ELSEVIER

Signal Processing 75 (1999) 1–10

SIGNAL PROCESSING

Repulsive attractive network for baseline extraction on document images

E. Öztop, A.Y. Mülayim, V. Atalay, F. Yarman-Vural*

Department of Computer Engineering, Middle East Technical University, TR-06531 Ankara, Turkey

Received 26 July 1996; received in revised form 9 November 1998

Abstract

This paper describes a new framework, called *repulsive attractive* (RA) *network for baseline extraction* on document images. The RA network is an energy minimizing dynamical system, which interacts with the document text image through the attractive and repulsive forces defined over the network components and the document image. Experimental results indicate that the network can successfully extract the baselines under heavy noise and overlaps between the ascending and descending portions of the characters of adjacent lines. The proposed framework is applicable to a wide range of image processing applications, such as curve fitting, segmentation and thinning. © 1999 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Diese Arbeit beschreibt ein neues Rahmenprogramm, genannt '*Repulsive Attractive* (RA) *network for baseline extraction*' bei Dokument Abbildungen. Das RA Netzwerk ist ein energieminimierendes dynamisches System. Dieses wird mit der Abbildung des Dokumenttextes durch die anziehenden und abstoßende Kräfte, definiert über Netzwerk-Komponenten, und dem Dokumentbild in Beziehung gesetzt. Experimentelle Resultate zeigen, daß das Netzwerk die Grundlinie zwischen aufsteigenden und absteigenden Buchstabenbereichen unter schwierigen Rauschbedingungen und bei Überlappung erfolgreich extrahieren kann. Das vorgestellte Programm ist anwendbar auf weite Bereiche der Bildverarbeitung, wie Kurvenanpassung, Segmentierung und Ausdünnung. © 1999 Elsevier Science B.V. All rights reserved.

Résumé

Cet article décrit un nouveau cadre de travail, appelé *réseau répulsif attractif pour l'extraction de la ligne de base* sur des images de documents. Let réseau répulsif attractif est un système dynamique minimisant une énergie, qui interagit avec l'image du texte du document par des forces attractives et répulsives définies sur les composantes du réseau et sur l'image du document. Des résultats expérimentaux indiquent que le réseau peut extraire avec succès les lignes de base sous un bruit important et en présence de recouvrements entre les parties ascendantes et descendantes des caractères de deux lignes adjacentes. Le schéma proposé est applicable à une large gamme d'applications de traitement d'images, telles l'ajustement de courbes, la segmentation et l'affinement. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Baseline; Repulsive attractive network; Hough transform; Energy minimization; Active model

* Corresponding author. E-mail: vural@ceng. metu.edu.tr

PII: S0165-1684(98)00220-5

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 slanted baseline and the documents are contaminated with noise.

In this study, a new method is presented for extracting the baselines. The method is an application of a framework called repulsive attractive (RA) network [11]. RA Network can be described as an energy minimizing dynamical system [9] and it is inspired from the universal law of gravitation where the masses attract each other according to their weight and distance among them. In this particular application in which the aim is to extract the baselines of a document image, 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. 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. Some other possible applications of the RA network in image processing can be thinning and segmentation.

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

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 [5], least-squares methods [1], horizontal projection profiles [7], run-length smearing and use of typographical information [8].

Among all, the Hough transform and its variants are the most widely used methods [10]. In these methods, the document image is first transformed to the (ρ,θ) plane. ρ indicates the length and θ indicates the slope of the normal line to a pixel in polar coordinates. Then, a two-dimensional search is accomplished for detecting the peaks in the transform plane. For a document image having curved baseline and skew angle Hough transform works successfully. However, it has certain drawbacks that may limit its use. One problem is the quantization of the input image by extracting the geometric features, such as center of mass of each character. This process introduces an uncertain amount of error into the result yielding unsatisfactory solutions for the characters with ascending and descending portions. Secondly, the complexity of the search for peaks in the (ρ,θ) plane is a serious drawback. In addition, it can be difficult to put a criterion for identifying the peaks [3,4].

There is a vast amount of variation of the least-squares methods for fitting lines or curves to a given set of data points [6]. 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 [12]. 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 [2].

Finally, use of typographical information for baseline extraction heavily depends on the skew angle, font style and size. It is particularly developed for machine printed texts [8].

3. The repulsive attractive (RA) network for baseline extraction

In this section, the repulsive attractive (RA) network for baseline extraction is described.

