WORD SPOTTING VIA SPATIAL POINT PROCESSES

Jeffrey C. O'Neill, Alfred O. Hero III, and William J. Williams

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109-2122

ABSTRACT

This paper presents a statistically based method for spotting target words in documents. The crux of the method is the representation of a word by a spatial (planar) point process evolving on a regular lattice of coordinate pairs. This is accomplished by extracting the coordinate pairs, i.e. pixel locations, where the binary bitmap values of the word are non-zero. With this representation the word is completely determined by the spatial intensity function, i.e. the unnormalized spatial probability density function, associated with the extracted set of coordinate pairs. In this work, we use a finite number of moments of the intensity function to characterize the word. Location and scale invariance is obtained by transforming the coordinate pairs to have zero mean and unit variance. Finally, optimal detection strategies are applied to the moments to make the decision.

1. INTRODUCTION

Given a document, it is frequently desirable to know whether the document contains a certain word or set of words [1, 2]. It would be useful if this process could be automated and work reliably on documents with different fonts, font sizes, and noise contamination, e.g. in faxed documents. This paper proposes a statistical method for doing this based on applying optimal detection strategies to a set of moments of the spatial intensity function associated with locations of non-zero valued pixels in the image. Throughout this paper we assume that individual words in the document have been isolated and placed in a rectangular window of specified length and width.

Let $W = W(\underline{x})$ be the image of a word where $\underline{x} = (x_1, x_2)^T$ is a spatial variable which indexes over the n non-zero pixels of the bitmap image, here assumed to lie on a regular lattice. As a first step we would like to find some transformation of the data that would give us location and scale invariance for an isolated word. This is accomplished by using the following spatial point process representation of a given word:

$$W(\underline{x}) = \sum_{i=1}^{N} \Pi(\underline{x} - \underline{x}_i) \tag{1}$$

where $\{\underline{x}_i\}_{i=1}^N$ are the spatial locations of pixels over which W is non-zero, and $\Pi(\underline{x})$ is a $1, 0$ valued rectangular function taking the value 1 on a square pixel at the origin within the rectangular window. Note that the coordinate pairs $\underline{x}_i$ and the integer N will vary depending on the particular word, the font, and any noise contamination. Hence $\{\underline{x}_i\}_{i=1}^N$ is properly modeled as the realization of a random spatial point process $\{\underline{X}_i\}_{i=1}^N$ with intensity function $\lambda(\underline{x})$. The intensity function can be viewed as an unnormalized probability density of the coordinate pairs and is completely characterized by its set of spatial moments $\mu_{p,q} = E[X_1^p X_2^q]$, $p, q = 1, 2, \ldots$. The first order marginal moments (means) $[\mu_{1,0}, \mu_{0,1}]^T$ give the mean location of the word within the rectangular window while the centered second order marginal moments (variances) $[\mu_{2,0} - \mu_{1,0}^2, \mu_{0,2} - \mu_{0,1}^2]^T$ give the scale. By subtracting the means from the pairs $\{\underline{X}_i\}_i$ and dividing by the square root of the variances we obtain a representation of the word which is invariant to scale (font size) and translation (spatial position). Given only a small set of the moments of these invariant coordinate pairs we can effectively discriminate between different words using detection techniques explained below.

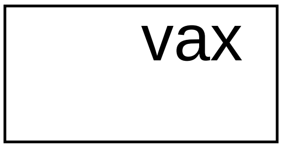
2. SPATIAL MOMENT ESTIMATION

To illustrate, consider the four words shown in Figure 1. Let $\underline{W}$, $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ represent the spatial point processes representing the four words shown in Figure 1a, b, c, and d, respectively. Let $\{\underline{w}_i\}$, $\{\underline{x}_i\}$, $\{\underline{y}_i\}$, and $\{\underline{z}_i\}$ denote realizations, i.e. coordinate pairs of the non-zero elements of the bitmaps. We transform each of the four sets of coordinate pairs so that the sample means and variances are zero and one respectively and denote them as $\hat{\underline{w}}_i$, $\hat{\underline{x}}_i$, $\hat{\underline{y}}_i$, and $\hat{\underline{z}}_i$.

This work was partially supported by the US Dept. of Defense, contract number MDA904-95-C-2157.

(a) $\underline{W}$.

(b) $\underline{X}$.



(c) $\underline{Y}$.

(d) $\underline{Z}$.

Figure 1: Four example words.

