Skew Angle Estimation of Urdu Document Images: A Moments Based Approach

R. J. Ramteke, Imran Khan Pathan, S. C. Mehrotra

***Abstract*— The performance of an OCR system will not be satisfactory for most of the scanned images without accurate skew correction. This paper presents the skew angle estimation and correction for Urdu document images script using moments method. The basic idea is to draw a random polygon over the text in document. This leads to thinning free preprocessing. The skew angle is calculated using Central moments and centroid of the document image. Experimental results are found to be satisfactory and compared with other skew detection techniques.**

***Index Terms*— Skew Estimation, Moments, Optical Character Recognition, Urdu Document.**

I. Introduction

System based optical character recognition has received considerable attention in recent years due to rising automations of various systems. The conversion of existing documents into electronic once for better archival, retrieval and maintenance is growing. Character recognition has variety of applications in various fields like, reading bank checks, reading postal zip code, reading passport numbers, employee code, office automation etc. Optical character recognition is the process of converting digital image of text, such as a scanned paper document or electronic fax file, into computer-editable text. Typical application of an OCR is to convert a scanned image of text on paper to a text document that can be further processed on a computer. It can be a great help to visually impaired when interface with voice synthesizer.

In general, there are five major stages in OCR processing:

1) Preprocessing
2) Segmentation
3) Representation
4) Training and Recognition
5) Post Processing

There are several problems encountered in processing a document. Preprocessing is the primary stage before starting actual character recognition. Initially we obtain a digitized raster image of the document using appropriate scanning method. When the document image is scanned, it may be skew because of some reasons. The skewed image will cause serious problems in document analysis. Some processing needs the documents without skew, such as character extraction and recognition, structural analysis, and so on. Although some applications can use directly the skew documents, they are too complex and inefficient. It is, therefore, often necessary to determine the skew angle and reconstruct the document. Skew detection and correction indeed helps the subsequent stages in document image analysis. Devising schemes which can detect skew angles accurately irrespective of the scripts and range of skew angles is a challenging task in the field of document image analysis.

Preprocessing focuses on enhancing the scanned image to make feature extraction easy and correct and to reimburse for the eventual poor quality of the scanned document. The all the other stages of OCR systems mainly depend upon the accuracy of preprocessing stage. During the scanning process, the whole document or a portion of it is fed through a scanner. The digital image of a document may be skewed / rotated arbitrarily because of direction in which it was placed on the platen when it was scanned or because of a document feeder malfunction. A significant skew in document can be detected by human vision easily and the skew correction can be made by re-scanning the document, whereas for mild skew it may not be possible to notice its skew as human vision system fails to identify it. Even a smallest skew angle existing in a given document image results in the failure of segmentation of complete characters from words or a text lines, as the distance between the character reduces. Further, most of the OCRs and document retrieval/display systems are very sensitive to skew in document images

Skew angle is the angle that the text lines in the digital image make with the horizontal direction. Therefore, skew estimation and correction are important steps before line and words segmentation. Various methods have been built for document skew angle estimation reported in the literature [1]. Chen Yi-Kai [2] has also proposed another method based on Fourier transform. In this method, the direction for which the density of the Fourier space is the largest gives the skew angle. The method requires the computation of the Fourier transform, which can be time consuming for a large image. Hong Yah [3] presented a method based on the cross-correlation between two lines in the image with a fixed distance. The correlation functions for all pairs of lines in the image are accumulated. The shift for which the accumulated cross-correlation function takes the maximum is then used for determining the skew angle. The image is rotated in the opposite direction for skew angle.

Skew angle detection is considered as a significant part of any Optical Character Recognition and document analysis

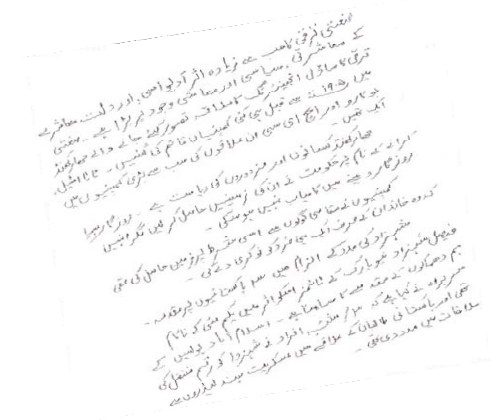
Manuscript received March 22, 2011.