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

The dynamics of the repulsive attractive network is determined by the following forces:

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

The repulsive attractive network can be associated with the document text image. The baselines, being curves, are approximated as connected line segments. The triple $(\mathscr{Y},g,\mathscr{U})$ of the RA network for baseline extraction is specified as follows (see Fig. 1):

- $\mathscr{Y}$ denotes the document image embedded into $\Re^2$. More precisely, $\mathscr{Y} = \{(x,y) \in \Re^2 \mid \exists$ two points (x_0,y_0), (x_1,y_1) on the image such that $x_0 \leqslant x \leqslant x_1$ and $y_0 \leqslant y \leqslant y_1\}$
- $\mathscr{U} = \{U_0,U_1,\ldots,U_n\}$ denotes the set of curves that approximates the baseline where n is the number of baselines, and the sub-units u_{ij} are the connection points on the curves.
- $g : \mathscr{Y} \to \mathscr{Z}$, where $g(p)$ is the intensity value of the image at pixel p if $p \in \mathscr{Z}^2$, 0 otherwise.

The internal force of the RA network enforces the sub-units of the same baseline to lie on a horizontal line. The forces acting on the document image is defined at three levels:

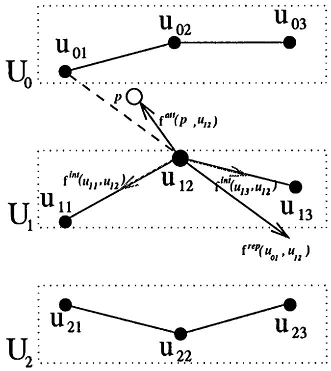


Fig. 1. The forces acting on the sub-unit u_{12} (only a sample from each type of force is shown).

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

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

the repulsive force between the sub-units of distinct baselines is

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

and the attractive force between the image pixels and sub-unit u_{ij} is

$$\boldsymbol{f}^{\text{att}}(p,u_{ij}) = \boldsymbol{u}(p,u_{ij})g(p)/(\delta^2(p,u_{ij}) + \eta), \quad \forall p \in \mathscr{Y}. \tag{3}$$

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

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

$$\boldsymbol{F}_{ij}^{\text{NET}-\text{int}} = \sum_{v \in \mathscr{I}_{ij}} \boldsymbol{f}^{\text{int}}(v,u_{ij}), \tag{4}$$

$$\boldsymbol{F}_{ij}^{\text{NET}-\text{rep}} = \sum_{v \in \mathscr{R}_{ij}} \boldsymbol{f}^{\text{rep}}(v,u_{ij}), \tag{5}$$

$$\boldsymbol{F}_{ij}^{\text{NET}-\text{att}} = \sum_{p \in \mathscr{Y}} \boldsymbol{f}^{\text{att}}(p,u_{ij}), \tag{6}$$

where

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

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

$$\boldsymbol{F}_{ij}^{\mathrm{NET}} = \alpha \boldsymbol{F}_{ij}^{\mathrm{NET-int}} + \beta \boldsymbol{F}_{ij}^{\mathrm{NET-rep}} + \gamma \boldsymbol{F}_{ij}^{\mathrm{NET-att}}, \qquad (7)$$

where α, β, γ are real coefficients.

The RA network framework defines the internal, repulsive and attractive forces among the sub-units and the pixels in order to extract the baselines on a document. A unit evolves dynamically via the movements of its sub-units. The displacement of the sub-units is determined by the forces acting on it and the total net force is used to update the position vector associated with each sub-unit. 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 performed on the vertical position of each sub-unit. The energy function is defined as the integral of the forces. At a local minimum of the energy function, the sub-units do not change their positions. Each baseline is assumed to be a local minimum of the energy function. If a unit is sufficiently close to a baseline, then it will be attracted by the local minimum and this situation is called as the local convergence. The global convergence yields the extraction of all the baselines on a document. In order to guarantee the global convergence, a virtual unit is allocated on top of the document. Also, bottom of the document is considered to behave as a virtual textline. As soon as a unit is stuck to the bottom of the image, the global convergence is achieved. This indicates that all of the existing baselines have been extracted.

