WORD SPOTTING VIA SPATIAL POINT PROCESSES, PRINCIPAL COMPONENTS, AND LINEAR REGRESSION[1]

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

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109-2122
email: hero@eecs.umich.edu

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 linear regression and principal component techniques are applied determine what combination of moments will serve to best detect a given word.[1]

1. INTRODUCTION

Given a document, it is frequently desirable to know whether the document contains a certain word or set of words [2, 1]. 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 linear regression and principal components analysis (PCA) 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 n 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 the isolated word. This is accomplished by using the following spatial point process representation of a given word

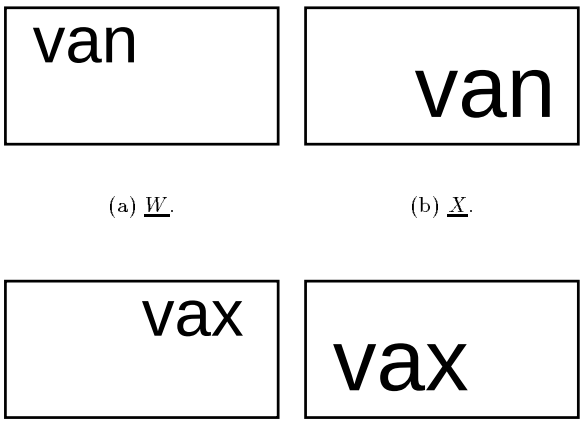
$$W(\underline{x}) = \sum_{i=1}^{N} \Pi(\underline{x} - \underline{x}_i) \tag{1}$$

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 un-normalized probability density of the coordinate pairs and is completely characterized by its set of spatial moments $\mu_{pl} = E[X_1^p X_2^q]$, $p, q = 1, 2, \ldots$. The first order marginal moments (means) $[\mu_{10}, \mu_{01}]^T$ gives the mean location of the word within the rectangular window while the centered second order marginal moments (variances) $[\mu_{20} - \mu_{10}^2, \mu_{02} - \mu_{01}^2]^T$ gives 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 a large set of the moments of these invariant coordinate pairs we can effectively discriminate between different words using regression 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,

[1] 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.

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 $\underline{\hat{w}}_i$, $\underline{\hat{x}}_i$, $\underline{\hat{y}}_i$, and $\underline{\hat{z}}_i$.

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 the higher order moments will not be identical and they can be discriminated.

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

$$\underline{m}_{pq} = \frac{E[X_1^p \; X_2^q]}{\sqrt{E[X_1^{2p}] \, E[X_2^{2q}]}} \tag{2}$$

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

To illustrate we calculated 10 different normalized moments, denoted (with abouse of notation) as $m_1, \ldots, m_{10}$, corresponding to all combinations of p and q such that $p + q <= 5$. The words 'van' and 'vax' were generated in four different font sizes. The ten moments were calculated for each of the eight 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 numbered 1, 4, and 6 discriminate between the two words quite well while other moments do not discriminate as well or not at all. Of course for a different pair of words, a completely different set of moments could provide the best discrimination.

3. LINEAR REGRESSION

In the above section we mentioned that the moments that distinguish a pair of words depends on the pair of words being compared. For a particular class of words it would useful to know which moments best distinguish words. To do this we used linear regression to relate bitmap images of different words to the set of associated moments.

Assume that L words in a class are to be discriminated. For each word we converted the original bitmap image into a list of non-zero coordinates, calculated m moments, and normalized by the sample mean and variance. Translation and scale invariant bitmaps were then recovered by mapping the words back to the image domain using the the normalized coordinate pairs. This ensured that different words are centered in the bitmap and have the same spread. We then put the bitmap images into column vector format by lexicographic ordering of the pixel locations column by column. For each of L words we denote these image vectors by $\underline{y}_1, \ldots, \underline{y}_L$ and we denote the vectors of associated moments by $\underline{m}_1, \ldots, \underline{m}_L$. Define the sample mean vectors: $\underline{\bar{y}} = \frac{1}{L} \sum_{i=1}^{L} \underline{y}_i$ and $\underline{\bar{m}} = \frac{1}{L} \sum_{i=1}^{L} \underline{m}_i$. Finally, form the matrix Y of bitmap vectors and the matrix of associated moment vectors as follows:

$$\begin{aligned} Y &= [\; \underline{y}_1 - \underline{\bar{y}} \quad \underline{y}_2 - \underline{\bar{y}} \quad \cdots \quad \underline{y}_L - \underline{\bar{y}} \;] \\ M &= [\; \underline{m}_1 - \underline{\bar{m}} \quad m_2 - \underline{\bar{m}} \quad \cdots \quad \underline{m}_L - \underline{\bar{m}} \;] \end{aligned}$$

The linear (affine) regression model which describes the relationship between the word bitmaps and the moments is [3]:

$$Y = A\,M + \underline{u}\,\underline{1}^T + \epsilon$$

where ϵ is a matrix of (white) noise residuals. The least squares solution to $\min \|Y - A\,M - \underline{u}\,\underline{1}^T\|$ gives a matrix of regression coefficients[2] A_{LS} and intercept u_{LS}:

$$\begin{aligned} A_{LS} &= Y\,M'\,(M\,M')^{-1} \\ \underline{u}_{LS} &= A_{LS}\,\underline{\bar{m}} + \underline{\bar{y}} \end{aligned}$$

The quantities A_{LS} and $\underline{u}_{LS}$ can now be investigated to tease out the combinations of moments which have the best discrimination power between bitmaps of the L words.

