plots are in agreement for both x and y directions.

Using the results of the training set, we find parameters to separate the decision regions. The absence of a secondary and tertiary peak, which maps to the origin (Region 0 in Fig. 17) defines the decision region corresponding to the Inkjet classification. The decision region corresponding to Xerographic media is the rectangular region between $0 <$ Peak2/Peak1 < 0.5 and $0 <$ Peak3/Peak2 < 0.7 (Region 1 in Fig. 17). The decision region for Lithographic media is then set as all the remaining areas on the 2-D plane (Regions $\{2, 3, 4\}$) in Fig. 17. The manner in which we arrange the axes of the scatter plots reveals the high frequency characteristics of the lithographic patterns (low noise, and stronger periodicity features), thus lithographic images reside at the larger valued regions of at least one of the two ratios.

4.2. *Final decision*

Final decision criteria proceeds by giving precedence to distance method which enables a three-way classification. Hence if the image marking process falls into any one of the three regions, i.e. regions $(0, 1, \{2, 3, 4\})$ in Fig. 17, in both x and y directions, then the image is classified as the corresponding media. If there is a discrepancy between the results of these two distance measures, then dispersion measure, A_H, i.e. the area under H_D is checked. If that gives xero/litho (regular patterns) decision, we can do a further classification as follows. If the periodicity measure of an image in one of the directions results in Region 1, and in the other direction results in Region 3, this implies that there is a strong third peak, i.e. a high frequency in one of the directions, hence the media can be classified as lithographic. In contrast, if the periodicity measure of the image in one of the directions results in Region 1, and the other in Region 2, this implies a still weak third peak, and a little stronger second peak. This is selected to come from a xerographic media, hence the

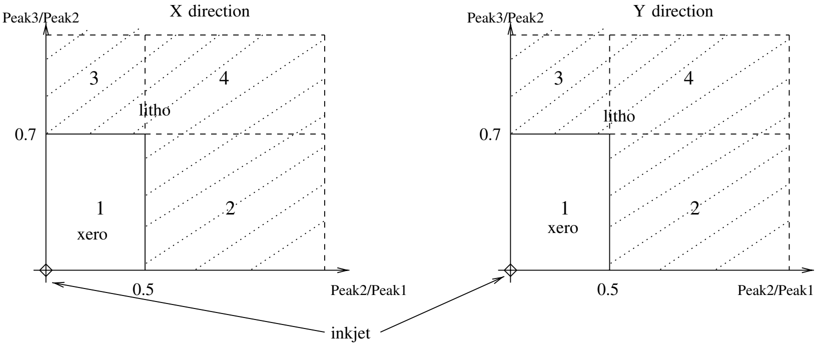


Fig. 17. Decision regions obtained at the end of the training phase.

classification. If no classification could be made up to this point, by periodicity measure, and the dispersion measure has classified as aggregated patterns, i.e. inkjet, then this result is accepted, i.e. decision is made in favor of inkjet media.

4.3. *Testing phase*

Test 1: In order to test our decision criterion and to see how default parameters work on previously unobserved images, we tested the classification method with a small test set of 30 scanned images which are reproduced on one of the media: inkjet, xerography, lithography and photo. The scatter plots for periodicity measure are given in Fig. 18. With the decision boundaries obtained from the training set, only one lithographic image from the test set (a lithographic image from a scanned

