REPULSIVE ATTRACTIVE NETWORK FOR BASELINE EXTRACTION ON DOCUMENT IMAGES

Erhan Öztop *Adem Y. Mülayim* *Volkan Atalay* *Fatoş Yarman-Vural*

Department of Computer Engineering,
Middle East Technical University, Ankara, Türkiye
vural@ceng.metu.edu.tr

ABSTRACT

This paper describes a new framework, called, *Repulsive Attractive (RA) Network for Baseline Extraction* on document images. The RA network is a self organizing feature detector which interacts with the document text image through the attractive and repulsive forces defined among the network components and the document image. Experimental results indicate that the network can successfully extract the baselines under heavy noise and with overlaps between the ascending and descending portions of the characters of adjacent lines. The proposed method is also applicable to a wide range of image processing applications, such as curve fitting, segmentation and thinning.

1. INTRODUCTION

It is well known that a crucial step in document image analysis is the identification of the baselines. On a document image, baselines not only give important information about the layout structure, but also provide effective clues to subsequent steps of document analysis, such as optical character recognition. Baseline extraction problem becomes complicated in handwritten documents where the ascending and descending portions of the characters between adjacent lines overlap. Additional complexity is introduced when the characters overlay on a curved or skewed baseline and the documents are contaminated with noise.

In this study, a new method is presented for extracting the baselines. The method is inspired from the universal law of gravitation where the masses attract each other according to their weight and distance among them. The pixels of the image and the baselines are considered as if they were masses; each pixel attracts the baselines proportional to its gray value and inversely proportional to the square of the distance between them. The baselines, themselves, repel each other with a magnitude inversely proportional to the square of the distance between them. This idea is implemented as a self organizing feature detector [1] called Repulsive Attractive (RA) network. The new method is applicable to a wide range of documents and overcomes many problems faced during the baseline extraction process, such as thresholding, noise sensitivity, intolerance to font-style variations and skew angles.

*This work is supported by TÜBİTAK-BİLTEN Information Technologies and Electronics Research Institute.

In the next section available techniques for baseline extraction problem are reviewed. In Section 3, baseline extraction using RA network is described. The network is tested on various text images in Section 4. Finally, Section 5 concludes the paper.

2. BASELINE EXTRACTION

In general, a baseline is defined to be a curve or line on which the characters overlay on a document image. Standard baseline extraction methods include Hough transform [2], least squares methods [3], horizontal projection profiles [4], run-length smearing and use of typographical information [5]. Among all, the Hough transform and its variants are most widely used [2], where the document image is transformed to the (ρ, θ) plane (ρ is the magnitude and θ is the angle for a pixel). For a document image having curved baseline and skew angle Hough transform works successfully. However, the quantization of the input image by extracting the geometric features, such as center of mass of each character introduces an uncertain amount of error into the result yielding unsatisfactory solutions for the characters with ascending and descending portions. The complexity of the search in the (ρ, θ) plane is another drawback. The search may even end up with a locally optimum solution [6]. There is a vast amount of variation of the least square methods for fitting lines or curves to a given set of data points which can be applied to baseline extraction problem [7]. The major limitation of these methods is the sensitivity to the noise contamination. The most popular baseline extraction method for printed text is the use of horizontal projection profile which simply obtains the histogram of the image on the y-axis and identifies the baseline as the peak points of the histogram [8]. Obviously, it is highly sensitive to skew angles which requires a strong preprocessing stage for normalization. A common baseline extraction method for binary images is to use run length smearing. It is based on the run length codes which consists of a start address of each string of 1's, followed by the length of that string. This method is, also, highly susceptible to the noise contamination and skew angle. Finally, use of typographical information for baseline extraction heavily depends on the skew angle, font style and size. It is developed for machine printed texts [5].