R. J. Ramteke is with the Department of Computer Science, North Maharashtra University, Jalgaon-425001, (M.S.) India. phone: +91-9890688672; e-mail: rakeshramteke@yahoo.co.in.

Imran Khan Pathan, is with Miliya College of Arts and Science, Beed, (M.S.), India. He is now pursuing his Ph.D. at North Maharashtra University, Jalgaon-425001, (M.S.) India. (e-mail: mca_imran2003@yahoo.co.in).

S. C. Mehrotra is with the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad-431001, (M.S.), India (e-mail: mehrotrasc@rediffmail.com).

system because correct skew angle has a direct effect on the segmentation and feature extraction stages and OCR system performance. Figure 1(a) shows the skewed Urdu handwritten document image and Figure 1 (b) shows Skewed Urdu news paper image.



(a)

(b)

Fig. 1 Skewed images of (a) Urdu handwritten document. (b) Urdu news paper image

Consequently Many techniques were implemented like Hough transform, Cross Correlation, Projection Profile, Fourier transform and K Nearest Neighbor (K-NN) clustering etc. Most of the skew detection methods have the following common features: (1) a prior text/graphics separation is necessary, which may take a significant amount of time, though it can be useful for the next steps; (2) large text areas have to be present on a page for an accurate estimation; (3) many techniques have been designed for high-resolution images ranging from 100 to 300 dpi [4].

In projection profile technique a series of projection profiles are obtained at a number of angles close to the expected orientation, and the variation is calculated for each of the profiles. The profile that gives maximum variation corresponds to the projection with the best alignment to the text line, this projection angle is called the skew angle [5]. The Hough transform [6, 7] is another popular technique for skew detection, this transform is often applied to a number of representative points of characters such as the lowermost pixels or centers of gravity. Each representative point (x,y) is mapped from the Cartesian space to the points (ρ,θ) in the Hough space by forming a set of lines coming through (x,y) with a slope ρ and distance θ from the origin. The skew corresponds to the angle associated with a peak in the Hough space. The high computational complexity of the Hough transform often imposes restrictions on the possible angle range.

This paper describes a simple method for the skew estimation and correction of Urdu document image using central moments and centroid of the scanned image. Random polygon drawn over the text in document image and skew angle is calculated. The detail about Urdu script is presented in section 2. Proposed methodologies are given in section 3. Experimental results are reported in section 4. Section 5 discusses the results and comparison and conclusion is reached in section 6.

II. URDU SCRIPT

India is multilingual country of more than 1.25 billion population with 18 constitutional languages and 10 different scripts. Urdu is one of the popular Indian scripts and there must be now at least sixty million people in South Asia who regard Urdu as their mother tongue. There must be twice as many, perhaps more, who understand Urdu and world even use it on occasion, in conversation if not in writing. While literary history of Urdu goes back to the fifteenth century, specimens of it begin to be found as early as the thirteenth. Presently, Urdu is the official language of Pakistan and one of the sixteen major languages constitutionally recognized in India. Urdu speakers and publications can also be found in substantial numbers in the Middle East, England and North America [8].

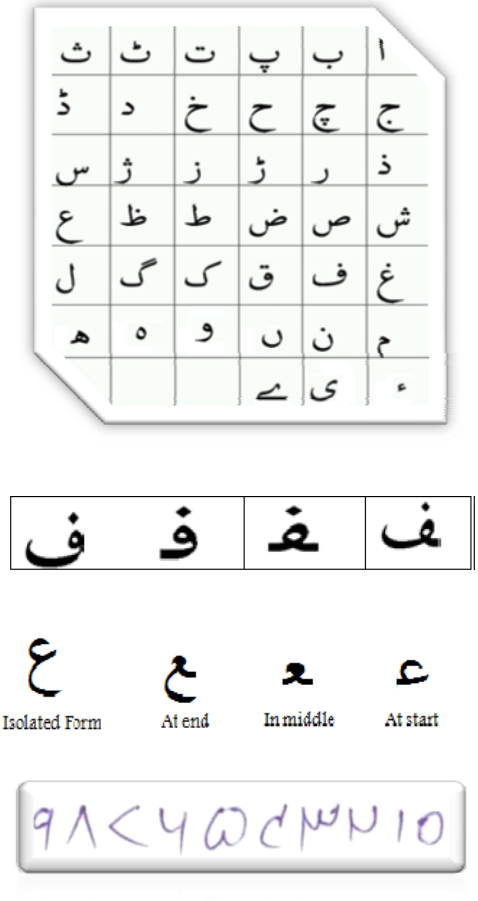


Fig. 2 Examples of Urdu alphabet and numerals

It is cursive script, written from right to left, like Arabic and Farsi but with some added alphabets. It is therefore the OCR

