EFFICIENT CLASSIFICATION OF SCANNED MEDIA USING SPATIAL STATISTICS

Gozde Unal
Siemens Corporate Research
Princeton, NJ 08540

Gaurav Sharma
University of Rochester
Rochester, NY 14627

Reiner Eschbach[*]
Xerox Corp.
Wester, NY 14580

ABSTRACT

We address the automatic classification of scanned input media in order to improve color calibration. Since scanner responses vary significantly according to the type of input, a media dependent color calibration for a scanner is desirable for accurately mapping scanner responses to a standard color space. To assist such media dependent calibration, 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 medium and the spatial statistics of corresponding images, which may be observed in the scanned images. Adopting two spatial statistical measures of dispersion and periodicity, and utilizing extensive training data, we determine well separated decision regions to classify the input medium with a high confidence level. Experimental results over an independent test data set validate the results.

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 purposes of copying or for creating an electronic representation for use in various applications. 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 [1, 2].

The use of a media dependent calibration requires identification of the input medium 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 medium type is a preferable alternative. This paper describes an automated medium classification system based on the scanned image data itself with no additional sensors.

Our classification technique is based on the strong correlation between the four main types of input media - photographic, lithographic, xerographic, and inkjet and the spatial characteristics of the corresponding reproduction processes. Photographic reproduction is 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 [3]. Lithographic reproduction typically uses a higher halftone frequency and has a lower noise than xerography. Analysis of the underlying halftone/contone spatial characteristics of scanned image data can therefore be used to identify the input medium with a reasonable degree of confidence. 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 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.

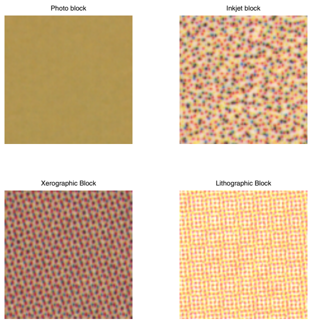

Fig. 1. Subregions from images on different media

Though classification based on 2-D power spectra has been proposed earlier [4, 5], real-time implementation requires classifiers that are much more computationally efficient. In this paper, we present one such method based on spatial analysis of point processes. We generate point patterns from small blocks of scanned data that are representative of the underlying halftone processes and analyze them using two spatial statistics, namely the dispersion measure, and periodicity measure. The measures are finally used in a decision criterion to classify the input medium type. Experimental studies show that the input media can be identified correctly to a high degree of confidence using the proposed method.

2. BACKGROUND ON SPATIAL STATISTICS

A simple 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 a study region. Formally a Homogeneous Poisson Process(HPP) is equivalent to CSR [6], characterizing absence of

[*] This work was conducted while all three authors were at Xerox Corp.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

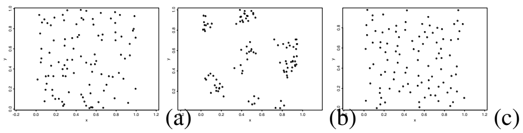

Fig. 2. Realizations of spatial point processes: (a) homogeneous Poisson process (b) aggregated (c) regular process.

structure in the data(Fig. 2(a)). Insight on the structure of a point pattern may be gained by testing the deviation from CSR. There are two types of deviations: If events at short distances occur more frequently than is expected under CSR, and the pattern has a more uneven intensity of points with local peaks at aggregations [7], the pattern is called **aggregated** (Fig. 2(b)). Patterns that have an evenness in distribution are called **regular** patterns. They exhibit more large inter-event distances than a CSR process (Fig. 2(c)).

First and second order properties often provide useful information on a stochastic process even when full characterization is difficult. First order properties may be analyzed, for instance, using (a) **area-based** methods that rely on frequency distribution of observed numbers of events in subregions (quadrats) of the study area or (b) **distance-based** metrics that use information on the distances between events to characterize the pattern.

Quadrat Methods: Quadrat sampling refers to collecting counts of events in quadrats. Under a hypothesis of CSR, the distribution of number of points per quadrat is Poisson with mean λ. A **dispersion measure** is given by

$$D = \frac{s^2}{\bar{x}} - 1, \qquad (1)$$

where $\bar{x}$ (s^2) is the sample mean (variance) of event counts in a quadrat. The expected value of the dispersion index is 0 for a random pattern (HPP), negative for regular processes, and positive for aggregated processes. The index therefore distinguishes two distinct types of departures from an HPP.

