Novel Approach for Baseline Detection and Text Line Segmentation

Mahdi Keshavarz Bahaghighat
Electrical and Computer Department,Raja University, Qazvin, Iran.

Javad Mohammadi
Islamic Azad university of Takestan,Electrical and Computer Department
Takestan, Iran.

ABSTRACT
Baseline detection and line segmentation are essential preprocessing steps of any OCR system. In this paper we have proposed a robust and fast method for base lines detection based on projected pattern analysis of Radon Transform. The algorithm have been tested on more than 350 samples including both printed and handwriting of Persian/Arabic, English and also multilingual documents. Obtained results indicate that in spite of narrow interline spaces and noisy components our method is capable to extract baseline in documents precisely. In addition, in the case of multi-frequencies pattern, it has been shown that proposed method can reach its performance to accurate detection of base lines.

General Term
Image Processing, Document Analysis.

Keyword
Optical Character Recognition, Document Analysis, Multilingual Documents, Radon Transform, Neural Networks

1. INTRODUCTION
Analysis of document images for information extraction has become very prominent in recent years. Wide variety of information, which has been conventionally stored on paper, is now being converted into electronic form for better storage and intelligent processing [7, 6, 4].In any OCR system, preprocessing including scanning, image enhancement, skew estimation and correction, base line extraction are the primary steps that playan important role[2,3,7] in performance of a OCR system.Baseline is the virtual line on which semi cursive orcursive text are aligned/ joined. Generally baseline is kept in mind during both writing and reading. Baseline detection is not only used for automatic character recognition but it is also necessary for human reading. Without baseline detection it is very difficult and big issue to read the text even for humanand error rate increase up to 10%[14] while the context sensitive interpretation is involved.On the other hand, inaccurate segmented text lines will cause errors in the recognition stage.The rest of the paper is organized as follows: Section 2 introduces backgrounds of related works
while section 3 remark properties of Persian/Arabic scripts. Afterwards, we briefly describe the Radon Transform and its features. We then in section5 develop the principal of proposed algorithm to estimate baseline in Persian/Arabic and text line segmentation in English documents base on strong features of Radon Transform in details. Finally, our conclusions are presented in Section 6.

2. RELATED WORKS
Generally, text line segmentation approaches can be grouped into the different strategies such as projection based, smearing, grouping, Hough-based, graph-based and Cut Text Minimization (CTM) approach and etc.
In projection-based approach the vertical projection profile is obtained by summing pixel values along the horizontal axis for each y value. From the vertical profile, the vertical gaps between the text lines can be determined. A profile curve can be obtained by projecting black/white transitions or the number of connected components. The profile curve is then analyzed to find its maxima and minima [15, 16, 17].
In smearing technique, consecutive black pixels along the horizontal direction are smeared. If the distance between the white space is within a predefined threshold, it is filled with black pixels. The bounding boxes of the connected components in the smeared image are considered as text lines[18].
On comparison, grouping approach involves building alignments by aggregating units in a bottom-up approach. Units such as pixels, connected components, or blocks are then joined together to form alignments.
Likforman-Sulem and Faure [19] proposed an approach based on perceptual grouping of connected components of black pixels. Text lines are iteratively constructed by grouping neighboring connected components based on certain perceptual criteria such as similarity, continuity and proximity. Therefore local constraints on the neighboring components are combined with global quality measures. To handle conflicts, thetechnique merges a refinement procedure combining aglobal and a local analysis. According to the authors theproposed technique cannot be used on degraded or poorlystructured documents, such as modern authorialmanuscripts.
Furthermore, the Hough transform is used for locating straight lines in images. In [21] an iterative hypothesis validation strategybased on Hough transform was proposed.
Based on the authors, this technique is able to detect text line in handwritten documents which may contain lines oriented in different directions, erasures and annotations between main lines.