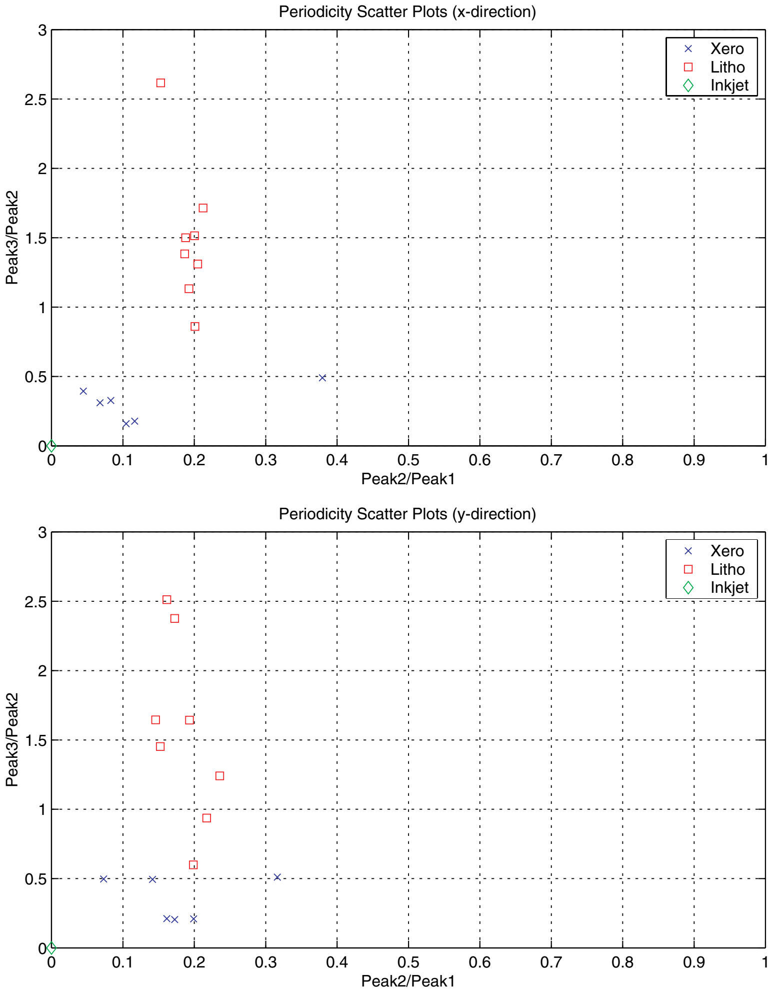


Fig. 18. Scatter plots for Peak3/Peak2 versus Peak2/Peak1 for images in the test set.

advertisement for furniture upholstery), results in its periodicity data in y direction falling into xerographic decision region. However, with the convention as explained in the previous paragraph, that is its other direction periodicity being in Region 2, can correctly classify it also as lithographic. The test set included one line screen xerographic image. Although it is classified as xerographic correctly, line screen information was lost due to a small additional peak very close to being shallow but not labeled as such by the threshold we have set for shallowness. Nevertheless, except for these two cases, all images in the test set are classified correctly. A collection of photographic images from an image repository at Xerox were also correctly identified.

Test 2: An extensive testing is performed using a number of print-outs from Xerox and HP printers. This test set consisted of 315 scanned documents printed from the following: Xerox 4925, Xerox Phaser440, Xerox DC-12, Hakuba400, and HP1200c. Three Xerox Scanners 705968, 708585 and 710350 were used. The original data printed on each of the different printers were taken from a photograph original, one from a newspaper, and also printing of same image by different print technologies. The dataset consisted of 66 inkjet, 123 xerographic, 97 lithographic, 29 photographic images of varying colors and content.

The confusion matrix in Table 1 summarizes the results for the second test dataset. We see that the confusion for photographic media was 9% towards lithographic media, with a correct classification rate of 91%. The inkjet classification was correct 88% of the time with the errors distributed among the other three media. For xerography, correct classification was 87% and confusion was mostly for inkjet. Lithographic media was confused with photographic with a 17% with a correct classification of 83%.

Since the test results reported above were conducted on actual scans, these included various forms of noise (scanner noise, printing process noise, etc). Therefore, the confusion matrix summarizes results that are representative of actual noisy images. We expect that better results may be obtained on synthetically generated images but these would not be representative of the actual performance in practice.

In this extensive dataset of Test 2, there were a larger number of different printer technologies and models on which the media were produced. In addition, three different scanners were used for the evaluation. This resulted in a minor increase in the classification errors compared to Test 1 that included one scanner and one representative printer of each type. We observed a few systematic sources for the

Table 1. Confusion matrix for the test set 2.