Distance Methods:

An alternative first order statistics is the mean nearest-event distance $\bar{y}$. Intuitively, small values of $\bar{y}$ indicate aggregation, whereas large values of $\bar{y}$ indicates regularity. Diggle [8] has suggested a test based on the entire empirical distribution function of nearest-event distances. 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 HPP. The significance of the test can be evaluated using Monte Carlo simulations.

Which Tests to Use?: The dispersion measure (1) is a strongly established statistic for testing CSR via quadrat count analysis. However, the choice of quadrat size is important in its computation. If the structure of the pattern is detectable using a single scale (quadrat size), then index of dispersion D is straightforward for testing two-way departures from CSR. D provides a **global** test to catch heterogeneity which usually manifests itself in terms of aggregation. Distance-based methods emphasize **local** characteristics, thus are more sensitive to aggregation and regularity, and they are less sensitive to the choice of correct scale. A combination of these two techniques is therefore used for media classification, as outlined in the following section.

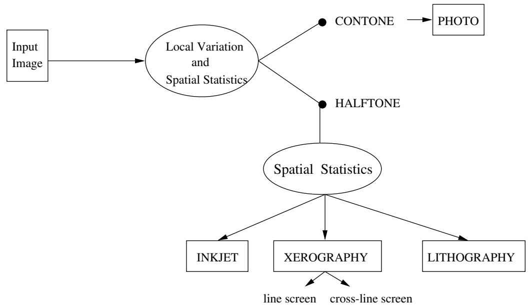

Fig. 3. Decision Tree for the Media Identification Problem.

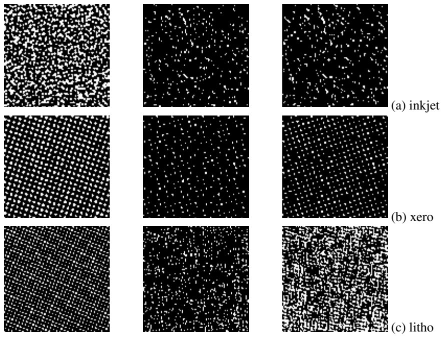

Fig. 4. Level sets obtained from image channels of different media. Columns 1: red, 2: green, 3: blue channel.

3. HALFTONE/CONTONE PATTERNS

The decision tree which is depicted in Fig. 3, explains the sequential decisions that are to be made for media identification problem. The first step differentiates between the two major types of processes, contone and halftone. Blockwise local variation of the scanned data is used to detect photographic media. If the standard deviation of the block is less than some experimentally predetermined threshold value, then the block is a candidate **constant** block with very small spatial variation, labeled as photographic. The remaining blocks are labeled as **varying**.

Point Pattern Generation: Each scanned image block is processed to extract a point pattern. Close investigation of halftone dot patterns in Fig. 1 reveal the fact that a group of dots with a certain color value when viewed separately from dots with other colors, can represent the underlying spatial halftone pattern. This is however equivalent to viewing a color level set of the color block. We generate level sets from each channel separately or from color level sets (additional details may be found in [9]). In Fig. 4 (a), the point patterns clearly show a dispersed and aperiodic structure in which there are some regions which do not contain events (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.

3.1. Dispersion Measure

The measure D in Eq. (1), is calculated over each point pattern extracted out of blocks over the image. 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 CSR: the regular patterns.

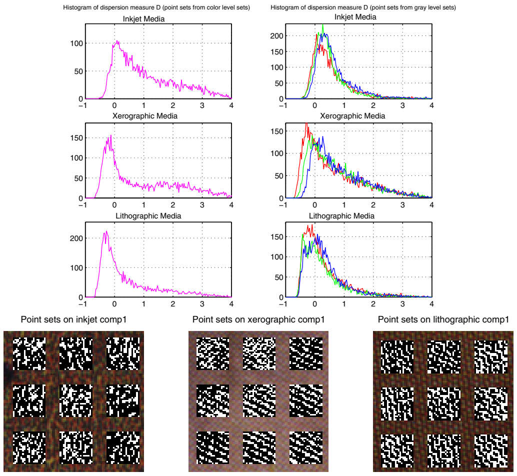

Fig. 5. Histograms of dispersion measure D calculated over an image on three different media (on different rows). (Column 1: from color level set,2: from gray level sets). Bottom:Sample point sets from the image.