systems used for Arabic and Farsi can't suit the need of Urdu script. Like other Indian scripts in Urdu also two or more characters may combine and create a complex shape called compound characters. But Urdu characters vary from each other on the basis of small changes in their shape and the hat features that carry very important information. Also depending on the positions (first, middle or last) in a word the basic shape of a character may be changed. Thus, OCR development for Urdu is more difficult than any European language script having a smaller number of characters [9]. Unlike to Devanagari script in which the head line (Shirorekha) is used to detect skew angle [10], the Urdu script has no such guide line. Moreover, the fashion in which the black pixels comprise the characters of Urdu script are through vertical expansion on both sides of base line i.e. the base line cannot be predicted easily, unlike Roman Script in which almost all characters are expanded above the base line. Therefore the skew detection of Urdu script is seems to be tougher task.

III. METHODOLOGY

The text image which is scanned at an angle can be rotated to a normal position following the series of steps. The steps to be followed for getting back to the normal position include:

1) Base line identification
2) Centroid calculation using moments
3) Skew angle estimation using central moments and correction.

The base line identification [11] is generally the most important step of the whole process. Baseline is the line along which the center of gravity or centre of mass or centroid of the word hangs. The text document is inscribed by considering the farthest pixel in the four directions and determined the rectangle. Text image may contain more than four corner points, in that case text image has to be considered as polygon. Although a rectangle is used as a polygon, it can be calculated as:

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i\ y_{i+1} - x_{i+1}\ y_i) \quad (1)$$

Where A is the Area of polygon,

A. Moments

In this section, moments based method presented in [12] is adopted for computing skew angle using extracted coordinate. Moments are measures of the pixel distribution around the center of gravity of the document image and allow capturing the global shape information. For discrete 2D image, the moments are evaluated using equation 1. Further, for the binary image, moments of the image function f(x, y) are evaluated in equation 2 as follows:

$$m_{pq} = \sum_{0}^{n}\sum_{0}^{n} x^p\, y^q \quad (2)$$

Where for p, q = 0, 1, 2,….and n * n is the size of the image.

From equation 2, central moments for discrete image can be calculated in equation 3, equation 4 for binary image

$$\mu_{pq} = \sum_{0}^{n}\sum_{0}^{n} (x-\overline{x})^p (y-\overline{y})^q f(x,y) \quad (3)$$

$$\mu_{pq} = \sum_{0}^{n}\sum_{0}^{n} (x-\overline{x})^p - (y-\overline{y})^q \quad (4)$$

$$m_{10} = \sum_{0}^{n}\sum_{0}^{n} x^1 - y^0 \quad (5)$$

$$m_{01} = \sum_{0}^{n}\sum_{0}^{n} x^0 y^1 \quad (6)$$

$$m_{00} = \sum_{0}^{n}\sum_{0}^{n} x^0 y^0 \quad (7)$$

Using equation 5, 6 and 7 we compute equation 8 and 9

$$\overline{x} = \frac{m_{10}}{m_{00}} \quad (8)$$

$$\overline{y} = \frac{m_{01}}{m_{00}} \quad (9)$$

$$\mu_{11} = \sum_{0}^{n}\sum_{0}^{n} (x-\overline{x})^2 - (y-\overline{y})^0$$

$$\mu_{11} = m_{20} - \frac{2m_{10}{}^2}{m_{00}} + \frac{m^2 10}{m_{00}}$$

$$\mu_{11} = m_{11} - \bar{x} m_{01}$$

$$\mu_{11} = m_{11} - \bar{y} m_{10} \quad (10)$$

$$\mu_{20} = \sum_{0}^{n}\sum_{0}^{n} (x-\overline{x})^2 - (y-\overline{y})^0$$

$$\mu_{20} = m_{20} - \frac{2m_{10}{}^2}{m_{00}} + \frac{m^2 10}{m_{00}}$$

$$\mu_{20} = m_{20} - \frac{m_{10}{}^2}{m_{00}}$$

$$\mu_{20} = m_{20} - \bar{x} m_{10} \quad (11)$$

$$\mu_{02} = \sum_{0}^{n}\sum_{0}^{n} (x-\overline{x})^0 - (y-\overline{y})^2 = m_{02} - \frac{m_{01}{}^2}{m_{00}}$$

