Cooperation of multi-layer perceptrons for the estimation of skew angle in text document images

Nadine Rondel, Gilles Burel

IEEE 3rd Int. Conf. on Document Analysis and Recognition (ICDAR), Vol. 2, pp. 1441-1444, Aug. 14-16, 1995, Montreal, Canada

Cooperation of Multi-Layer Perceptrons for the Estimation of Skew Angle in Text Document Images

Nadine RONDEL[†‡] & Gilles BUREL[†]
[†] Thomson Broadband Systems, Av de Belle Fontaine, 35510 Cesson-Sévigné, France
[‡] SEFT, 18 rue du Dr. Zamenhof, 92130 Issy-Les-Moulineaux, France

Abstract— *Estimating the Skew Angle in text document images can be a crucial problem in Optical Character Recognition. Based on a new sensor array processing technique, an original solution to Skew Angle Estimation (SAE) is proposed. Thanks to the reformulation of the SAE problem in the framework of Angle of Arrival theory, a fast and accurate method*[1] *is presented, that is based on the cooperation of two neural networks. The first neural net is a three-layer perceptron receiving on input the values of the correlation matrix of the signals; the output is a "rough" estimation of the angle to estimate. This gross estimate is then used to initialize the weights of a second multi-layer perceptron (MLP). The second MLP is built in order to perform a Maximum Likelihood-like optimization, therefore reaching good performances. The system, though trained on simulated radar data, shows good performances on noisy handwritten texts.*

Keywords— *Skew Angle Estimation (SAE), Handwritten texts, Angles of Arrival (AOA), Array-processing, Cooperation of Neural Networks, Maximum Likelihood.*

I. Introduction

One major problem for automatic document analysis is the inclination of text lines. An interest has been shown recently for the estimation of the skew angle, particularly using the Hough transform techniques (Hinds *et al.*, 1990). But, as Aghajan & Kailath (1993b) underline, the technique is computationally expensive and requires a wide memory space. In (Aghajan *et al.*, 1993a) a different approach, based on the reformulation of the problem in an array processing framework is proposed. This approach provides very good results on handwritten typed text, but requires a pre-processing step (in order to reduce the noise, as well as to enhance the linearity of the text lines), which can be difficult to apply to handwritten text.
In order to avoid pre-processing, as well as to be able to deal with both typed and handwritten text, a new method is proposed, that is fast and robust. The system utilizes array-processing theory and cooperation of neural networks. For sake of simplicity, we describe the case where all lines have the same orientation: generalization to many orientations is straight-forward. Consider the scanned image ($L \times K$ pixels) of a text. The problem is to estimate the skew angle θ of the lines (fig 1) and, if required, the number of lines and their respective offset x_i^0. The paper

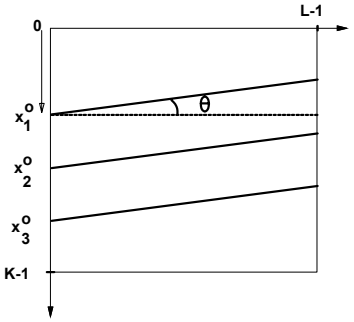

Figure 1: *The skew angle problem*

is organized as follows: part II is an introduction to array processing and Angle Of Arrival (AOA) estimation. The similarity with Skew Angle Estimation (SAE) is shown in Part III. Then, theoretical considerations yield to the proposed approach (part IV). A summary is given in part V, and part VI shows results obtained on handwritten data.

II. Introduction to AOA estimation

Consider a narrow-band plane wave of wavelength λ and frequency ν impinging on a linear sensor array at angle θ (fig 2). The spatial delay between two successive sensors is $d \sin\theta$ and the phase delay is:

$$\omega = \frac{2\pi d}{\lambda} \sin\theta \tag{1}$$

Taking the reference s at the first sensor, the complex envelope of the signal received at the l-th sensor is:

$$y_l = e^{-j\omega l} s + n_l \tag{2}$$

where n_l represents the noise. The signal vector received on an array of L sensors at time t is therefore:

$$\boldsymbol{z_t} = [y_0, ..., y_l, ..., y_{L-1}]^T = \boldsymbol{a}(\omega) s_t + \boldsymbol{n_t} \tag{3}$$

[1] IEEE Third Int. Conf. on Document Analysis and Recognition (ICDAR'95), August 14-16th, 1995, Montreal, Canada