In the case of graph-based approach, a method based on a shortest spanning tree search was presented in [20]. The principle of the method consisted ofbuilding a graph of main strokes of the document imageand searching for the shortest spanning tree of this graph.
This method assumes that the distance between the wordsin a text line is less than the distance between two adjacenttext lines.
On the other hand, in[22] a new two-stage method for estimating and correcting the baseline of handwritten sub words in Farsi and Arabic text lines was introduced. It based on the template matching algorithm; the candidate baseline pixels were detected. The writing path and the baseline of the sub words are estimated in the first and second stages of the proposed algorithm, respectively. After the estimation in each stage, the baseline then was adjusted in the correction phase. Experimental results show the effectiveness of this approach in adjusting the baseline close to the correct position.
Finally, a novel piece-wise painting scheme was proposed by [8,9] to prepare patches of black and white blocks all along the text line, identify some candidate points, regress a curve through these candidate points to trace the baseline which is subsequently stretched straight horizontally and subsequently we de-tilt the characters to align the text-line with the horizontal imaginary baseline properly.

3. PROPERTIES OF PERSIAN /ARABIC

In general Persian/Arabic text either machine printed or handwritten is written cursively and from right to left. These letters are normally connected to the baseline.
According to Table1 an Arabic letter might have up to four different shapes, depending on its relative position in the word and this increases the number of classes from 28 to 100.Furthermore, Persian languages has four additional symbols which are shown in Table2 . In fact, Persian writing uses letters which consist of 32 basic letters, ten numerals, punctuation marks, spaces, and special symbols.
In order to make quick comparison between Persian/Arabic and English documents, Table3 have been drawn as well.
The problem of recognizing off-line Persian handwritten words is important in office automation, as well as in many other applications. Using the analytical approach to extract features included in Persian characters seems to be most appropriate due to the nature of Persian handwritten characters.
The handwritten Persian character has no fixed pattern, but has fixed geometrical features. The shapes of handwritten Persian characters differbetween writers, but fortunately the geometrical features are always the same.

4. RADON TRANSFORM

The radon function computes projections of an image matrix such as f along specified directions.
In fact, a projection of f(x,y) is a set of line integrals. The Radon function computes the line integrals from multiple sources along parallel paths, or beams, in a certain direction. The beams are spaced 1 pixel unit apart. To represent an image, the radon function takes multiple, parallel-beam projections of the image from different angles by rotating the source around the center of the image [1, 5, 10]. The Figure1 shows a single projection at a specified rotation angle.

Table1. Arabic alphabet in all its form

Name	Isolated	Initial	Medial	Final
Alif	ا	ا	ـا	ـا
Ba	ب	بـ	ـبـ	ـب
Ta	ت	تـ	ـتـ	ـة
Tha	ث	ثـ	ـثـ	ـث
Jeem	ج	جـ	ـجـ	ـج
Ha	ح	حـ	ـحـ	ـح
Kha	خ	خـ	ـخـ	ـخ
Dal	د	د	ـد	ـد
Zal	ذ	ذ	ـذ	ـذ
Ra	ر	ر	ـر	ـر
Zay	ز	ز	ـز	ـز
Sin	س	سـ	ـسـ	ـس
Shin	ش	شـ	ـشـ	ـش
Sad	ص	صـ	ـصـ	ـص
Dhad	ض	ضـ	ـضـ	ـض
Tta	ط	طـ	ـطـ	ـط
Az	ظ	ظـ	ـظـ	ـظ
Ain	ع	عـ	ـعـ	ـع
Ghain	غ	غـ	ـغـ	ـغ
Fa	ف	فـ	ـفـ	ـف
Qaf	ق	قـ	ـقـ	ـق
Kaf	ك	كـ	ـكـ	ـك
Lam	ف	لـ	ـلـ	ـل
Mim	م	مـ	ـمـ	ـم
Nun	ن	نـ	ـنـ	ـن
Ha	ه	هـ	ـهـ	ـه
Waow	و	ـو	----	ـو
Ya	ي	يـ	ـيـ	ـي

Table2. Four additional Persian alphabet

Name	Isolated	Initial	Medial	Final
Pe	پ	پـ	ـپـ	پ
Che	چ	چـ	ـچـ	چ
Gaf	گ	گـ	ـگـ	گ
Je	ژ	ژ	ژ	ژ