	Photo	Inkjet	Xero	Litho
Photo	0.9091	0	0	0.0909
Inkjet	0.0407	0.878	0.0163	0.0650
Xero	0	0.1237	0.8660	0.0103
Litho	0.1724	0	0	0.8276

misclassifications in this dataset. In particular, there were newspaper images, which were detected as photographic although they were truly lithographic. The primary reason for this misclassification is that the newspaper images consisted mainly of text whereas the algorithm was designed for the classification of pictorial images. In addition, errors were observed for tests where images were rotated on the scanner platen in the scanning process by angles of 30, 45, and 60 degrees. In this rotated scan scenario, the scan consisted of a rectangular region including the image and a relatively large portion of the scanner backing — which was close to black or white. These large background regions caused problems for the classification algorithm and in some cases, lithographic media were classified as photographic. Both these problems are, however, not fundamental to the algorithm and are readily remedied when the proposed algorithm is used within a larger system – as would typically be the case. In these situations, typically scan document analysis techniques are utilized to first segment the page into a number of different regions corresponding respectively to text, pictorial, and graphics content (See for example[34]). The proposed algorithm would then be applied, without incurring these parasitic problems, to the individual pictorial regions.

The computational performance of the algorithm was also evaluated. For the images in Test Set 2, using a C language implementation on a 2.0 GHz Pentium processor, the average, median, maximum, and minimum computation times for the classification of a scanned image were 1.34, 1.4, 2.54, and 0.08 seconds, respectively.

5. Conclusions

The four primary color image reproduction technologies, viz. photography, lithography, xerography and inkjet printing, employ processes with clearly distinguishable spatial statistics. In this paper, we exploit this fact to develop a fully automated and computationally efficient approach for the classification of input media type based on the spatial statistics of the scanned image.[32] These spatial statistics are founded on the mathematical tools provided by the spatial statistics literature such as the dispersion measure which models the global first order properties of a spatial point process and the periodicity measure, i.e. the inter-event distance measure, which models the local first order properties of a spatial point process. The point patterns are extracted from small color blocks and utilized as samples from a stochastic spatial point process. The system allows classification of scanned images into the four main categories based on scanned image data alone without the use of any additional sensors. Experimental results demonstrated that the proposed classifier is efficient and achieves reasonable and acceptable accuracy.

Appendix A: Impact of Noise on Periodicity

In order to gain additional insight into how noise can lead to differing observable spatial statistics, we consider a simple stochastic model for halftones where the

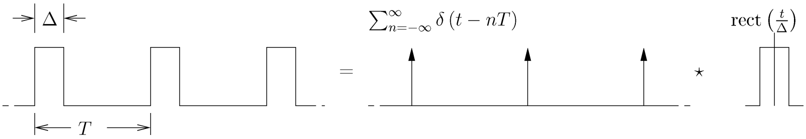

Δ
T
$\sum_{n=-\infty}^{\infty} \delta(t-nT)$
rect $\left(\frac{t}{\Delta}\right)$
=
$\star$

Fig. 19. Model for a halftone (one dimensional illustration, noiseless case).

halftone (representative of a constant image region) is generated by the convolution of a "spot" function with a periodic lattice representative of the halftone. This model has been extensively used for the analysis of halftone periodicity.[3,23,25,30,32] In order to illustrate the concept, we consider a simple one-dimensional version of the model illustrated in Fig. 19, where a periodic function representing a halftone with period T and area coverage $\Delta < \frac{T}{2}$ is represented as a convolution of a pulse $g(t)$ with a periodic impulse train. The noiseless halftone signal is thus represented by a convolution as

$$x(t) = g(t)\star \sum_{n=-\infty}^{\infty} \delta(t-nT) = \sum_{n=-\infty}^{\infty} g(t-nT) \tag{A.1}$$

where $\star$ denotes convolution and in Fig. 19, $g(t) = \mathrm{rect}\left(\frac{t}{\Delta}\right)$ with $\mathrm{rect}(t)$ defined as the standard "rect" function:

$$\mathrm{rect}(t) = \begin{cases} 1 & \text{if } -\dfrac{1}{2} \le t \le \dfrac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{A.2}$$