Note that $\underline{\hat{w}}_i$ and $\underline{\hat{x}}_i$ each represent the same word in a different location and a different font size. However since we have normalized each of them to have the same mean and the same variance they should have identical moments. On the other hand, $\underline{\hat{w}}_i$ and $\underline{\hat{y}}_i$ also have the same mean and variance, but since they represent different words, the higher order moments will not be identical and they can be discriminated.

We will perform discrimination based on empirical estimates of the centered and normalized moments:

$$m_{p,q} = \frac{E\left[(X_1 - E[X_1])^p (X_2 - E[X_2])^q\right]}{\sqrt{E\left[(X_1 - E[X_1])^{2p}\right] E[(X_2 - E[X_2])^{2q}]}} \quad (2)$$

where $p > 0$ and $q > 0$. We also tried using unnormalized moments, however experiments indicated that the normalized moments provided better discrimination performance.

To illustrate, we calculated four different normalized moments, $m_{1,1}$, $m_{2,1}$, $m_{1,2}$, and $m_{3,1}$, of the words "van" and "vax" in seven different font sizes. The four moments were calculated for each of the fourteen words and are shown in Figure 2. For comparison purposes, each column of Figure 2 is scaled linearly so that the moments range from zero to one. From the figure, it can be seen that moments $m_{1,1}$ and $m_{3,1}$ discriminate between the two words quite well while moments $m_{2,1}$ and $m_{1,2}$ do not discriminate as well. Of course, for a different pair of words, a completely different set of moments could provide the best discrimination. In this paper we will construct an optimal test function for word spotting which is based on all available moments.

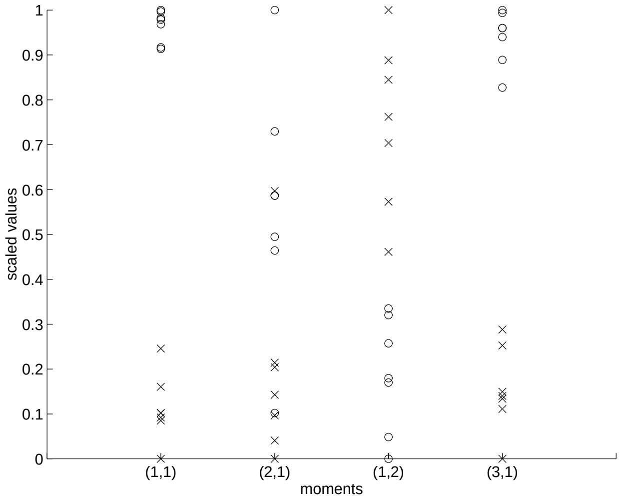


Figure 2: Comparison of the moments for the words "van" and "vax". In each column, the seven "x" represent seven font sizes of "van" and the seven "o" represent the seven font sizes of "vax".

3. OPTIMAL DETECTION STRATEGY

In this paper we focus on the simple problem of distinguishing two words, "van" and "vax" in additive noise, independent of their font size and spatial location. In particular, define the two composite hypotheses H_0 : *van + noise* and H_1 : *vax + noise*. Should it exist, the best test of these hypotheses would be a function of the statistical distribution of the measured spatial point process and would attain the highest possible probability of detection (power) $P(\text{decide } H_1 | H_1 \text{ true})$ subject to any user-specified level of false alarm $P(\text{decide } H_1 | H_0 \text{ true})$. However, the statistical distribution of the spatial point process is difficult to estimate and we will instead focus on constructing a test based on the empirical moments of the point process as described above.

Motivated by the fact that the estimated moments are computed as sums of a large number of binary random variables, we propose the model that over the ensemble of all font sizes the vector of estimated moments $\underline{m}$ is a Gaussian random vector with unknown mean $\underline{\mu}$ and covariance matrix K. We thus have the equivalent set of (*simple*) hypotheses for testing "van" (H_0) against "vax" (H_1):

$$\begin{aligned} H_0: &\quad E[\underline{m}] = \underline{\mu}_0, \ \mathrm{cov}[\underline{m}] = K_0 \\ H_1: &\quad E[\underline{m}] = \underline{\mu}_1, \ \mathrm{cov}[\underline{m}] = K_1 \end{aligned}$$

where $\underline{\mu}_0$, K_0 and $\underline{\mu}_1$, K_1 will be respectively estimated from training sequences of the aggregated population