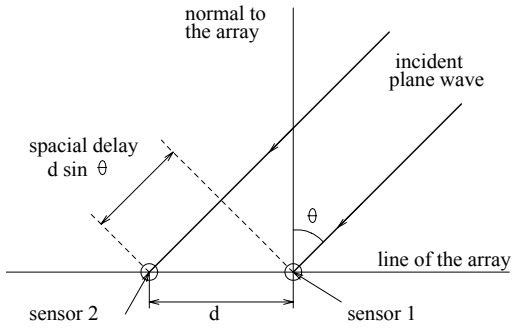

Figure 2: *Delay induced by a plane wave impinging upon a linear antenna*

where $\boldsymbol{a}(\omega) = [1, e^{-j\omega}, ..., e^{-j(L-1)\omega}]^T$ is the so-called "steering-vector", and $\boldsymbol{n_t}$ is the $L \times 1$ noise vector.
We take N snapshots, and build the snapshot matrix $M = [\boldsymbol{z_1}, ..., \boldsymbol{z_t}, ..., \boldsymbol{z_N}]$. High resolution methods like TLS-ESPRIT (Roy & Kailath, 1989) use singular value decompositions of the estimated correlation matrix of the signals $\hat{R} = \frac{1}{N} MM^H$ to estimate ω, and then θ (eq. 1).

III. Similarities between AOA and SAE

A. Aghajan, Khalaj and Kailath method

(Aghajan *et al.*, 1993a) shows how to link the AOA problem to the SAE problem. This method is dedicated to binary (preprocessed) text images. It was extended to other applications in (Aghajan & Kailath, 1993b), such as finding the orientation of lines on preprocessed aerial images. The method requires a pre-processing stage: a horizontal blurr to connect components, then a thresholding to binarize the pixels and finally a (positive) edge detection to eventually transform the image in a series of almost straight parallel thin lines. Then, if there are k non-zero pixels on the l-th column, at rows $q_1, ..., q_k$ respectively, the l-th sensor receives $y_l = \sum_{i=1}^{k} e^{-j\mu q_i}$, where μ is a parameter. Denoting x_i^0 the offset of line i in the first column, we get $q_i \simeq x_i^0 + l \tan\theta$ in the l-th column. Hence we have:

$$y_l = e^{-j\mu l \tan\theta}.s + n_l \tag{4}$$

where n_l stands for the noise and $s = \sum_{i=1}^{k} e^{-j\mu x_i^0}$. Now, the link with the AOA problem is easy to draw: it is clear that any method able to estimate ω from equation (2) is able to estimate $\mu \tan\theta$ in equation (4). Taking profit of this strong similarity, the AOA dedicated TLS-ESPRIT algorithm is used to estimate θ.

B. Proposed approach

As shown in (Aghajan *et al.*, 1993a), the method described above provides very good results on preprocessed typed text. Here, we propose a method which can deal directly with raw data (especially from handwritten documents), without any preprocessing step.

In order to match the usual array-processing hypotheses, consider that a linear array of L sensors is set at the top of the image, such that sensor l is dedicated to the l-th column; the "signal" impinging on this sensor is the content of its column. Since AOA methods are narrow-band methods, the data should be the complex envelopes of the signals received at each sensor, at a given frequency ν. In the sequel, we will note $Y_l(.)$ the signal on column l, such that $Y_l(k)$ is the grey level of pixel (k, l), and y_l the complex envelope of $Y_l(.)$ at frequency ν:

$$y_l = \frac{1}{\sqrt{K}} \sum_{k=0}^{K-1} Y_l(k) e^{-j2\pi\nu k} \tag{5}$$

Given the y_l, the problem is now to estimate the skew angle θ, as well as (if needed) the number of lines and their respective offsets. In classical array processing the frequency ν is fixed by the physical constraints (for example, it is the frequency of the radar). On the contrary, as far as text skew angle estimation is concerned, no constraint is set on ν. Aghajan & Kailath consider $\mu = 2\pi\nu$ instead of ν, and advise to take $\mu = 1$, but do not really base their choice on any particular theoretical ground.
If a different point of view is taken, and if basic Signal Processing Theory is considered, it seems a judicious idea to pick the frequency ν that contains the maximum energy of the signals. Let us call $\hat{y}_l(n)$ the Discrete Fourier Transform of column l: $\hat{y}_l(n) = \frac{1}{\sqrt{K}} \sum_{k=0}^{K-1} Y_l(k) e^{\frac{-j2\pi nk}{K}}$.
Symmetries of the DFT allow to consider only $1 \leq n \leq K/2$ (the continuous component $n = 0$ is not useful). The optimal frequency is $\nu = n_0/K$, where n_0 maximizes $\sum_{l=0}^{L-1} |\hat{y}_l(n_0)|^2$. Since n_0 is the number of cycles in a column, it is an estimation of the number of lines.
Equation (5) is used to compute the signals received at the frequency ν on each of the L sensors. As far as AOA is concerned, the number of sensors must be greater than the number of angles (which is 1 here), and many snapshots are needed. In the case of a single snapshot on a very long array (as is the case here), the so-called "spatial smoothing" technique is used: a virtual inter-sensor distance d is chosen, as well as a virtual number m of sensors, and the virtual snapshot matrix is built as follows:

$$M = [z_0, ..., z_{d-1}] = \begin{bmatrix} y_0 & y_1 & \cdots & y_{d-1} \\ y_d & y_{d+1} & \cdots & y_{2d-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{(m-1)d} & y_{(m-1)d+1} & \cdots & y_{md-1} \end{bmatrix} \tag{6}$$

The number of snapshots is d, and needs to be greater than m so that the $m \times m$ correlation matrix $\hat{R} = \frac{1}{d} MM^H$ is a good estimate of the true correlation matrix.
Let us now look for the optimal value for the virtual inter-sensor distance d: in a text with a skew angle θ we have

$Y_l(k) \simeq Y_0(k + l\tan\theta)$, which yields, for the complex envelope, to $y_l \simeq y_0 e^{j2\pi\nu l\tan\theta}$. If we change the indexing of the sensors such that $\tilde{y}_{l,i} = y_{ld+i}$, we can write $\tilde{y}_{l,i} \simeq \tilde{y}_{0,i} e^{j2\pi\nu ld\tan\theta}$ which is to be compared to AOA equation $y_l \simeq y_0 e^{-j\omega l}$. Hence we have:

$$\omega = -2\pi\nu d\tan\theta \tag{7}$$

If we assume that the skew angles to estimate all fit within $-63^o \leq \theta \leq 63^o$ (it is very unlikely to have a skew angle higher than that), then $-2 \leq \tan\theta \leq 2$, yielding to $-4\pi\nu d \leq \omega \leq 4\pi\nu d$. To avoid ambiguity, ω must stay in the interval $[-\pi, \pi]$, hence we get:

$$d = \frac{1}{4\nu} \tag{8}$$

Estimation of the offsets: Keeping the frequency related to the estimated number of lines n_0 only, and denoting $y_0 = \hat{y}_0(n_0)$, the inverse DFT provides:

. $\quad Y_0(k) \simeq \frac{2|y_0|}{\sqrt{K}} cos\left\{2\pi\nu\left(k + \frac{Arg(y_o)}{2\pi\nu}\right)\right\}$

Hence, for a text darker than the background, the offset of the first line is estimated by:

$$\hat{x}_0 = min_{(h)}\left\{x = \frac{-Arg(y_0) + \pi + 2h\pi}{2\pi\nu}; x > 0\right\} \tag{9}$$

For the $n_0 - 1$ following lines: $\hat{x}_p^o = \hat{x}_0 + \frac{p}{\nu}$

IV. AOA estimation using Neural Networks

Classical AOA methods (MUSIC, ESPRIT) are fast, but suboptimal. The Maximum Likelihood Estimator (MLE) is an optimal method, but it implies a large computational load, because it requires to examine every possible solution. A new method is exposed here, that takes profit of the cooperation of two neural networks, and that theoretically reaches the performances of the optimal MLE. Furthermore, this method is as fast as the classical suboptimal methods.

A Multi-Layer Perceptron (MLP) is a feed-forward neural network whose neurons are arranged in succeeding layers. The links between neurons are parameters called "weights". Training a network consists in tuning the weights in order to minimize the discrepancy between the desired output and the obtained output.