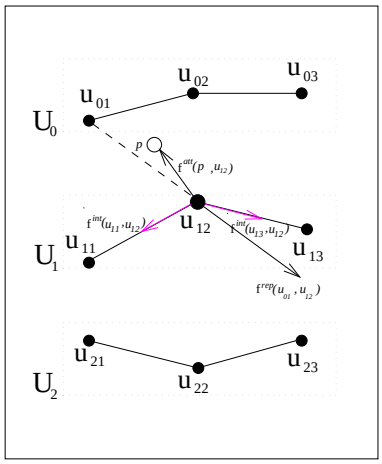

Figure 1. Forces acting on the subunit u_{12} (only a sample from each type of force is shown).

3. REPULSIVE ATTRACTIVE (RA) NETWORK FOR BASELINE EXTRACTION

A Repulsive Attractive Network is identified by the triple $(\mathcal{Y}, g, \mathcal{U})$ where $\mathcal{Y}$ is a vector space, g is a real-valued function defined on $\mathcal{Y}$ and $\mathcal{U}$ is the set of units. A *unit* $U_i \in \mathcal{U}$ is composed of *subunits*, u_{ij}. Each subunit is associated with a position or a weight vector in $\mathcal{Y}$.

The dynamics of the Repulsive Attractive Network is determined by the following forces.

- The internal force, $\mathbf{f}^{int}$ that exists among the subunits belonging to the same unit. This force gives units the tendency to have certain shape or orientation.
- The repulsive force, $\mathbf{f}^{rep}$ that exists among the subunits of different units.
- The attractive force, $\mathbf{f}^{att}$ that is exerted by the points of $\mathcal{Y}$ with a magnitude proportional to the value of g at these points.

The Repulsive Attractive Network is associated with a document image in the following manner. The baselines, being curves, are approximated as connected line segments. The triple $(\mathcal{Y}, g, \mathcal{U})$ and the internal force for the RA Network for baseline extraction is specified as follows (See Figure 1).

- $\mathcal{Y}$ denotes the document image embedded into $\Re^2$. More precisely, $\mathcal{Y} = \{(x, y) \in \Re^2 \mid \exists$ two points $(x_0, y_0), (x_1, y_1)$ on the image such that $x_0 \leq x \leq x_1$ and $y_0 \leq y \leq y_1\}$
- $\mathcal{U} = \{U_0, U_1, \ldots, U_n\}$ denotes the set of curves that approximates the baseline where n is the number of baselines, and the subunits u_{ij} are the connection points on the curves.
- $g : \mathcal{Y} \rightarrow \mathcal{Z}$, where $g(p)$ is the intensity value of the image at pixel p if $p \in \mathcal{Z}^2$, 0 otherwise.
- The internal force of the RA network enforces the subunits of the same baseline to lie on a horizontal line.

The forces acting on the document image is defined at three levels.

At level 1, the internal, repulsive and attractive forces among the subunits and pixels are defined as follows. For each subunit $u_{ij} \in U_i$, the received internal force between the subunits of the same unit is

$$\mathbf{f}^{int}(u_{il}, u_{ij}) = \mathbf{u}(u_{il}, u_{ij})/\delta^2(u_{il}, u_{ij}), \forall u_{il} \in U_i, j \neq l \quad (1)$$

the repulsive force between the subunits of distinct baselines is

$$\mathbf{f}^{rep}(u_{lk}, u_{ij}) = \mathbf{u}(u_{ij}, u_{lk})/\delta^2(u_{lk}, u_{ij}), \forall u_{lk} \in U_l, i \neq l \quad (2)$$

and the attractive force between the image pixels and subunit u_{ij} is

$$\mathbf{f}^{att}(p, u_{ij}) = \mathbf{u}(p, u_{ij})g(p)/(\delta^2(p, u_{ij}) + \eta), \forall p \in \mathcal{Y} \quad (3)$$

In the above equations δ is the Euclidian distance function, $\mathbf{u}(P_1, P_2)$ is the unit vector directed from P_1 towards P_2 and η is a real constant.