In the presence of noise, each of the halftone spots will be different and the observed halftone function will therefore not be strictly periodic. The model of (A.1) may appropriately be modified to represent the noisy halftone as

$$x(t) = \sum_{n=-\infty}^{\infty} g_n(t-nT) \tag{A.3}$$

where $g_n(\cdot)$ represents the stochastically varying halftone pulses that are different from realization to realization. Signals of this form have been extensively studied in digital communications. Under the assumption of stationarity, the power spectrum of the signal consists of a discrete line spectrum and a continuous spectrum. It can be shown that (see, for example [Ref. 23, pp. 202–205]), the discrete spectrum is determined entirely by the mean $\bar{g}(t) = \mathrm{E}\{g_n(t)\}$ as

$$S_x^d(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \left|\bar{G}\left(\frac{n}{T}\right)\right|^2 \delta\left(f - \frac{n}{T}\right) \tag{A.4}$$

where $\bar{G}(\cdot)$ denotes the Fourier transform of $\bar{g}(t)$.

We next compare the two scenarios for halftones with an area coverage Δ, where we assume $\Delta < \frac{T}{2}$: noiseless and noisy halftones. The first clean noiseless halftone

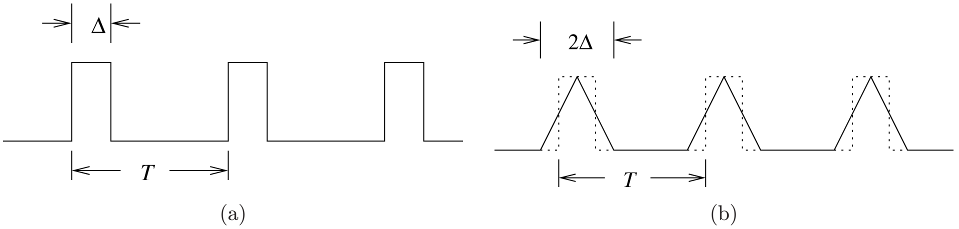


Fig. 20. Mean $\bar{g}(t)$ of halftone pulses for (a) noiseless halftone $h_1(t)$ and (b) a noisy halftone $h_2(t)$.

case is illustrated in Fig. 20(a). In this case, we have $\bar{g}(t) = h_1(t) = \mathrm{rect}\left(\frac{t}{\Delta}\right)$. For the second, we assume that halftone noise in the periphery of the printed regions (such as is typical for xerographic printing systems) causes the spreading of $g_n(t)$ from the ideal shape Fig. 20(b). For simplicity of analysis, we further assume that the net effect of the spreading produces a mean value

$$\bar{g}(t) = h_2(t) = \mathrm{rect}\left(\frac{t}{\Delta}\right) \star \mathrm{rect}\left(\frac{t}{\Delta}\right). \tag{A.5}$$

The mean $\bar{g}(t)$ of halftone pulses for our two idealized situations is illustrated in Fig. 20. Note that our assumption ensures that the two halftones have the same spatial average value of $\frac{\Delta}{T}$ in the two cases. Using (A.4), we now obtain the discrete component of the power spectrum (in our idealized scenario) for the noiseless case as

$$S_x^{1,d}(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \mathrm{sinc}^2\left(\frac{n\Delta}{T}\right) \delta\left(f - \frac{n}{T}\right) \tag{A.6}$$

and for the noisy case as

$$S_x^{2,d}(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} \mathrm{sinc}^4\left(\frac{n\Delta}{T}\right) \delta\left(f - \frac{n}{T}\right) \tag{A.7}$$

Note that the fall-off in the discrete spectrum is more rapid for the noisy case than for the noiseless case (as $\mathrm{sinc}^4(\cdot)$ instead of $\mathrm{sinc}^2(\cdot)$). It can be observed that the noise leads to faster fall-off in higher order spectral lines therefore causing weaker periodicities at higher frequencies. Motivated by the difference in spectral characteristics induced by the noise, we adopted the discriminating statistic presented in Secs 3.4 and 3.4.1, which uses the ratio of the second peak strength to the first peak strength in order to distinguish between the noise characteristics of the halftones.

Acknowledgments

We thank Bill Fuss at Xerox Research Center, Webster, NY, for his assistance in various parts of the research effort we report here, and also for several useful discussions. In particular, we are thankful to him for refining and optimizing the software

implementation and for obtaining several of the results reported for the extensive test dataset.