Table3. Differences between Latin and Persian Writing

	English	Persian
Direction	from left to right	from right to left
Connection	In general each character is connected to the next character with diagonal strokes	Persian letters are normally connected tothe baseline with horizontal strokes
Character	English characters	Persian letter might

versions	have few shape variations	haveup to four differentshapes, depending on its relative position in theword
Features	English Writing has specific geometrical features	Persian writing has a unique feature for each character, especially curves and dots
Segmentation	Any analytical segmentation approach can segment the handwriting into different letters or sub-letters	The letters or segmented sub-letters are different from segments in English

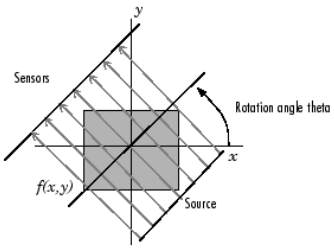


Fig1.Parallel-Beam Projection at Rotation Angle Theta

Projections can be computed along any angle. In general, the Radon transform of f(x,y) is the line integral of f parallel to the y´-axis[11, 12, 13]:

$$R_\theta(x') = \int_{-\infty}^{+\infty} f(\rho^1_{x,y,\theta}, \rho^2_{x,y,\theta})dy \qquad (1)$$

$$\rho^1_{x,y,\theta} = x'\cos(\theta) - y'\sin(\theta) \qquad (2)$$

$$\rho^2_{x,y,\theta} = x'\sin(\theta) + y'\cos(\theta) \qquad (3)$$

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \qquad (4)$$

Figure2 indicates geometry of Radon Transform .

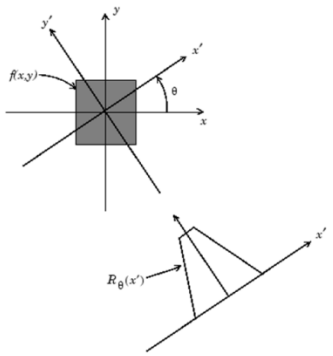


Fig2: Geometry of Radon Transform

Usually local region in the document image has a consistent orientation and frequency. Therefore, it can be modeled as a surface wave that is characterized completely by its dominate orientation and frequency pattern. This approximation model is useful enough for our purpose to evaluate the features of Radon Transform. A local region of the image can be modeled as a surface wave[25] according to Eq.5:

$$I(x, y) = A\cos(2\pi f(x\cos(\theta) + y\sin(\theta))) \qquad (5)$$

Then an example image and its projection by Radon Transform are shown in figure3.

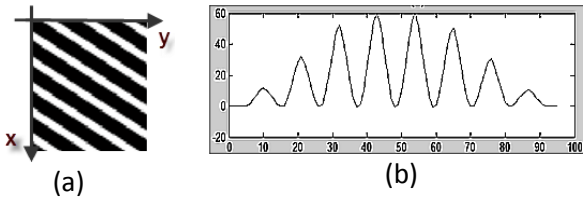


Fig 3: (a) A well-defined 65×65 synthetic image with $\theta = 60$.
(b) Radon Transform of image, R(x).

As we can see in the projection function in figure3.(b), providing that it have been projected on correct orientation which is parallel to the local orientation of input image, it can approximately treat as a sinusoidal plane wave.
Experiments were then conducted with a Gaussian noisy elements applied to the images.Although SNR of noisy image is relatively high, figures 4 And 5 indicate that the presence of noise does not considerably impact the projected pattern. This means that it can noticeably tolerate added noise

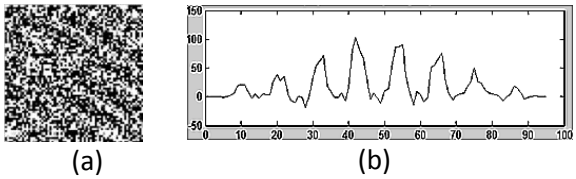
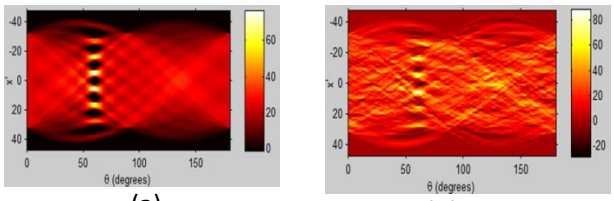


Fig4:(a) Noisy image, Gaussian Noise with SNR=-5dB.
(b) Radon Transform



(a) (b)

Fig5:(a) Radon Transform Map of noise free image (b) Radon Transform Map of noisy image with SNR=-5dB.