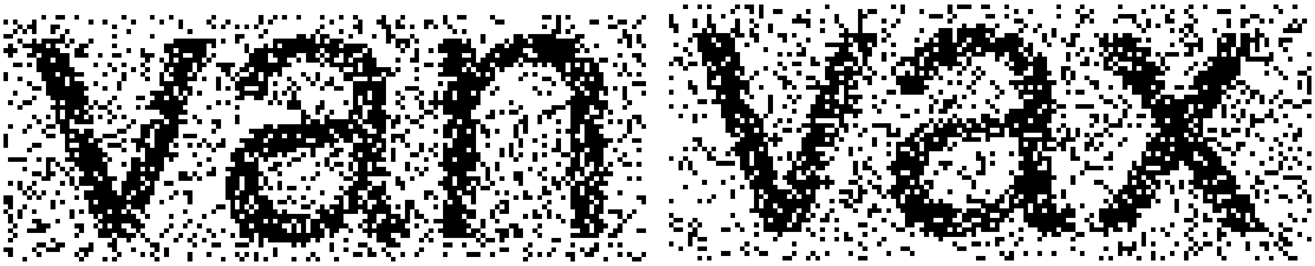


Figure 3: Examples of “van” and “vax” with 20% salt-and-pepper noise.

of “van” and “vax” at all expected font sizes.

The most powerful test between H_0 and H_1 is the likelihood ratio test which simplifies to comparing the difference between two quadratic forms to a threshold: *Decide* H_1 *if*

$$(\underline{m} - \underline{\mu}_0)^T K_0^{-1}(\underline{m} - \underline{\mu}_0) - (\underline{m} - \underline{\mu}_1)^T K_1^{-1}(\underline{m} - \underline{\mu}_1) > \gamma \quad (3)$$

otherwise choose H_0.

Here γ is selected to ensure a given level α of false alarm probability: $P(\text{Decide } H_1|H_0) = \alpha$.

4. NUMERICAL RESULTS

The test (3) was used to discriminate between the word classes {“van” in 13 font sizes} and {“vax” in 13 font sizes}. A 20% level of salt-and-pepper noise was added, i.e. on average one in five pixels were flipped from 1 to 0 or 0 to 1. Two realizations of the noisy bitmaps for “van” and “vax” in 12 point font (images are not to scale) are shown in Figure 3.

To estimate the mean vectors and covariance matrices, we used a training set of 13 different font sizes for each word class. Each word was combined with 20 independent noise realizations for a total of 260 realizations per word. Subsequently, for each realization, a vector of sample moments:

$$\underline{\hat{m}} = [m_{1,1} \;\; m_{2,1} \;\; m_{1,2} \;\; m_{3,1} \;\; \cdots]$$

was estimated by using sample means to estimate $m_{p,q}$ as defined in equation 2. The sample mean vectors, $\underline{\mu}_0$ and $\underline{\mu}_1$, and sample covariance matrices, K_0 and K_1, were then computed for each word class and substituted into the test statistic 3. The performance of any detector is completely characterized by the receiver operating characteristic (ROC) curve which is the plot of $P(\text{decide } H_1|H_1 \text{ true})$ against $P(\text{decide } H_1|H_0 \text{ true})$ (denoted $P(\text{detection})$ and $P(\text{false alarm})$ in Figure 4) [3]. To compute the ROC curve for the detector we used 13 different font sizes of “van” and “vax”, each with 52 unique realizations of the 20% salt-and-pepper noise – corresponding to a total of 676 realizations for computing estimates of the probability of false alarm and the probability of detection for each threshold, γ.

Figure 4 shows the behavior of the ROC curve as the number of moments used in the detection scheme, i.e. the dimension of the vector $\underline{m}$, increases from 1 to 18. From Figure 4, it is seen that the performance of the detector improves as the number of moments used increased to 18. Although not shown on the figure, this trend continued until the number of moments exceeded approximately $0.2 \times n$, where n is the total number of training samples available to estimate the mean and covariance of the moment vector (here $n = 260$).