$$\mu_{02} = \mu_{02} - \bar{y} m_{01} \quad (12)$$

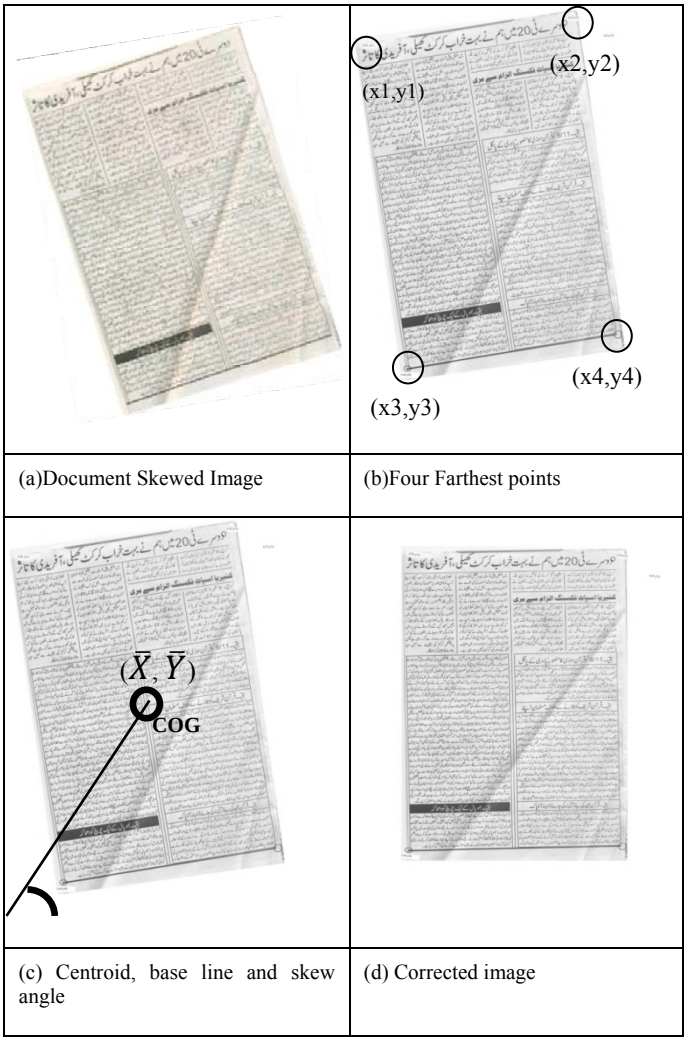
The theta (skew angle) is computed using equation 10,11and 12 and it results in equation 13.

$$\Theta = \frac{1}{2}\ tan^{-1}\left[\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right] \quad (13)$$

Equation 13 gives the θ for the set of points, which is an estimated skew angle.

IV. SKEW CORRECTION ALGORITHM

The proposed method comprises of subsequent seven steps, where Urdu text image is an input and predictable output will be skewed free Urdu text image. The following figure 4 shows the steps of the processing.



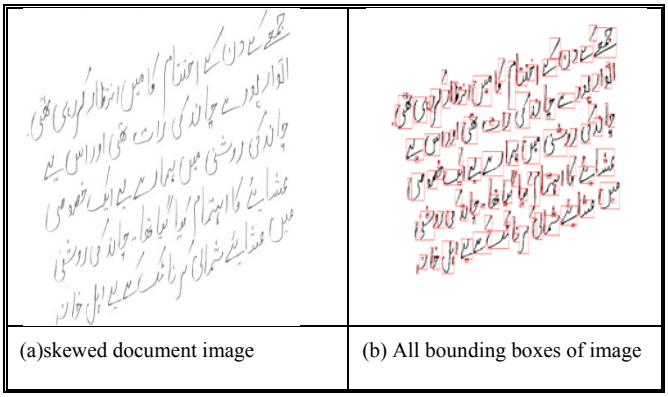
(a)Document Skewed Image | (b)Four Farthest points