4. Algorithm for baseline extraction

In order to simulate the system described in the preceeding section, the following algorithm is used:

```
1. Choose network parameters: $\alpha$, $\beta$, $\gamma$.
   Initialize $\rho$, $\varepsilon$ and $M$.
   Initialize $\mathscr{U} = \{u_0\}$, $u_0$ located at the top edge of the image.
2. Let $\delta = 0$
3. For $k = 1,2,\ldots,M$ do
   3.1. For each sub-unit $u_{ij}$ of the newly added unit
        3.1.1. Compute $\boldsymbol{F}_{ij}^{\mathrm{NET}}$
        3.1.2. Let $F_y$ be the vertical component of $F_{ij}^{\mathrm{NET}}$
        3.1.3. Let $\Delta y = \rho F_y$
        3.1.4. Update the vertical position of sub-unit $u_{ij}$ by adding $\Delta y$
        3.1.5. Let $\delta = \delta + \Delta y^2/M$
4. If $\delta > \varepsilon$ goto step 2, else continue
5. If the coordinates of the last added unit collapse with the bottom of the document image then stop.
   Else add a new unit to $\mathscr{U}$, located just below the last added unit.
6. Goto step 2.
```

At the initial step, parameter values for internal, repulsive and attractive forces are chosen and the parameters that control the convergence are initialized.

The behavior of the forces in the network are determined by the values of the parameters α, β, γ. The internal force, enforces the sub-units 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 neighboring sub-units, 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. Repulsive force guarantees that the distance between the sub-units of different units on the same column does not come close to 0. 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 sub-units 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.

ρ is a constant for controlling 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 converge. ε is a constant to control the local convergence. It should be close to 0. The very first unit is introduced just below the top of the document image. At the second step, coordinates of the sub-units of the newly introduced unit are updated M times. This is repeated until the local convergence is reached. The local convergence criterion checks whether there is any change in the positions of the sub-units. This is evaluated by computing the average square displacements after a set of M updates. The algorithm terminates when the coordinates of the lastly introduced unit collapse with the bottom of the document as a result of a local convergence. Otherwise, a new unit is added just below the last one.

5. Results

The performance of the RA network as a baseline extractor is tested on various document images by a simulation program using the C language, under X Windows environment.

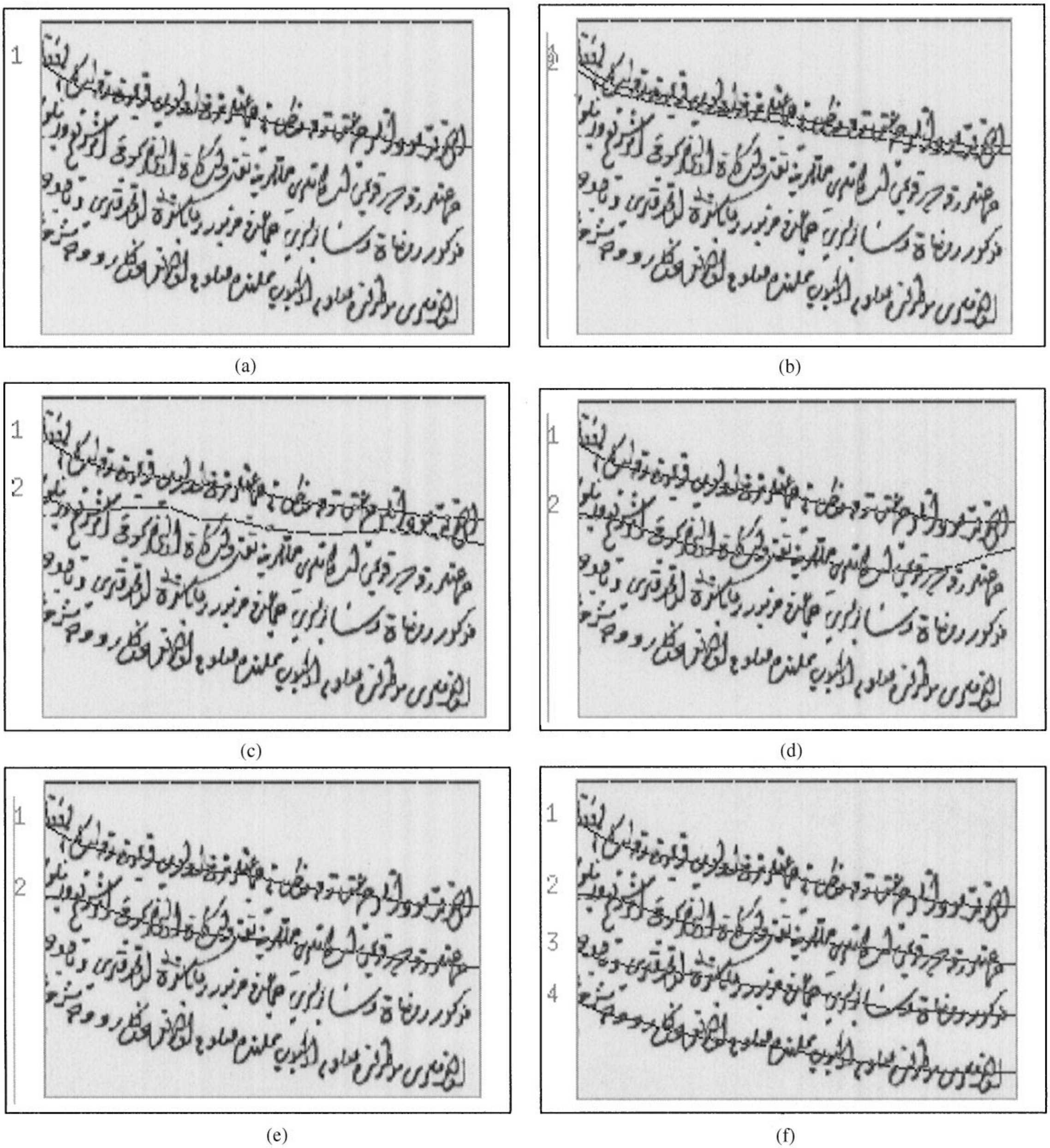
The gray-level images are selected among the documents from the ancient Ottoman archives (written in Arabic) and from the Latin handwritten and printed documents. The heavy noise on the documents are due to the aging, ink smearage and low quality of the paper. Furthermore, the handwritten documents have 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 for noise reduction and binarization. The sample documents are directly fed to the RA network, for baseline extraction. In most of the documents, the noise pixels have relatively light gray values compared to text pixels and the number of the text pixels exceed the number of the noise pixels. In this case, the attractive force of the text pixels dominates the attractive force of the noise pixels. As a result the RA network becomes insensitive to noise.