Rough estimation MLP

This MLP is a 3-layer perceptron with m^2 inputs, about m non-linear hidden neurons, and one output. It receives on input the correlation matrix (normalized by the mean of its diagonal elements), and it is trained to provide on output an estimate $\tilde{\omega}$ of ω. The training base is composed of simulated plane waves arriving at different angles θ on an array of $m = 10$ sensors. 181 values of ω are taken in the interval $[-\pi, \pi]$ and there are 5 examples for each ω, generated at a Signal to Noise Ratio of 10 dB.

Fine estimation MLP

The rough estimation $\tilde{\omega}$ is then refined by the MLP shown on figure 3. Input and output layers both contain m

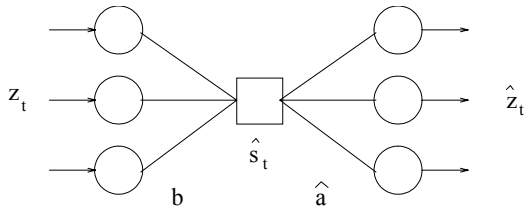

Figure 3: *The fine estimation MLP*

units, whereas the hidden layer is composed of only 1 unit. The weights and the neurons can take complex values and the equations of the network are: $\hat{\boldsymbol{s}}_t = \boldsymbol{b}^T \boldsymbol{z}_t$ and $\hat{\boldsymbol{z}}_t = \hat{\boldsymbol{a}}\hat{s}_t$. The output weight vector $\hat{\boldsymbol{a}}$ is *constrained* to represent the steering vector $\boldsymbol{a}(\hat{\omega})$ corresponding to parameter $\hat{\omega}$. $\boldsymbol{a}(\hat{\omega})$ is initialized with the "rough" estimation $\tilde{\omega}$ provided by the first MLP, and $\boldsymbol{b}_{init}$ is computed as the pseudo-inverse of $\boldsymbol{a}(\tilde{\omega})$.

The snapshots are presented in a random order to the input layer, and the output layer is forced to reproduce the input vector as closely as possible. (Burel & Rondel, 1993) provides the proof that minimizing $E\{\|\hat{\boldsymbol{z}}_t - \boldsymbol{z}_t\|^2\}$ is equivalent to optimizing a Maximum Likelihood Criterion. A gradient algorithm dedicated to constrained complex neural networks is used (Rondel & Burel, 1995). Since the network is initialized in the very neighbourhood of the solution, convergence is fast and avoids local minima. Another originality of this network is that the desired estimation is contained in the weights instead of being provided on the output layer.

V. A summary of the method

1. take a text image with K rows and L columns.
2. compute the DFT $\hat{y}_l(n)$ of the columns.
3. the number n_0 of lines in the text is the value of n which maximizes $\sum_{l=0}^{L-1} |\hat{y}_l(n)|^2$.
4. the optimal frequency (max. energy) is $\nu = n_0/K$.
5. compute the sensor outputs (eq. 5) and the optimal inter-sensor distance $d = 1/4\nu$, in order to build the snapshot matrix M as in (6), then the correlation matrix $\hat{R} = \frac{1}{d} MM^H$.
6. use the first MLP to obtain a rough estimation $\tilde{\omega}$ from the normalized matrix $\tilde{R}$.
7. initialize the second MLP with $\tilde{\omega}$ and allow 15 iterations (each iteration consists in the presentation of the d snapshots).
8. extract the "refined" estimation $\hat{\omega}$ out of the output weight vector $\boldsymbol{a}(\hat{\omega})$, and compute θ using eq. 7.
9. the offset of the first line $\hat{x}_0$ is given by equation (9).
10. the offsets $\hat{x}_p^o$ of the other lines equal $\hat{x}_0 + \frac{p}{\nu}$.

VI. Experimental results

Experiments conducted on a window extracted from the scanned image of a handwritten poem are shown. Obviously, using the full image can provide more accurate estimations, but we show that good results can be obtained even on a small window. The skew angles to estimate are respectively $\theta = 0^o, 15^o$, and 30^o (fig 4). It must be stressed that the skew angle can hardly be defined with a better accuracy than a few degrees, since the text is written without constraint. The obtained estimates are