5. PROPOSED ALGORITHM

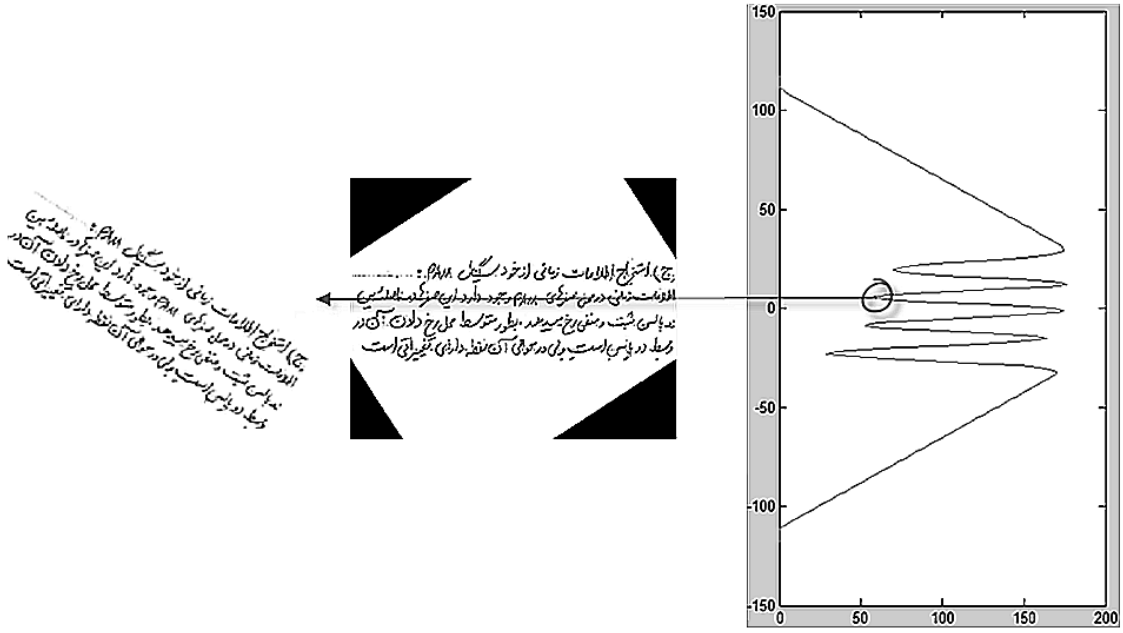
After introducing the theory of Radon Transform, now we have tendency to proposethe main steps of our developed algorithm to baseline detection and text line segmentation in the following stages:

1) Binarisation: This process involves examining the grey-level value of each pixel in the enhanced image, and, if the value is greater than the global threshold, then the pixel value is set to a binary valueone; otherwise, it is set to zero. The outcome is a binary image containing two levels of information, the foreground ridges and the background valleys[3, 7].

2) Orientation estimation: The skew is estimated from the binary image by applying the method mentioned in [2].

3) Rotation Compensation: it is necessary to shift the angle in anticlockwise direction by Nearest Neighborhood Method[2](see Figure6(b)).

4) Base line extraction: Detecting baseline is one of the main majorities in preprocessing OCR system.
We propose novel approach for base line detection base on extraction local minima in projected pattern of Radon Transform. An example of a projected waveform is shown in Figure6(c). This projection forms an almost sinusoidal-shape wave with the local minimum points corresponding to the base line in the document image (note encircled point in Figure6(c)). Therefore, the base line easily is then computed by counting the number of pixelsbetween consecutive minima points and their locations in the projected waveform.



(a) (b) (c)

Fig6: Base line extraction base on finding local minimum in projected pattern of Radon transform
(a) Multilingual handwritten Persian-English Image(b)Skew corrected image (c) Radon Transform of skew corrected image

To make more evaluation, further examples are shown in Figure7. Obtained results clearly indicate that proposed method compared to other well-known existing methods not only is capable to extract text lines in any document easily and precisely without requiring to large numbers of processing modules but also it can extract the positions of all lines in a given document simultaneously which it means that it will strongly reduce the total process time.