In the RA network for baseline extraction, the sub-units are evenly placed on a unit. The choice of the number of sub-units is not a critical parameter. A complex baseline, which does not have a constant curvature, needs relatively more sub-units than a simple baseline with constant curvature. For example, a linear baseline can be extracted simply by two sub-units. The number of sub-units may initially be chosen large. Then, by checking curvatures locally on the constructed unit, a kind of pruning algorithm on the sub-units can be applied until any change is detected on the curvatures. The network is composed of units and sub-units. There exists interactions among all of the sub-units, and also among all of the sub-units and the pixels of the image. Therefore, whenever a new unit is introduced, its sub-units affect the sub-units of already existing units and vice versa. However, since the existing units are stuck to the local minima, they do not exhibit important changes. Thus, the update is performed only for the sub-units of a newly introduced unit. The first unit appears just below the top of the document. It will be repelled by the top and attracted by the first textline. As soon as the local convergence is obtained for the first unit, second unit is initialized just below the first one. At this moment, repulsive force from the first unit pushes the second one, so that the second unit is captured by the attractive force of the second textline. More specifically, the displacement is determined by the net force which is composed of internal, attractive and repulsive forces. Internal forces and attractive forces are almost the same for the units discussed above. However, since these two units repel each other, the first unit goes up and the second unit goes down. The first unit cannot go further, because it is also repelled by the top of the document.

In the implementation of the algorithm, M is taken as 10 and ε is taken as 0.25 which is decided experimentally. During the simulations, the value of η was fixed to 10, which is relatively large compared to other parameters in Eq. (3). If η is chosen relatively small, then vertical displacements may become large (particularly when the distance between a sub-unit and a pixel is close to 0) which yields jumps rather than smooth displacements. Thus, the unit does not get stuck to a local minimum. The position at which a new unit is added 'just below' the previous one depends on the baseline spacing and the resolution of the document image. However, in the implementation, it is fixed to a value of 20 pixels which works fine for a moderate range of documents. The parameters α, β, γ and the number of sub-units will be discussed below in detail.