At level 2, the internal, repulsive and attractive forces are aggregated to generate net forces:

$$\mathbf{F}_{ij}^{NET-int} = \sum_{v \in \mathcal{I}_{ij}} \mathbf{f}^{int}(v, u_{ij}) \quad (4)$$

$$\mathbf{F}_{ij}^{NET-rep} = \sum_{v \in \mathcal{R}_{ij}} \mathbf{f}^{rep}(v, u_{ij}) \quad (5)$$

$$\mathbf{F}_{ij}^{NET-att} = \sum_{p \in \mathcal{Y}} \mathbf{f}^{att}(p, u_{ij}) \quad (6)$$

where

$$\mathcal{I}_{ij} = \{u_{ik} \mid k \neq j\} \text{ and } \mathcal{R}_{ij} = \{u_{lk} \mid i \neq l\} .$$

At level 3, the net forces are further aggregated to generate the total net force:

$$\mathbf{F}_{ij}^{NET} = \alpha \mathbf{F}_{ij}^{NET-int} + \beta \mathbf{F}_{ij}^{NET-rep} + \gamma \mathbf{F}_{ij}^{NET-att} \quad (7)$$

where α, β, γ are real coefficients.

The total net force is used to update the position or weight vector associated with each subunit. In this particular application, only the vertical components are taken into account. The vertical component is simply multiplied by a constant scale value to find out the amount of change to be done on the vertical position of each subunit. In order to simulate the system described above the following algorithm is used.

ALGORITHM:

0. Choose network parameters: α, β, γ.
Initialize $\mathcal{U} = \{u_0\}$, u_0 *is at the top of the image.*
Initialize ρ, ϵ *and* M.

1. Let $\delta = 0$.

2. *For k=1,2,...M do*
   - 2.1 *Choose a subunit u_{ij}, randomly,*
   - 2.2 *Compute $\mathbf{F}_{ij}^{NET}$,*
   - 2.3 *Let F_y be the vertical component of $\mathbf{F}_{ij}^{NET}$,*
   - 2.4 *Let $\Delta y = \rho F_y$,*
   - 2.5 *Update the vertical position of subunit u_{ij} by adding Δy,*
   - 2.6 *Let $\delta = \delta + \Delta y^2/M$.*
3. *If $\delta > \epsilon$ goto step 1, else continue.*
4. *If all of the baselines are extracted then stop.*
   else add a new unit to $\mathcal{U}$, located just below the last added unit.
5. *Goto step 1.*

In the above algorithm, the parameter ρ controls the step size of the position updates. When ρ gets small, the simulation yields finer results in terms of the shape of the baseline. However, it causes slow convergence. ρF_y should not exceed 1. ϵ is a constant to control convergence. It should be close to zero. M defines the number of random subunit selections per update session. It should be set to a value that is greater than the total number of subunits in the network.

4. EXPERIMENTAL RESULTS

First the behavior of the forces in the network and the effect of the parameters α, β, γ on the shape of the baseline are discussed. The internal force enforces the subunits of the same unit to lie on a horizontal line. Therefore, when the coefficient of the internal force α, is large relative to β and γ, the RA network looses its ability to detect curved baselines. On the other hand, if α is taken to be relatively small, then the local properties of the document image dominates the effect of neighbouring subunits, causing zigzags on the baseline. If the coefficient of repulsive force β, is relatively large, then the influence of text pixels on the baseline decreases. On the other hand, as β gets smaller, multiple units tend to capture the same baseline. In fact this parameter is proportional to the font size. The amount of attraction that the text pixels exert on the subunits is controlled by the parameter for the attractive force, γ. If γ is taken to be smaller compared to α and β, then the input to the network is received as an empty page image. If γ is large then the attractive force overrides the repulsive force. This may cause accumulation of more than one unit on the same line.