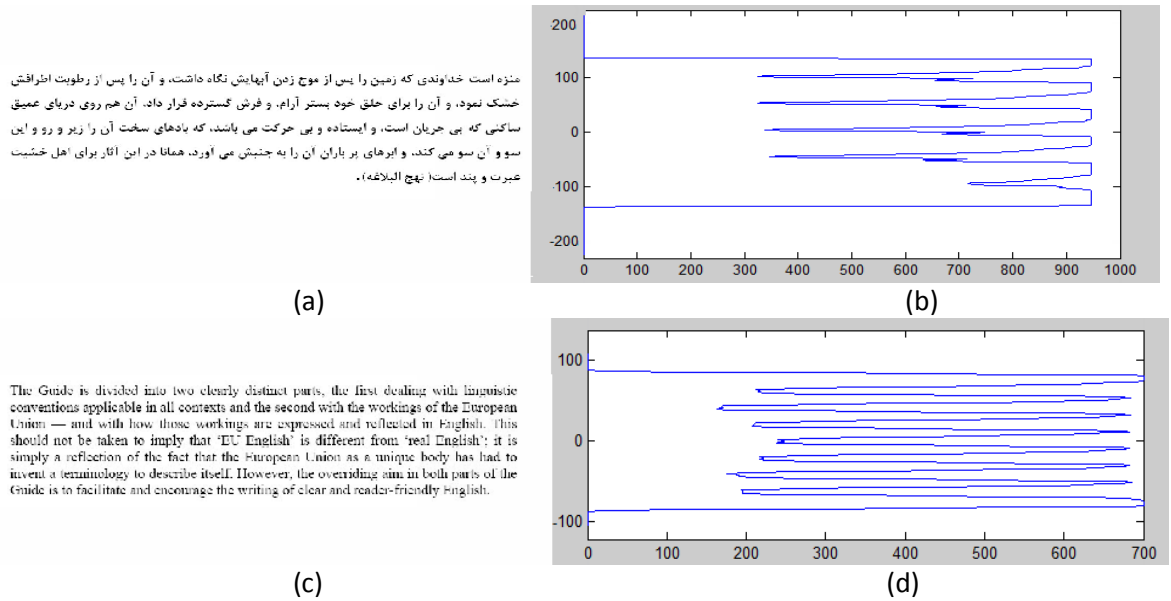


Fig7:Some examples of baselinedetection and line segmentation for printed documents
(a),(b)Persian documentand itsRadon Transform
(c),(d)Low-quality English document imageand itsRadon Transform

Now consider the situation in which noise was added to input image (see Figure8(a)).
As shown in Figure 8(b), noisy elements in the projection will result in the production of false local minima, which may impact the location of the true minimum points. These false minima can then cause an inaccurate estimation of base line. However, as illustrated by Figure 8(c), if the projection is smoothed by Savitzky-Golay digital smoothing filter[24] prior to estimating baselines,the noisy
elements will be eliminated, so leaving only the true local minimum points.
This additional step has shown to be useful in reducing noise effects in the projection, and consequentlyit will be essential to improve the accuracy of the process.