The evolution of the RA network for a handwritten Ottoman text is presented in Fig. 2. Taking the parameters $\alpha = 600$, $\beta = 550$, $\gamma = 50$ along with 15 sub-units on each unit enables the network to capture the high curvature in the baselines. Fig. 2(a) shows the state of the network, when the local convergence is achieved for the first unit. The second unit is introduced just below the first one as



(a) (b) (c) (d) (e) (f)

Fig. 2. The evolution of the RA network for a handwritten Ottoman text.

indicated in Fig. 2(b). Fig. 2(c) and (d) present some intermediate results for the second unit. The baseline extracted by the second unit is shown in Fig. 2(e). Finally, Fig. 2(f) presents the state when the global convergence is achieved.

In order to show the experimental convergence of the system, evolution of the position updates for an Ottoman text (cf. Fig. 2) is given as a plot in Fig. 3. The y-axis is the average square change in the displacement of the sub-units of a unit while the x-axis represents the set of iterations ($M \times$ number of subunits $= 10 \times 15 = 150$). After the first set of iterations, the average square change drops drastically, which indicates in fact, the local convergence. In the algorithm, this fact is tested by the comparison to ε which is fixed to 0.25. Local convergence is achieved in about 10 ms on a PowerPC processor-based computer.

One may expect that the values of the parameters, particularly those of α, β and γ, depend on the complexity of the document which may be affected by the alphabet in which the document is written. The complexity may include: noise level, curvature, change of curvature, regularity of writing, overlaps on the ascending and descending portions of the characters, etc. However, during the experimentation it is noted that the RA network is not sensitive to the changes in the parameters. A wide range of values of the network parameters is capable of baseline extraction. For example, the triple ($\alpha = 5000$, $\beta = 550$, $\gamma = 50$) with 15 sub-units successfully extracs baselines of all of the presented documents.

Fig. 4 shows a handwritten Ottoman text. Note that the baselines are quite close to each other. There is a substantial amount of overlap between the adjacent baselines because of ascending and descending portions of the characters which is quite common in handwritten documents. This situation is highly difficult to handle with standard baseline extraction methods.

Fig. 5 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. Fig. 6 shows a printed Latin text. The RA network easily finds the baselines with the above indicated parameter values.

Finally, the result of one of the most popular baseline extraction method, namely the horizontal projection profiles is demonstrated for comparison

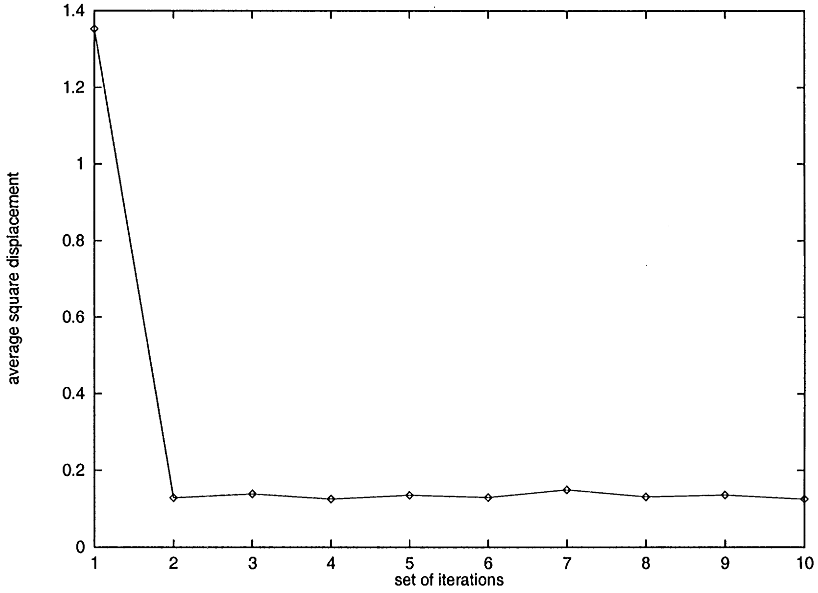


Fig. 3. Average square change in the displacement (change in vertical position) of the sub-units of a unit.