[2] M' denotes the transpose of the matrix M.

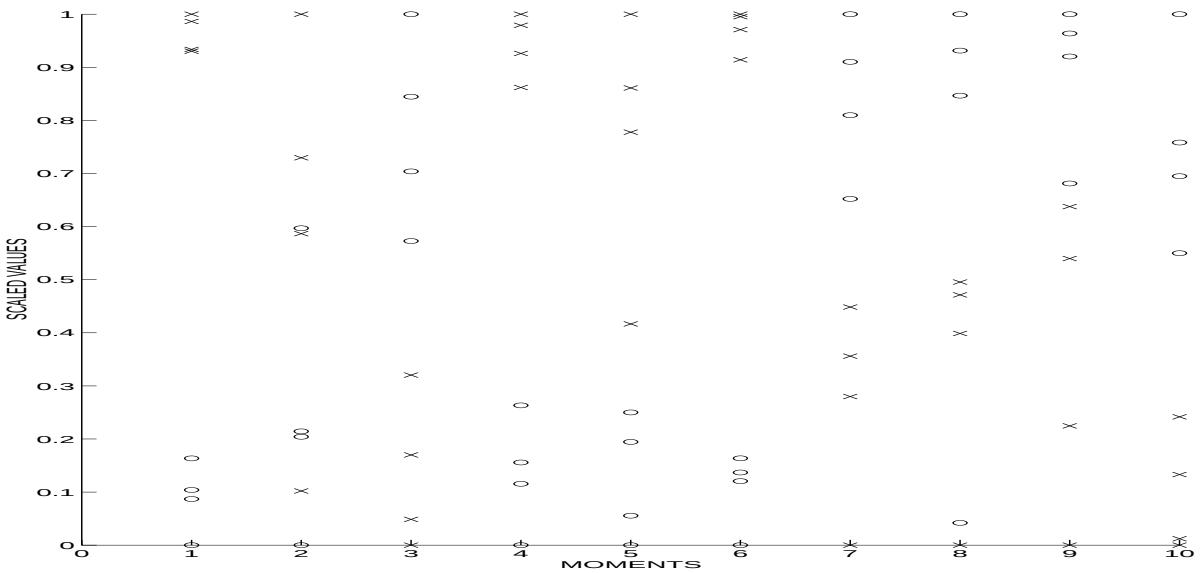


Figure 2: Comparison of the moments for the words 'van' and 'vax'. The four 'o' represent four font sizes of 'van' and the four 'x' represent the four font sizes of 'vax'.

4. SVD ANALYSIS

As written in (3) element ij of the matrix A_{LS} indicates how well the j-th element of the moment vector discriminates at the i-th pixel between the L word bitmaps. The large number (nL) of coefficients in the matrix A_{LS} makes the results difficult to interpret. We circumvent this difficulty by using the singular value decomposition (SVD) to extract the coefficients of A_{LS} which relate the moments directly to the L words.

The SVD of Y is $Y = U\Lambda V$ where U and V are square orthogonal matrices which span the spatial domain (dimension n) and the word class domain (dimension L), respectively. $\Lambda = [D : 0]^T$ is a rectangular matrix of dimension $n \times L$ where D is a an $L \times L$ diagonal matrix of singular values. Substitution of the SVD of Y into the regression equation and multiplication by the inverse $U^{-1} = U^T$ of U gives the modified regression equation

$$\begin{bmatrix} DV \\ 0 \end{bmatrix} = U^T A\, M + U^T \underline{u}\, \underline{1}^T + \tilde{\epsilon}$$

where $\tilde{\epsilon}$ is a (white) noise residual. From this equation we recognize that only the first L rows of $U^T A$ and $U^T \underline{u}$ are related to discrimination between the L words, which are encapsulated by the matrix DV. Hence, neglecting the last $n - L$ rows of the modified regression equation and solving for the least squares solution we obtain the $L \times m$ coefficient matrix A_{LS} and the $L \times 1$ intercept vector: $\tilde{A}_{LS} = U_L^T Y\, M' (M\, M')^{-1}$, $\underline{\tilde{u}}_{LS} = \tilde{A}_{LS}\, \underline{\bar{m}} + U_L^T \underline{\bar{y}}$, where U_L denotes the first L columns of the matrix U. $\tilde{A}_{LS}$ and $\underline{\tilde{u}}_{LS}$ are now much easier to interpret since they relate the moments directly to each of the L words. In particular, a large magnitude of the ij-th element of A_{LS} indicates that the j-th element of the moment vector $\underline{m}$ has strong discrimination power for the i-th word. We have used this modified regression to probe discrimation ability of variaus moments. Numerical results will be presented in the full paper.

5. REFERENCES

[1] F. R. Chen, L. D. Wilcox, and D. S. Bloomberg, " Word spotting in scanned images using hidden Markov models," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Minneapolis, Vol. V, pp. 1-4, 1993.

[2] S. Kahan, T. Pavlidis, and H.S. Baird, "On the recognition of printed characters of any font and size," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, Vol. 9, No. 2, pp. 274-288, Mar. 1987.

[3] D. F. Morrison, *Multivariate Statistical Methods*, McGraw-Hill, 1976.