From Figure 4 it is seen the detector does not show significant improvement as the number of moments is increased from one to three. The reason for this is that the moments $m_{2,1}$ and $m_{1,2}$ are not useful for distinguishing “van” and “vax” as shown in Figure 2. When moment $m_{3,1}$ is added, there is significant improvement since this moment discriminates “van” and “vax” very well.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a method based on higher order moments for spatial point processes for detecting words in low SNR which is robust to changes in font size and spatial location. The detection scheme presented here uses the noisy images without any preprocessing. Median filters and other methods could be used to reduce the amount of noise in the image, but in situations with a very low to signal to noise ratio (SNR), these preprocessing methods become less effective. Since this scheme works well on the unprocessed images in low SNR, this is where it will have the greatest advantage. Another issue that needs to be addressed is the separation of the noisy words in a document. Here we have assumed that the words have already been separated into individual bitmaps. This may not be an

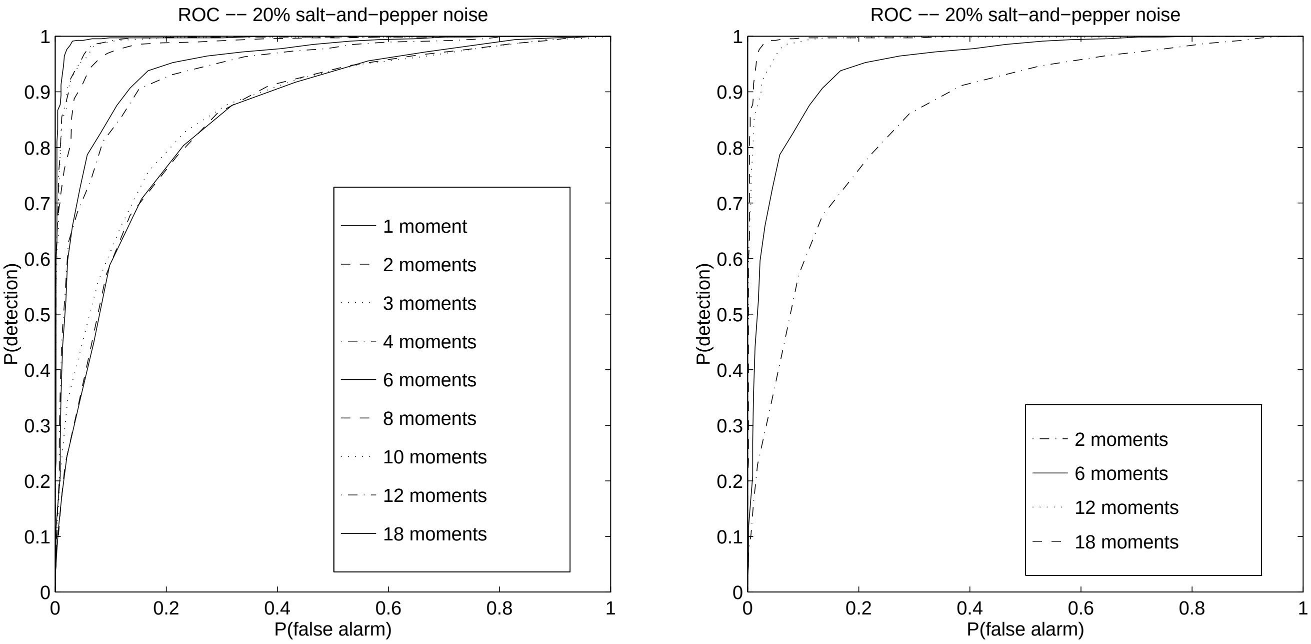


Figure 4: Two sets of ROC curves for the detector. On the left are ROC curves for the detector using 1, 2, 3, 4, 6, 8, 10, 12, and 18 moments. On the right are curves for just 2, 6, 12, and 18 moments.

easy problem and needs to be investigated.

It will be important to investigate detector performance in lower SNR under more challenging types of noise such as fax noise. Other useful, but difficult, additions to this method would be to make the detector invariant to the font (helvetica vs. `courier`) and font style (*italic* vs. **bold**).

Finally we note that in the low noise case, the (noncentered) first-order marginal moments ($m_{0,1}, m_{1,0}$) of the spatial point process give the center of mass of the word which can be used to shift the bitmap to the center of the image plane. Likewise the second order (centered) marginal moments ($m_{0,2}-m_{0,1}^2, m_{2,0}-m_{1,0}^2$) give the vertical-horizontal spatial extent which can be used to scale the bitmap at a standard scale.

6. REFERENCES

[1] F. Chen, L. Wilcox, and D. Bloomberg, "Word spotting in scanned images using hidden Markov modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, volume V, pp. 1–4, Minneapolis, MN, 1993.

[2] S. Kahan, T. Pavlidis, and H. Baird, "On the recognition of printed characters of any font and size," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 9, no. 2, pp. 274–288, Mar, 1987.

[3] H. L. Van-Trees, *Detection, Estimation, and Modulation Theory: Part I*, Wiley, New York, 1968.