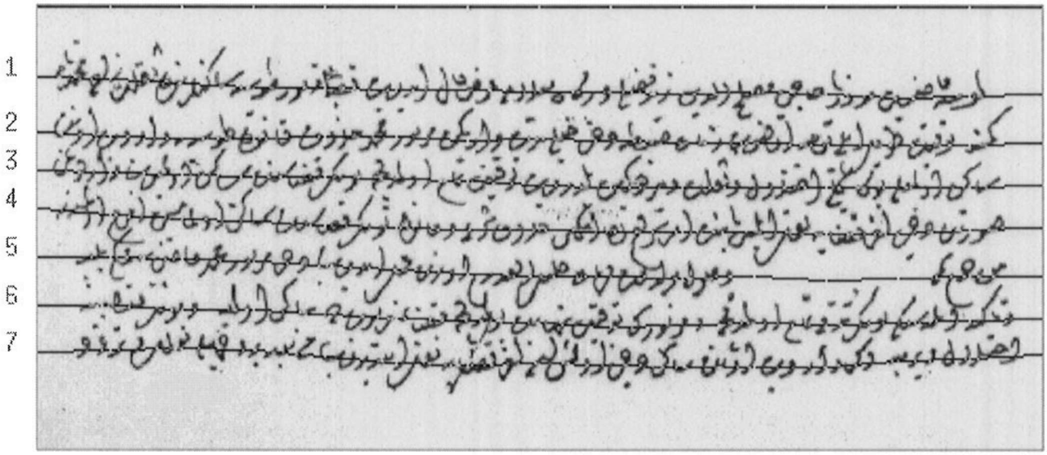
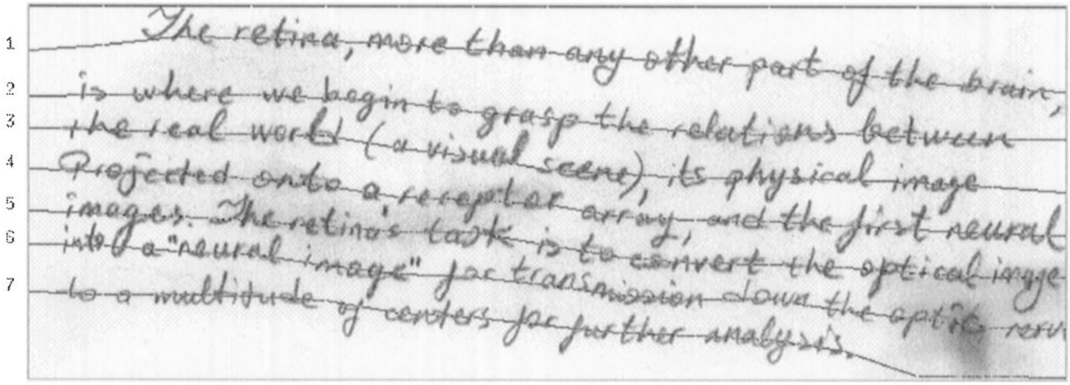
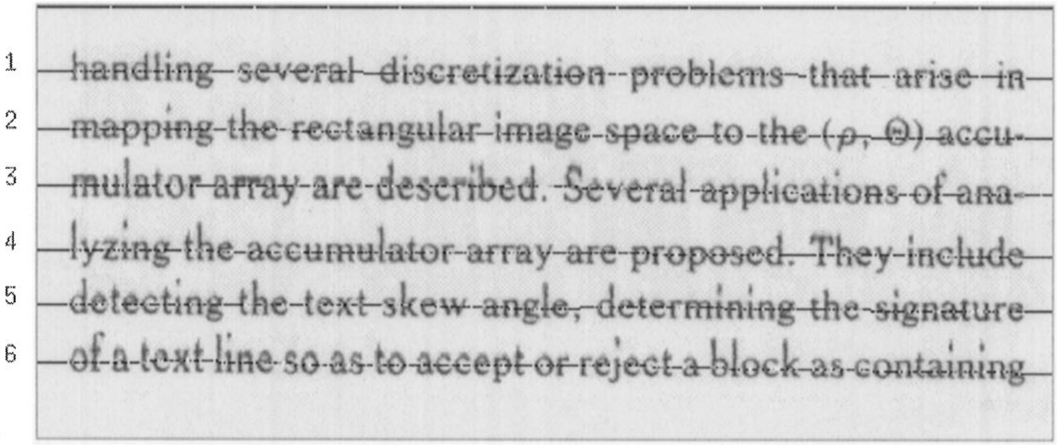


Fig. 4. Extracted baselines by RA network for a handwritten Ottoman text.



The retina, more than any other part of the brain,
is where we begin to grasp the relations between
the real world (a visual scene), its physical image
Projected onto a receptor array, and the first neural
images. The retina's task is to convert the optical image
into a "neural image" for transmission down the optic nerv
to a multitude of centers for further analysis.