References

1. M. Acharyya and M. Kundu, Document image segmentation using wavelet scale-space features, *IEEE Trans. Circuits and Syst. Vid. Technol.* **12**(12) (2002) 1117–1127.
2. R. Bala and G. Sharma, System optimization in digital color imaging, *IEEE Sig. Proc. Mag.* **22**(1) (2005) 55–63.
3. F. Baqai and J. Allebach, Computer-aided design of clustered-dot color screens based on a human visual system model, *Proc. IEEE* **90**(1) (2002) 104–122.
4. R. A. Calvo and H. A. Ceccatto, Intelligent document classification, citeseer.ist.psu.edu/calvo00intelligent.html (2000).
5. P. Clark and F. Evans, Distance to nearest neighbor as a measure of spatial relationship in populations, *Ecology* **35**(4) (1954) 445–453.
6. N. Cressie, *Statistics for Spatial Data* (John Wiley and Sons, New York, 1991).
7. F. David and P. Moore, Notes on contagious distributions on plant populations, *Ann. Bot. Lond.* **18** (1954) 47–53.
8. R. de Queiroz and R. Eschbach, Fast segmentation of the jpeg compressed documents, *J. Electron. Imag.* **7**(2) (1998) 367–377.
9. P. Diggle, *Statistical Analysis of Spatial Point Patterns* (Academic Press, London, 1983).
10. R. Duda and P. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
11. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic Press, Boston MA, 1990).
12. A. Gatrell, T. Bailey, P. Diggle and B. Rowlingston, Spatial point pattern analysis and its application in geographical epidemiology, *Trans. Insti. British Geographers* (1996).
13. C. M. Hains, S. Wang and K. T. Knox, Digital color halftones, in *Digital Color Imaging Handbook*, ed. G. Sharma (CRC Press, Boca Raton, FL, 2003), Chapter 6.
14. R. Haralick, K. Shanmugam and I. Dinstein, Textural features for image classification, *IEEE Trans. Syst. Man, and Cybern.* **3** (1973) 610–621.
15. H. R. Kang, Color scanner calibration, *J. Imag. Sci. Technol.* **36**(2) (1992) 162–170.
16. S. Kumar, R. Gupta, N. Khanna, S. Chaudhury and S. Joshi, Text extraction and document image segmentation using matched wavelets and mrf model, *IEEE Trans. Image Process.* **16**(8) (2007) 2117–2128.
17. K.-H. Lee, Y.-C. Choy and S.-B. Cho, Geometric structure analysis of document images: A knowledge-based approach, *IEEE Trans. Patt. Anal. Mach. Intell.* **22**(11) (2000) 1224–1240.
18. S.-W. Lee and D.-S. Ryu, Parameter-free geometric document layout analysis, *IEEE Trans. Patt. Anal. Mach. Intell.* **23**(11) (2001) 1240–1256.
19. Y. Li and A. Jain, Classification of text documents, *Comput. J.* **41**(8) (1998) 537–546.
20. S. Marinai, S. G. M. and G. Soda, Artificial neural networks for document analysis and recognition, *IEEE Trans. Patt. Anal. Mach. Intell.* **27**(1) (2005) 23–25.
21. G. Nagy, Twenty years of document image analysis in PAMI, *IEEE Trans. Patt. Anal. Mach. Intell.* **22**(1) (2000) 38–62.
22. T. Pavlidis and J. Zhou, Page segmentation and document classification, *CVGIP: Graph. Mod. Imag. Process.* **1**(54) (1992) 484–496.
23. J. Proakis, *Digital Communications*, 4th ed (McGraw-Hill, New York, 2001).