In order to test the performance of RA network as a baseline extractor, some simulation experiments have been done. The gray level images are selected among the documents from the ancient Ottoman archives and from the Latin handwritten documents. The heavy noise on the documents are due to the aging, ink smearage and low quality of the paper. Furthermore, the handwritten documents have high curvature on the baselines with considerable overlaps on the ascending and descending portions of the characters.

One of the major superiority of the RA network is that it does not require any preprocessing stage to remove the noise or to binarize the documents. Therefore, the sample documents are directly fed to the RA network, for baseline extraction.

Figure 2 shows a handwritten Ottoman text. Note that the text lines are quite close to each other. This situation is higly difficult to handle with standard baseline extraction methods, mentioned in section 2. As it can be seen from the figure the extracted baselines are very close to the expected ones. During the experiment, it is noted that the RA network is not sensitive to the changes in parameters. For example, equally acceptable results can be obtained using the triples, each representing (α, β, γ): (5000,550,50), (5000,550,35), (2500,550,50), (5000,800,50) with 15 subunits per unit.

Figure 3 shows a handwritten Latin text with heavy noise and fair amount of overlap between the characters of adjacent lines. The simulated RA network, however is quite insensitive to such disturbances and captures the curvatures of the baselines. The parameters are selected as $\alpha = 5000$, $\beta = 550$, $\gamma = 40$ and 20 subunits are employed per unit

5. CONCLUSION

In this study, the RA network for baseline extraction on document images is described. This network may be incorporated into a complete document processing system as a part of a page layout extractor or it may be used to provide information to a character recognizer.

There are many advantages of RA network baseline extraction method: RA network completely eliminates the thresholding problem. Moreover, employing gray-level images improves the performance in contrast to the other methods. One of the most serious problems in many image processing tasks is noise sensitivity. All of the existing methods for baseline extraction suffers from this problem. However, since the attraction of the text pixels dominate the attraction of the noise pixels on the baseline, the proposed method is insensitive to noise. Not only a given set of parameters works fine for a wide range of documents, but also small variations in these parameters does not affect performance of the method. The text lines are not restricted to straight lines which is a major advantage for handwritten documents. Additional difficulty in identification of baselines is generally introduced in handwritten documents by the overlapping portions of ascending and descending characters between the text lines. The RA network baseline extraction method handles elegantly the problem.

Finally, the principles used in the RA network baseline extraction method can be used to develop new techniques for various image processing tasks, such as segmentation and thinning.

REFERENCES

[1] Kohonen, T., *Self Organization and Associative Memory*, Berlin, Springer-Verlag, 1989.

[2] Leavers V.F., "Which Hough Transform?", *CVGIP: Image Understanding*, vol.58, no. 2, 1993, pp. 250-264.

[3] Aghajan, H. K., Kailath, T., "SLIDE:Subspace-Based Line Detection", *T-PAMI*, 1994, vol. 16, no. 11, pp. 1057-1073.

[4] Jain, A. K., *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice Hall, 1989.

[5] Kanai, J. "Text-line Extraction Using Character Prototypes", *Workshop on Syntactic and Structural Pattern Recognition*, New Jersey 1990, pp. 182-191.

[6] Atalay V., Özçilingir M., Yalabık N., "Computer Recognition of Ottoman Text" *Proc. ISCIS V*, Capadoccia, Türkiye, 1990.

[7] Haralick, R.M., Shapiro, L.G., *Computer and Robot Vision*, vol.1, pp.602-627, Addison-Wesley, 1992.

[8] Yarman-Vural, F. and Atıcı, A. "A Segmentation and Feature Extraction Algorithm for Ottoman Cursive Script", *Proceedings on Turkish Artificial Intelligence and Neural Network Conference, (TAINN)*, Ankara, Türkiye, 1994.