Fig. 5. Baseline extraction by RA network for a handwritten Latin text.



handling several discretization problems that arise in mapping the rectangular image-space to the (ρ, Θ) accumulator array are described. Several applications of analyzing the accumulator array are proposed. They include detecting the text skew angle, determining the signature of a text line so as to accept or reject a block as containing

Fig. 6. RA network extracts baselines easily for a printed Latin text.

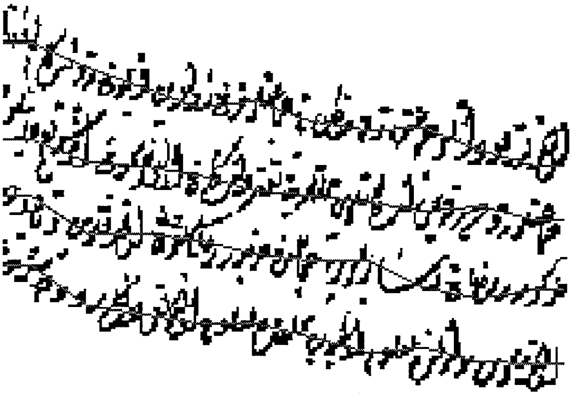


Fig. 7. Baseline extraction by using horizontal projection profiles.

purpose. In Fig. 7, the Ottoman handwritten document is presented with the extracted baselines by a simple projection method where the image is divided into vertical strips and a local maximum is sought inside the strips. This method requires binarization of the document image. Additionally, the method depends on many parameters such as the search for a local maximum value of the projections. It is clear that the curvature of the extracted baselines do not exhibit a stable behavior on the document image.

RA network or any of the methods mentioned in Section 2 can be applied in a straightforward manner to the baseline extraction problem of machine printed text documents. However, the major problem is the baseline extraction in complex environments, for example, when handwritten text has irregularities or high skew or even when there is overlapping of ascending and descending portions of characters of adjacent lines. In this case, algorithms do not provide meaningful results. Additionally, a quantitative assessment of the performance of the employed method is very difficult, simply because the identification of the real baseline becomes subjective, depending on the problem domain.

6. Conclusion

In this study, the RA network for baseline extraction on document images is described. It is based on an energy minimization principle. 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 binarization 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. The existing methods for baseline extraction also 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. In the proposed method, the baselines are not restricted to straight lines which is the situation for most handwritten documents. Additional difficulty in identification of baselines is introduced in hand written documents by the overlapping portions of ascending and descending characters between the baselines. The RA network baseline extraction method handles the problem of overlapping elegantly.

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] H.K. Aghajan, T. Kailath, SLIDE: subspace-based line detection, T-PAMI 16 (11) (1994) 1057–1073.

[2] A. Amin, Off-line arabic character recognition: the state of the art, Pattern Recognition 31 (5) (1998) 517–530.

[3] V. Atalay, M. Özçilingir, N. Yalabık, Computer recognition of Ottoman text, in: Proc. ISCIS V, Capadoccia, Turkey, 1990.

[4] R.O. Duda, P.E. Hart, Use of Hough transform to detect lines and curves in pictures, Comm. ACM 15 (1) (1972) 11–15.

[5] L.A. Fletcher, R. Kasturi, A robust algorithm for text string separation from mixed text/graphic images, T-PAMI 10 (1988) 910–918.

[6] R.M. Haralick, L.G. Shapiro, Computer and Robot Vision, Addison-Wesley, Reading, MA, 1992, Vol. 1, pp. 602–627.

[7] A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[8] J. Kanai, Text-line extraction using character prototypes, in: Workshop on Syntactic and Structural Pattern Recognition, Newjersey, 1990, pp. 182–191.

[9] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, Internat. J. Comput. Vision 1 (1988) 321–331.

[10] V.F. Leavers, Which Hough transform?, CVGIP: Image Understanding 58 (2) (1993) 250–264.

[11] E. Öztop, A.Y. Mülayım, S.S. Gül, F. Yarman-Vural, V. Atalay, Repulsive attractive network for baseline extraction, Technical Report TR.95-1, Dept. of Comp. Eng., METU, February 1995.

[12] A.A. Atıcı, F.T. Yarman-Vural, A heuristic algorithm for character recognition of Arabic script, Signal Processing 62 (1997) 87–99.