For real scanner data, considerable variation may be seen in the D values over an image. For a scanned inkjet media, the normalized histogram of D over blocks of the image is a peaked curve which is positively skewed. Xerographic and lithographic media, on the other hand yield a peaked D distribution which is negatively skewed. Xerographic and lithographic media, on the other hand produce a peaked D distribution which is negatively skewed. These expected results are observed in Fig. 5(top)

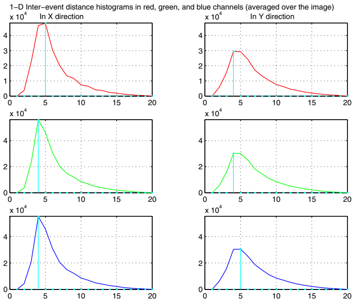

Fig. 6. Inter-event distance histograms on inkjet medium.

A few sample blocks from a composite-content image on inkjet, xerographic, and lithographic media along with the point patterns generated through color level sets are shown in Fig. 5(bottom). To account for the noise effects of real data and the positive-biasedness of the dispersion measure, the decision to distinguish regular patterns from Poissonian or aggregated patterns, the areas of the histogram $\int_{-\beta}^{0} H_D(x)dx$, and $\int_{0}^{\beta} H_D(x)dx$ are compared. $\beta \in (0, 1)$ is a parameter which will be chosen as 0.5.

3.2. Periodicity Measure

For computational simplicity, we utilize a slightly modified 1-D version of the inter-event distance spatial statistic. Histograms of inter-event distances in 1-D in x and y directions are calculated and accumulated over blocks of the image. Instead of MC simulations, we determine a sample average of the empirical distribution function of 1-D inter-event distances which provides a global estimate of the distribution of halftone dot patterns for the specific media (reproduction technology).

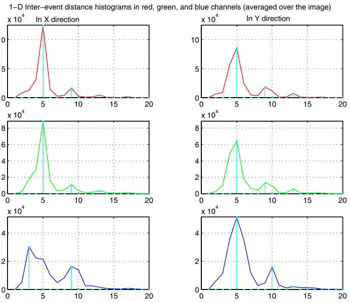

Fig. 7. Inter-event distance histograms on xerographic medium.

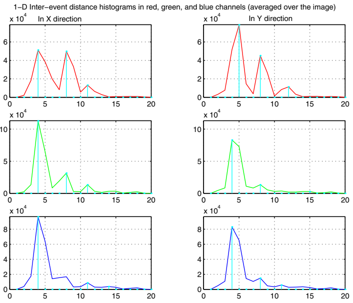

Fig. 8. Inter-event distance histograms on litho. medium.

The inter-event distance captures inherent periodicity (or lack thereof) in the input scan data. It can be observed as expected from the unimodal density plots in Fig. 6 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. Inter-event distance histograms for the composite image reproduced on xerographic medium are shown in Fig. 7. The presence of secondary and tertiary peaks in these plots 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.

For the composite image reproduced on lithographic medium, inter-event distance histograms in x and y directions are plotted in Fig. 8, and observed to display a distinctively trimodal character. Existence of a strong second peak in addition to a relatively strong third peak when compared to those of xerographic medium is an indication of stronger periodicity characteristics.

The stronger tertiary peak of the lithographic medium in comparison to xerography may be attributed to the higher noise at the microscopic scale in xerographic printing in comparison to lithographic reproduction. This causes weak periodicities at larger distances (corresponding to higher halftone harmonics) when compared to those of lithography.

4. EXPERIMENTAL RESULTS

A set of 42 images on different media (inkjet, xerographic, lithographic, and photo) were scanned at 600dpi to create a training data set. The resolution is one determining factor in choosing the quadrat size in order to calculate dispersion measure. For our problem, 2×2, 3×3 quadrat sizes are reasonable choices to capture point pattern structure over small blocks of size 20×20. For lower resolution images, 4×4, or 5×5 quadrat size may be needed in order to display larger patterns of halftone dots.

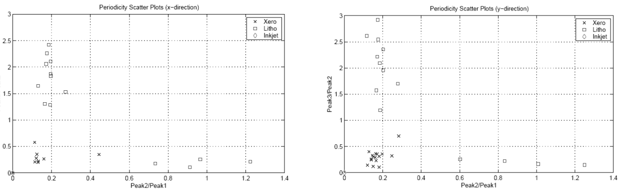

Fig. 9. Peak3/Peak2 vs. Peak2/Peak1 scatter plots (Training).