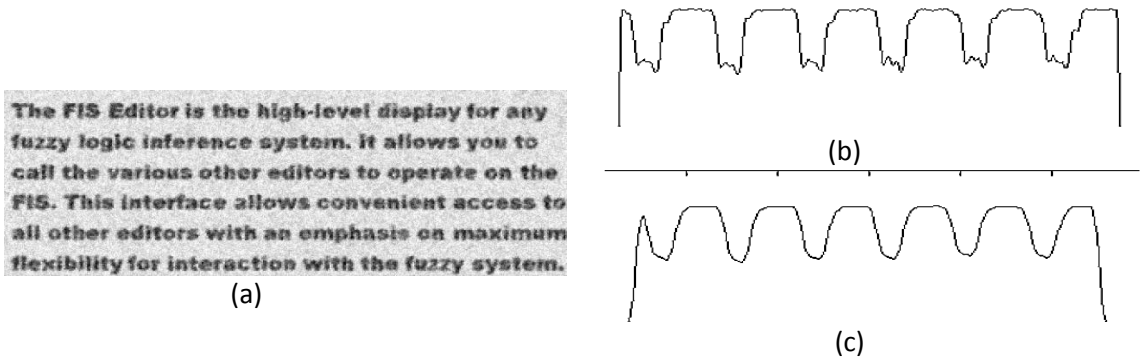
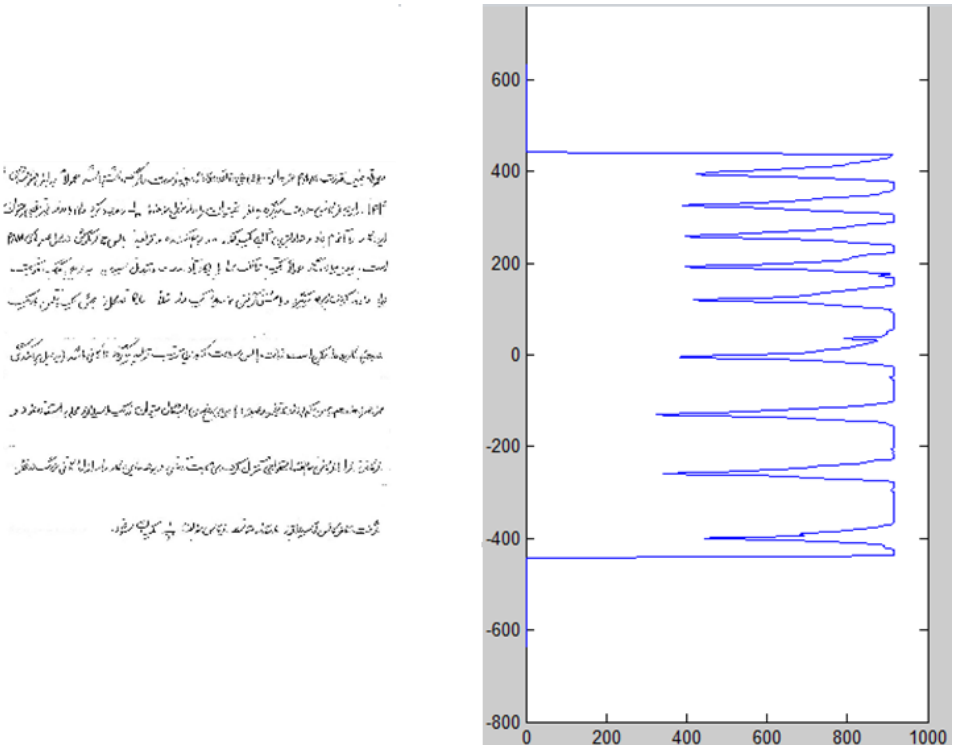


Fig8: The effect of smoothing the projection prior to inter- line detection.
(a) Noisy image (b) Radon Transform before smoothing (c) Radon Transform after smoothing

Furthermore, sometimes document images may have several frequencies in their patterns. This fact has been illustrated in Figure9(a) as an example of low-quality Persian handwritten input image . It can be viewed in Figure9(b) that its Radon Transform



(a) (b)

Fig9: Base line detection in low quality and multi-frequencies Persian handwritten document (a) Low-qualityPersian handwritten input image (b) Radon Transform

projection has two dominate frequencies according to upper and lower parts of input image in Figure9(a). In the flowing stage we intend to distinguish between these parts by using Distributed Time-Delay Neural Network(TDNN)[23].

The FTDNN had the tapped delay line memory only at the input to the first layer of the static feed forward network. The distributed TDNN was first introduced in [23] for phoneme recognition.It is possible for distributed TDNN to distribute the tapped delay lines throughout the network. The original architecture was very specialized for that particular problem. Figure10 shows general two-layers distributed TDNN.