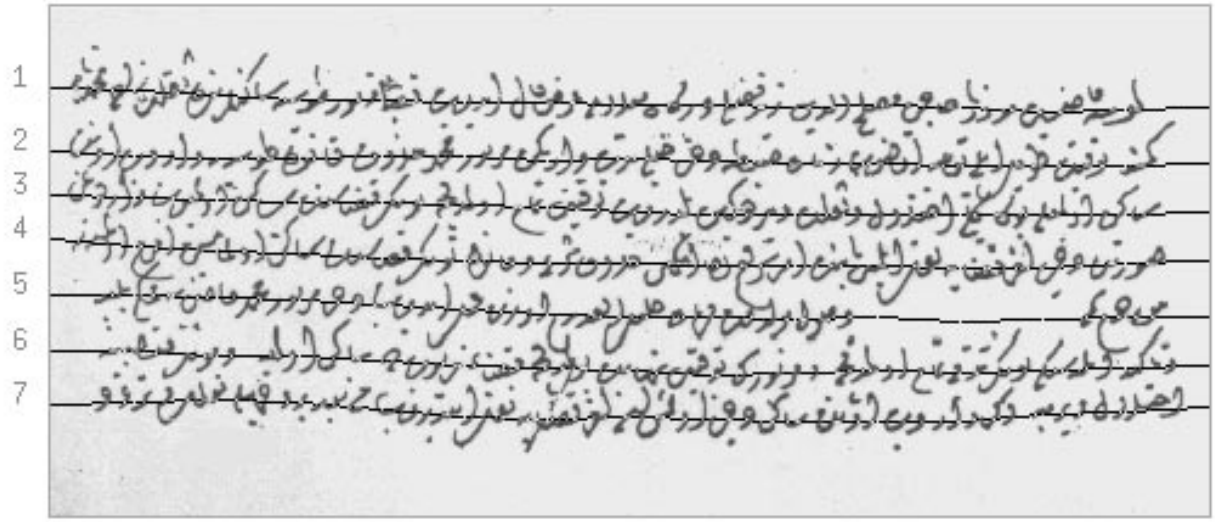

Figure 2. Extraction of baselines for a handwritten Ottoman text.

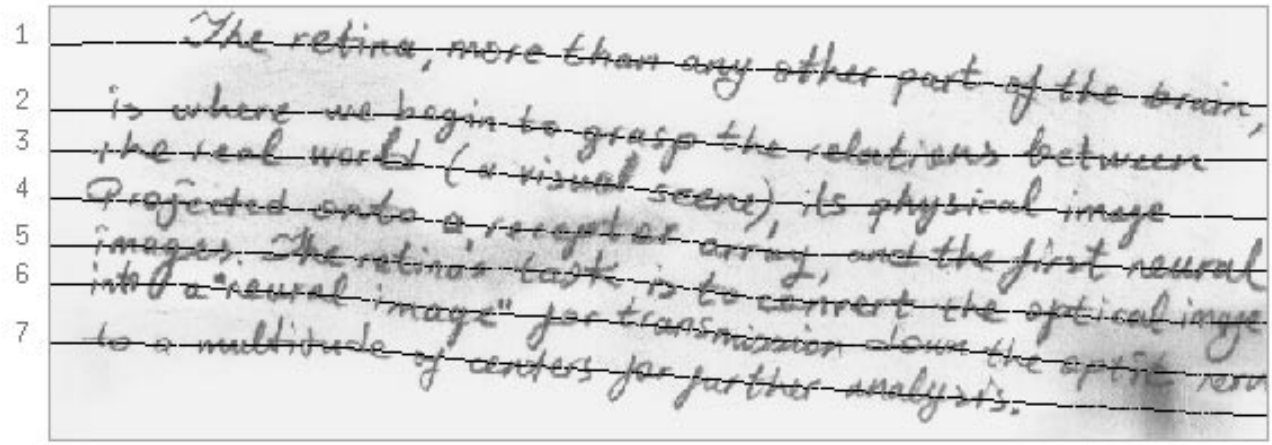

Figure 3. Extraction of baselines for a handwritten latin text with noise.