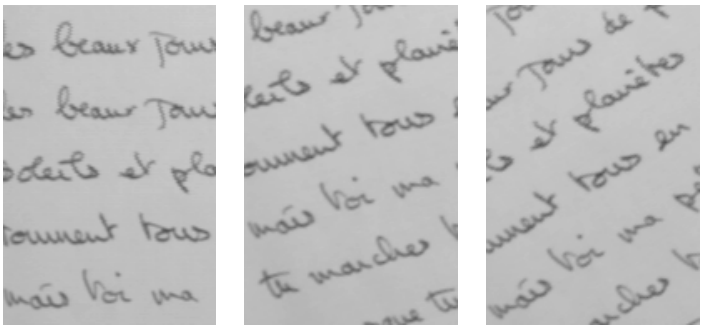

Figure 4: *Images at* $0^o, 15^o$ *and* 30^o *skew angles*

displayed in the chart below.

image	n_0	ν	d	θ (rough)	θ (fine)
0^o	5	0.026	10	0.9^o	0.4^o
15^o	5	0.026	10	8.7^o	15.0^o
30^o	4	0.021	12	35.6^o	32.9^o

For comparison, the estimations of θ obtained using (Aghajan *et al.*, 1993a) directly on raw data are respectively -0.3^o, -0.4^o, and 1.1^o. The fact that the estimated skew angles are all around 0^o may be partially explained by the fact that the parameter μ is fixed to 1: at such a high frequency, it is merely noise that is considered, and there is no particular skew on noise. However, this comparison is not fair, because, Aghajan's method requires a preprocessing step, while our approach doesn't need any.

Figure 5 shows the estimated lines on the 15^o image. As one can see, the performances of the method are quite good, even in the case of handwritten unpreprocessed text. It must be stressed that handwritten text is always more difficult than typed text, because it less regular: for example, handwritten lines are not regularly spaced, the lines are not exactly straight, and the skew angle may vary from one line to another.

VII. Conclusion

An original method for estimating the Skew Angle of text document images has been presented . Based on the co-

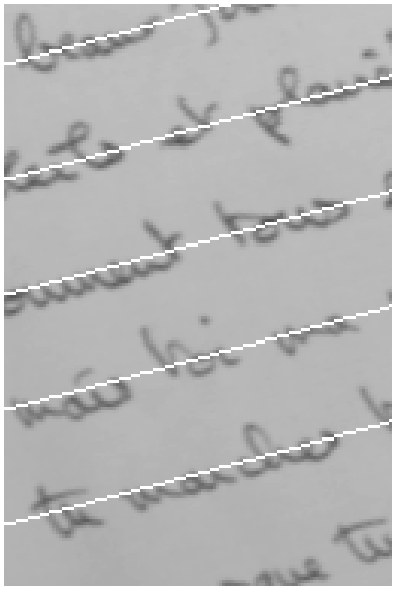

Figure 5: 15^o *image with the representation of its estimated skew angle and line offsets.*

operation of two neural networks, the method takes advantage of the similarity of the given problem with array-processing Angle of Arrival formalism. Experimental results show the accuracy of the method in the case of unpreprocessed handwritten text lines. An improvement of the method is the estimation of a number of lines which is not an integer (not presented here due to lack of space).

REFERENCES

Aghajan, H.K., Khalaj, B.H. & Kailath, T. (1993a). Estimation of Skew Angle in Text Image Analysis by Sensor Array Processing Techniques. *SPIE Vol. 1906. Character Recognition Technologies*, pp. 49-60, 1993.

Aghajan, H.K. & Kailath, T. (1993b). Sensor Array Processing Techniques for Super Resolution Multi-Line-Fitting and Straight Edge Detection. *IEEE Transactions on Image Processing*, No 4, (Vol. II pp. 454-465), Oct. 1993.

Burel, G. & Rondel, N. (1993). Neural Networks for Array Processing: from DOA Estimation to Blind Separation of Sources. *IEEE/SMC Conference, Invited Paper*, Le Touquet, France, 17-20 Oct 1993.

Hinds, S.T., Fisher, J.L. & d'Amato, D.P. (1990). A Document Skew Detection Method using Run-Length Encoding and the Hough Transform. *Int. Conf. on Pattern Recognition* (Vol. I pp. 464-468). Atlantic City, NJ, 1990.

Rondel, N. & Burel, G. (1995). Cooperation of multi-layer perceptrons for angles of arrival estimation. *IEE 4th Conf. on Artificial Neural Nets*, Cambridge, UK, June 26-28th, 1995.

Roy, R. & Kailath, T. (1989). ESPRIT:Estimation of Signal Parameters via Rotational Invariance Techniques. *IEEE Trans on ASSP*, Vol. 7, No 37, pp. 984-995, July 1989.