(c) Centroid, base line and skew angle | (d) Corrected image

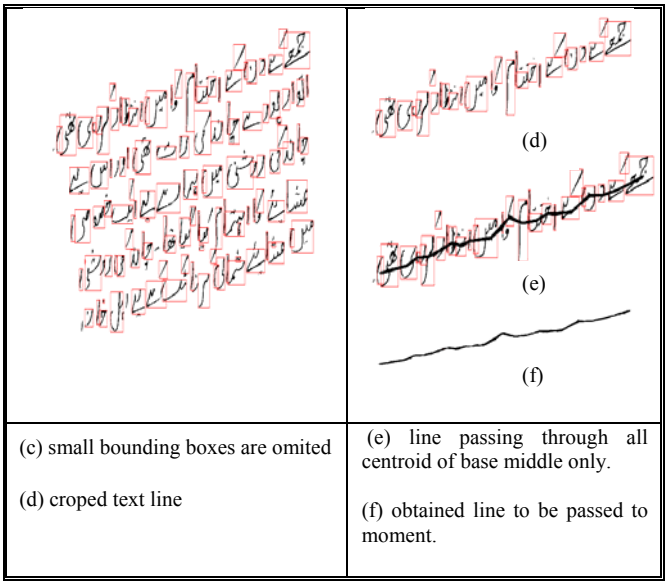
Fig.3 Example of processing steps

A. *Algorithm -I*

1) Analyzing document image whether some skewed is present or not as shown in fig 3 (a)
2) Four farthest points are determined in possible four directions as shown in fig 3(b).
3) Area of polygon of image is calculated using equation number three.
4) Centroid – Centre of Gravity of image is calculated using values of $\bar{x}$, $\bar{y}$ coordinates of all four point with help of equation 8 and 9 respectively fig 3(c).
5) The moments are calculated up to second order.
6) Find the angle θ using equation 13, which is in fact the skew angle of the image fig 3(c).



(a)skewed document image | (b) All bounding boxes of image



(c) small bounding boxes are omited

(d) croped text line

(e) line passing through all centroid of base middle only.

(f) obtained line to be passed to moment.

Fig. 4 Processing steps to get line passes through all centroid

B. *Algorithm -II*

1) Region properties of skewed image shown in figure 4(a) are derived to acquire bounding boxes of all characters.
2) Using region properties of images bounding boxes for all isolated character elements are obtained as shown in figure 4(b).
3) Small bounding boxes are omitted to abstract actual words and characters and skip top or bottom featuring elements like zabar, zer, pesh, dot etc as shown in figure 4(c).
4) Obtained four farthest points of first line and fetch first line separately as shown in figure 4(d).
5) A line is drawn through all centroids of middle elements bounding boxes where top and bottom elements are omitted as shown in figure 4(e).
6) Coordinates of obtained line passes through centroids are used to calculate moments as shown in figure 4(f).
7) Average angle of Algorithm I and Algorithm II is calculated and the document is rotated in clockwise or anticlockwise direction with determined skew angle as shown in figure 3(d).

V. Experimental Results and Discussion

The algorithm has been tested on a set of 125 different types of Urdu documents images. Various handwritten text, Urdu news paper, magazine, text books images are tested using Intel core 2 duo 2.09 GH on MATLAB 7.4 with different skewed angle, samples are shown in figure 4. The proposed method attempted on both handwritten and the printed documents. By analyzing the result in table I it clear that the proposed method work much better on printed documents as compare to handwritten document images. This method works for different types of documents with various image resolutions by several effects on processing speed of the system. A better document analysis can improve the system efficiency and system could work more efficiently on both left and right skewed images with any of orientation of image. Fig. 5 shows various images used as input to the system and the results are reported in the tables.