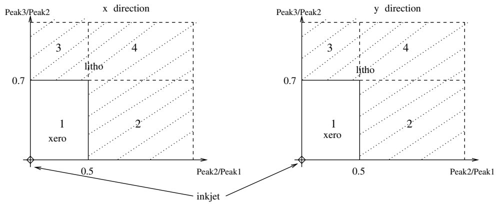

Fig. 10. Decision Regions

Our observations from the previous section lead us to use 2-D decision planes which summarize the information obtained from the inter-event distance histograms of an image in two directions. The final test for a three-way classification of the input media follows by plotting Peak 3, the third detected peak, over Peak 2, vs. Peak 2 over Peak 1. Strength of Peak 3 and Peak 2 produced by lithographic media provides good means of its identification. Thus, the quantities Peak2/Peak1 and Peak3/Peak2 (for both x and y directions) are obtained as the resulting distance measures to be used in the final decision phase. As can be observed from the scatter plots in Fig. 9, there's a distinct separation between different halftone media. Inkjet region is the single point at the origin (see Region 0 in Fig. 10). Xerographic region resides in the rectangular region between $0 <$ Peak2/Peak1 < 0.5 and $0 <$ Peak3/Peak2 < 0.7 (Region 1 in Fig. 10). Lithographic region is set as all the remaining areas on the 2-D plane (Regions $\{2, 3, 4\}$) in Fig. 10.

The final decision criteria gives precedence to the distance method which enables a three-way classification. Hence if image falls into any one of the three regions, i.e. regions $(0, 1, \{2, 3, 4\})$ in Fig. 10, in both x and y directions, then the image is classified as the corresponding medium. If there's discrepancy between the results of these two distance measures, then dispersion measure, 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 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., high frequency in one of the directions, hence the medium 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 might come from a xerographic medium, hence the 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.

In order to test our decision criterion and to see how default parameters work generally, we run with a new test set of 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. 11. With the decision boundaries obtained from the training set, only one lithographic image from the test set (a lithographic image from an ad on upholstery for furniture), results in its periodicity data in y direction fall into xerographic decision region. However, with the convention we take as explained in the previous paragraph, that is its other direction periodicity being in Region 2, can correctly classify it also as lithographic. A collection of photographic images from an image repository at Xerox were also correctly identified.

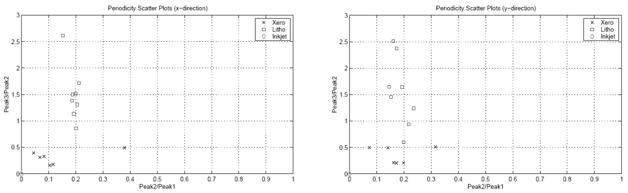

Fig. 11. Peak3/Peak2 vs. Peak2/Peak1 scatter plots (Test).

5. CONCLUSIONS

The four primary color image reproduction technologies, viz. photography, lithography, xerography, and inkjet printing, employ processes with clearly distinguishable spatial 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. The system allows classification of scanned images in to these four categories based on scanned image data alone without the use of any additional sensors. Experimental results indicate that the proposed classifier is efficient and accurate.

6. REFERENCES

[1] H. R. Kang, "Color scanner calibration," *J. Imaging Science and Technology*, vol. 36, no. 2, pp. 162–170, Mar./Apr. 1992.

[2] G. Sharma and S. Wang, "Spectrum recovery from colorimetric data for color reproductions," in *Proc. SPIE: Color Imaging: Device Independent Color, Color Hard Copy, and Applications VII*, R. Eschbach and G. G. Marcu, Eds., Jan. 2002, vol. 4663, pp. 8–14.

[3] C. M. Hains, S. Wang, and K. T. Knox, "Digital color halftones," in *Digital Color Imaging Handbook*, G. Sharma, Ed. CRC Press, Boca Raton, FL, 2003, Chapter 6.

[4] G. Sharma, "Methods and apparatus for identifying marking process and modifying image date based on image spatial characteristics," United States Patent No. 6353675, 05 Mar. 2002.

[5] R. Bala and G. Sharma, "System optimization in digital color imaging," *IEEE Sig. Proc. Mag.*, special issue on Color Imaging (accepted for publication to appear late 2004/early 2005).

[6] N. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, New York, 1991.

[7] G. Upton and B. Fingleton, *Spatial Data Analysis by Example*, John Wiley & Sons, New York, 1985.

[8] P.J. Diggle, *Statistical Analysis of Spatial Point Patterns*, Academic Press, London, 1983.

[9] G. Unal, G. Sharma, and R. Eschbach, "Classification of scanned media using spatial statistics," Tech. Rep., Xerox Corp., 2001.