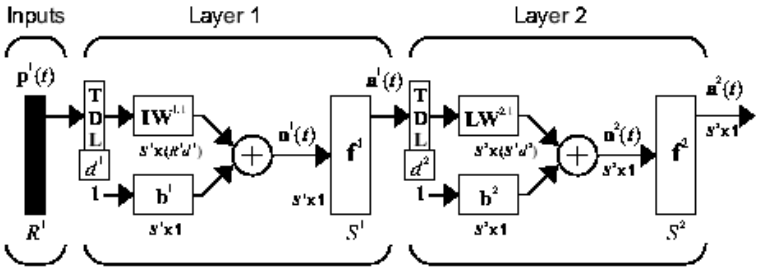


Fig10: Network structure for Distributed Time-Delay Neural Network

This network can be used for a simplified problem that is similar to phoneme recognition. The network will attempt to recognize the frequency content of an input signal. ConsiderFigure9(b) or its equivalent Figure 11(a) as a input signal to defined network in which one of two frequencies is present at given range corresponding to multi-frequencies pattern of document image input in fjgure9(a).

Target network output are defined to be 1 when the input is at the low frequency and −1 when the input is at the high frequency(according to two parts of input document image in figure9(a)) .

After training networks with epochs = 120 and simulation with obtained weights,Figure11have been shown for monitoring the trained network output. The result indicates that network is able to distinguish between two and morephonemes accurately.

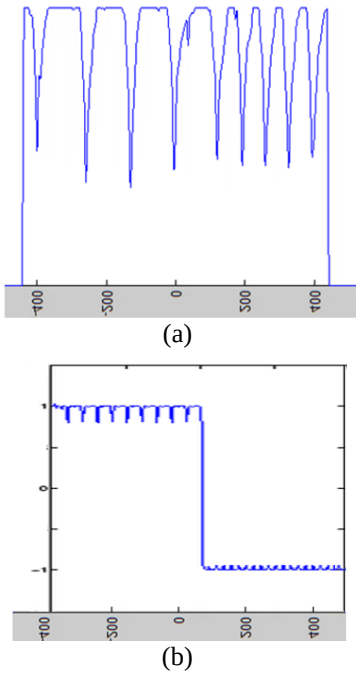


Fig11: Monitoring the trained network output.
(a) Projected pattern of Radon Transform
(b) Output of Neural Network

6. CONCLUSION

In this literature, the baseline detection and line segmentation method are discussed in detail. The proposed method was based on strong features of Radon Transform in terms of robustness and time efficiency. It just depends on analyzingprojected pattern of Radon Transform to find local minima in it.

The algorithm have been tested on more than 350 samples including both printed and handwriting of Persian/Arabic, English and also multilingual documents.The main prominence of our approach rather than the other methods is that in our algorithm, the baseline is estimated for each line simultaneously which can noticeably increase the timing performance. In addition, in the case of multi-frequencies pattern, it has been shown that proposed method can reach its performance to accurate detection of base lines.

7. ACKNOWLEDGMENTS

Authors would like to thank Mehri Noormohammadi for her supports, helps and considerations.

REFERENCES

[1] Z.A. Khan, W. Sohn,." Real Time Human Activity Recognition System basedon Radon Transform", IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" AIT, 2011.

[2] H.K.Chethan, G.Hemantha Kumar,2010. Graphics separation and skew correction for mobile captured documents and comparative analysis with existing methods, International Journal of Computer Applications(IJCA) (0975 – 8887), Volume 7– No.3.

[3] N. Priyanka, S.Pal, R. Mandal,"Line and Word Segmentation Approach for Printed Documents" IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" , RTIPPR, 2010.

[4] S.Akram, M.U.Din Dar, A.Quyoum, "Document Image Processing - A Review", International Journal of Computer Applications (IJCA) (0975 – 8887) Volume 10– No.5, November 2010.

[5] A.Bouchemha, A.Nait-Ali,N. Doghmane,"A Robust Technique to Characterize the Palmprint usingRadon Transform and Delaunay Triangulation" , International Journal of Computer Applications (0975 – 8887)Volume 10- No.10, November 2010.

[6] S.Prasad,V.K.Singh,A. Sapre, "Handwriting Analysis based on Segmentation Method for Prediction of Human Personality using Support Vector Machine", International Journal of Computer Applications (0975 – 8887)Volume 8– No.12, October 2010

[7] M.Hangarge,B.V.Dhandra,"Offline Handwritten Script Identification in Document Images", International Journal of Computer Applications (0975 – 8887) Volume 4 – No.6, July 2010.