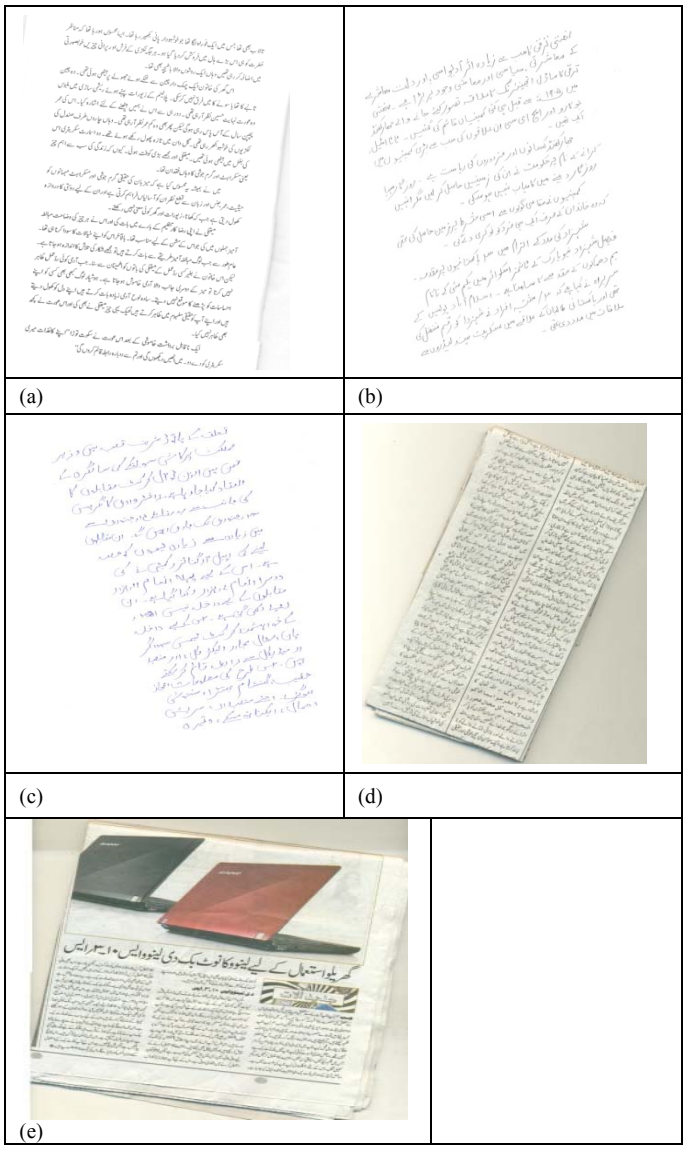


Fig. 5 Some sample images used as an input

The results are reported in following tables. The skew angle calculated by the proposed method shown in Table I. We have used four types of document samples for testing i.e. Handwritten rolled page, newspaper, magazine page and text book page. The Table II depicts the overall performance of the system. Mean and standard deviation is calculated for the different samples based on the true angle in 8, 9 and 10 degrees. The sample images were captured manually for these specific degrees.

TABLE I. SKEW ANGLE CALCULATED BY PROPOSED METHOD

Image	Actual skew angle	Skew angle identify by proposed method
Fig 5 (a)	8	10
Fig 5 (b)	9	10
Fig 5 (c)	10	10
Fig 5 (d)	23	25
Fig 5 (e)	15	17

TABLE II. RESULTS WITH RESPECT TO DOCUMENT'S TYPE

Document Type	Sample Images	True Angle in degree	Mean	SD
Handwritten Page	20	8	8.095	0.889
	15	9	8.8	1.146
	10	10	10.6	1.173
News paper	5	8	7	1
	10	9	9.2	1.398
	10	10	9.7	1.567
Magazine	10	8	9.9	1.286
	10	9	9.3	1.418
	10	10	9.4	0.843
Text Book	10	8	7.8	1.229
	10	9	8.6	0.699
	5	10	10	2.54

VI. CONCLUSION

A simple, efficient and fast document image skew detection and alteration technique is presented. Centroid of image is calculated by determining four farthest points of image and skew angle is finding through base line. Further, the moments are evaluated up to second order to achieve the skewed angle. The proposed method work more efficiently on images of printed news papers, magazines, books as well as handwritten documents. The performance and accuracy of proposed method can be improved on noise free images. The noises and the variation in the document resolution are still the main challenges facing in the Urdu skew detection and correction methods. The Proposed method was applied on different types of 125 documents which give different accuracy parentage as shown in table I and II, where skewed angle of printed document was calculated more accurate as compare to handwritten document images and concluded that a better document analysis can improve the system efficiency.