24. P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez and T. Tuytelaars, A thousand words in a scene, *IEEE Trans. Patt. Anal. Mach. Intell.* **29**(9) (2007) 1575–1589.
25. T. Rao and G. Arce, Halftone patterns for arbitrary screen periodicities, *J. Opt. Soc. Am. A* **5**(9) (1988) 1502–1511.
26. B. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, Cambridge, 1996).
27. G. Sharma, Methods and apparatus for identifying marking process and modifying image date based on image spatial characteristics, US Patent No. 6353675, 05 March 2002.
28. G. Sharma and S. Wang, Spectrum recovery from colorimetric data for color reproductions, *Proc. SPIE: Color Imaging: Device Independent Color, Color Hard Copy, and Applications VII*, eds. R. Eschbach and G. G. Marcu, Vol. 4663 (2002), pp. 8–14. [Online]. Available: http://www.ece.rochester.edu/gsharma/papers/specrecei02.pdf.
29. M. Tuceryan and A. Jain, Texture analysis, in *The Handbook of Pattern Recognition and Computer Vision*, eds. C. Chen, L. Pau and P. Wang (World Scientific, 1998), Chapter 2.
30. R. Ulichney, *Digital Halftoning* (MIT, MA.: The MIT Press, 1987).
31. G. Unal, G. Sharma and R. Eschbach, Efficient classification of scanned media using spatial statistics, *IEEE Int. Conf. Image Processing* (2004), pp. 2395–2398.
32. G. Unal, G. Sharma and R. Eschbach, Systems and methods for estimating an image marking process using event mapping of scanned image attributes, Published US Patent Application, 20050134934.
33. G. Upton and B. Fingleton, *Spatial Data Analysis by Example* (John Wiley and Sons, New York, 1985).
34. Y. Yang, An evaluation of statistical approaches to text categorization, *Inform. Retri. J.* **1**(1–2) (1999) 67–88.
35. H. Cheng and C. A. Bouman, Multiscale bayesian segmentation using a trainable context model, *IEEE Trans. Image Proc.* **10**(4) (2001) 511–525.

Gozde Unal received her Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 2002, and then was a postdoctoral fellow at the Georgia Institute of Technology, Atlanta, GA, USA, for a year. Between 2003 and 2007, she worked as a research scientist at Siemens Corporate Research, Princeton, NJ, USA. She joined the faculty of Sabanci University, Istanbul, Turkey in Fall 2007, where she is currently an assistant professor. She is a Senior Member of the IEEE, and an Associate Editor for *IEEE Transactions on Information Technology in Biomedicine*.

Her current research is focused on medical image analysis, segmentation, registration, and shape analysis techniques with applications to clinically relevant problems in MR, CT, US, and intravascular images.

Reiner Eschbach received his Ph.D. in physics in 1986. In 1988 he joined the Xerox Innovation Group where he is now a Research Fellow.

His research interests include image processing, color imaging, document automation, digital halftoning and compression. He is the former Secretary and Publications Vice President of the IS&T Board of Directors and Fellow of the Society for Imaging Science & Technology (IS&T)

Gaurav Sharma is an associate professor at the University of Rochester in the Department of Electrical and Computer Engineering and in the Department of Biostatistics and Computational Biology. He is also the Director for the Center for Emerging and Innovative Sciences (CEIS), a New York state funded center for promoting joint university-industry research and technology development, which is housed at the University of Rochester. He received the B.E. degree in electronics and communication engineering from Indian Institute of Technology Roorkee (formerly Univ. of Roorkee), India in 1990; the M.E. degree in electrical communication engineering from the Indian Institute of Science, Bangalore, India in 1992; and the M.S. degree in applied mathematics and Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh in 1995 and 1996, respectively. From Aug. 1996 through Aug. 2003, he was with Xerox Research and Technology, in Webster, NY, initially as a member of research staff and subsequently at the position of principal scientist.

Dr. Sharma's research interests include media security and watermarking, color science and imaging, genomic signal processing, and image processing for visual sensor networks. He is the editor of the *Color Imaging Handbook*, published by CRC press in 2003. He is a senior member of the IEEE, a member of Sigma Xi, Phi Kappa Phi, Pi Mu Epsilon, IS&T, and the signal processing and communications societies of the IEEE. He was the 2007 chair for the Rochester section of the IEEE and served as the 2003 chair for the Rochester chapter of the IEEE Signal Processing Society. He currently serves as the chair for the IEEE Signal Processing Society's Image Video and Multi-dimensional Signal Processing (IVMSP) technical committee. He is a member of the IEEE Signal Processing Society's Information Forensics and Security (IFS) technical committee and an advisory member of the IEEE Standing Committee on Industry DSP. He is an associate editor the *Journal of Electronic Imaging* and in the past has served as an associate editor for the *IEEE Transactions on Image Processing* and *IEEE Transactions on Information Forensics and Security*.