[8] P. Nagabhushan, A.Alaei, "Tracing and Straightening the Baseline inHandwritten Persian/Arabic Text-line: A NewApproach Based on Painting-technique", (IJCSE) International Journal on Computer Science and EngineeringVol. 02, No. 04, 2010, 907-916.

[9] A.Alaei, P. Nagabhushan, U.Pal, "A New Text-line Alignment Approach Based on Piece-wise Painting Algorithm for Handwritten Documents", 2011 International Conference on Document Analysis and Recognition, IEEE DOI 10.1109/ICDAR.2011.73.

[10] VahidKiani, Reza Pourreza& Hamid Reza Pourreza, Offline Signature Verification Using Local Radon Transform and Support Vector Machines,International Journal of Image Processing (IJIP) Volume(3), Issue(5).

[11] J.Mohammadi, M.Keshavarz Bahaghighat ,R.Akbari , ,Vehicle Speed Estimation Based On The Image Motion Blur Using RADON Transform , 2010 2nd International Conference on Signal Processing Systems (ICSPS).

[12] F. Hjouj, D.W. Kammler. "Identification of Reflected, Scaled, Translated, and Rotated Objects From Their Radon Projections". IEEE Transactions on Image Processing, 17(3):301-310,2008.

[13] M. R. Hejazi, G. Shevlyakov, Y-S Ho. "Modified Discrete Radon Transforms and Their Application to Rotation-Invariant Image Analysis". IEEE 8th Workshop on Multimedia Signal Processing, 2006.

[14] M.I.Razzak, M.Sher ,S. A.Hussain," Locally baseline detection for online Arabic scriptbased languages character recognition", International Journal of the Physical Sciences Vol. 5(7), pp. 955-959, July 2010.

[15] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "ArabicHand-written Text-line Extraction", in Proceedings of theSixth International. Conference on

Document Analysis andRecognition, ICDAR 2001, Seattle, USA, September 10-132001, pp. 281–285.

[16] M. Arivazhagan, H. Srinivasan, S. N. Srihari, "A Statistical Approach to Handwritten Line Segmentation", in Proceedings of SPIE Document Recognition and Retrieval XIV , San Jose, CA, February 2007.

[17] G. Tímár, K. Karacs, Cs. Rekeczky, "Analogic Preprocessing and Segmentation Algorithms For Offline Handwriting Recognition", Proceedings of IEEE CNNA'02,World Scientific 2002, pp.407-414.

[18] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "A new algorithm for detecting text line in handwritten documents," in International Workshop on Frontiers in Handwriting Recognition, 2006, pp. 35–40.

[19] L. Likforman-Sulem, C. Faure, "Extracting text lines in handwritten documents by perceptual grouping", Advances in handwriting and drawing: a multidisciplinary approach,C. Faure, P. Keuss, G. Lorette and A. Winter Eds, Europia, Paris, 1994, pp. 117-135.

[20] I.S.I. Abuhaiba, S. Datta, M.J.J. Holt, "Line Extraction and Stroke Ordering of Text Pages", Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, Canada, 1995, pp. 390-393.

[21] L. Likforman-Sulem, A. Hanimyan, C. Faure, "A Hough based algorithm for extracting text lines in handwritten documents", Third International Conference on Document Analysis and Recognition, Vol. 2, August 1995, pp. 774-777.

[22] M. Ziaratban and K. Faez, "A Novel Two-Stage Algorithm forBaseline Estimation and Correction in Farsi and Arabic Handwritten Text line," Proc. of International Conference on Pattern Recognition(ICPR'08) , 2008, pp. 1-5.

[23] Waibel, A., T. Hanazawa, G. Hilton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, 1989, pp. 328–339.

[24] J.Luo, K.Ying, P.He, J.Bai ,"Properties of Savitzky–Golay digital differentiators", Elsevier Inc. Digital Signal Processing 15 (2005) 122–136.

[25] Sharat S. Chikkerur, Alexander N. Cartwright and Venu Govindaraju, "Fingerprint Image Enhancement Using STFT Analysis", PHD thesis , Center for Unified Biometrics and Sensors University at Buffalo, NY, USA (2003).