REFERENCES

[1] L.A. Fletcher and R. Kasturi, "A robust algorithm for Text string separation from mixed text/Graphics image", IEEE Trans. on PAMI, Vol.10, pp. 910-918, Nov. 1988.

[2] Chen Yi-Kai. Skew Detection and Reconstruction Based on Maximization ol Variance of Transition- Counts. Pattern Recognition, 2000,33(5) 195-208.

[3] Hong Yan. Skew Correction of Document Images Using Interline Cross-Correlation. CVGIPI Graphical Models and Image Processing, 1993, $$ (6) : 538- 543

[4] S. Baird, "The skew angle of printed documents," Proc. of Society of Photographic Scientists and Engineers, vol. 40, pp. 21-24, 1987.

[5] Nadir Durrani "Typology of word and automatic word segmentation in Urdu Text Corpus", Ph.D. Thesis, at National University of Computer and Merging Sciences Lahore, Pakistan, 2007.

[6] S.N. Srihari, and V. Govindaraju, "Analysis of textual images using the Hough Transform", Machine Vision Application, 2, pp. 141-153, DOI: 10.1007/BF01212455, 1989.

[7] Nandini N., S.Murthy K., G. Hemantha Kumar, "Estimation of Skew Angle in Binary Document Images Using Hough Transfiorm", Journal of World Academy of Science Engineering and Technology, issue 42, pp. 44-49, 2008.

[8] C.M. Naim, South Asia Language & Area Center University of Chicago, "Introductory Urdu – Volume I, National Council for Promotion of Urdu Language, Ministry of HRD Govt. of India".book preface

[9] R. J. Ramteke, Sayyad shafiyoddin, Imran Khan Shaikh,"Urdu Word typology and word segmentation problems" International journal of computer and business intelligence, ISSN 0975-945X, vol 1, No.1, pp.31-37, January 2010.

[10] Rajiv Kapoor, Deepak Bagai, T.S. Kamal, "Skew Angle Detection of a Cursive Handwritten Devanagari Script Character Image", Journal of Indian Institute of Science, 82, pp 161-175, 2002.

[11] Atallah Mahmoud Al-Shatnawi and Khairuddin Omar, "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity", Journal of Computer Science 5 (5), pp. 363-368, 2009

[12] M.Aradhya V N, Hemantha Kumar G. and Shivakumara P, "Skew Estimation for Binary Document Images based on Thinning and

Dr. R. J. Ramteke, 1972, Associate Professor, North Maharashtra University, Jalgaon, India has received his M.Sc. and Ph.D. (Computer Science) in 1995 and 2007 respectively from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. His areas of interest are pattern recognition, Image Processing, Document Analysis, Soft computing, etc.

He is recipient of UGC Teacher Fellowship during 2004-2006. He has published more than 50 research article in National/International journal, conferences. He has participated in various academic courses, International & National Conferences/Seminars and worked as a Session Chair, Program Committee member, Member of organizing committee, Resource Person.
Dr. Ramteke was student member of IEEE for 2005-06, 2006-07, (No.80239919). He is member of International Association of Engineers (MIAENG) (No. 100424) and Senior Member International Association of Computer Science and Information Technology (IACSIT) (No. 80331580)

Imran Khan Pathan, 1981, Assistant Professor, Miliya College of Arts, Science and Management Science, Beed, (Maharashtra) India. He has received his Graduate and M.C.A. in 2002 and 2005 respectively from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. His areas of interest are Image Processing, Document Analysis, etc.
Currently, He is pursuing Ph.D. degree in North Maharashtra University, Jalgaon, India.

Dr. S. C. Mehrotra, 1951, F.N.A, Sc., FIETE has received M.Sc. (Physics) from Allahabad University in 1970 and PhD from University of Texas (Austin) in 1975. He is recipient of UGC award, Which Foundation fellow 1972 (U.S.A.), Dutch Fellowship 1980 (Netherlands), NASA Fellowship 1981 (New York), Alexander Von Fellowship 1983 (W. Germany) Etc. Presently working as a Professor & Head in Department of Computer Science and Information Technology. Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, (India)