*International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)*
*Volume-3 Number-2 Issue-10 June-2013*

# Angular Skew Correction Algorithm for Handwritten Hindi Text

**Rohit Sharma[1], Utkarsh Mathur[2], Naveen Srivastava[3]**

## Abstract

***In large scale document digitalization of hand written and printed documents, they are scanned and stored in digital form. Since many of these documents are hand written they have errors like angular skewness of words. Skew detection and removal is a part of pre-processing before using OCR software to digitalize the documents. One type of skewness that is most difficult to detect and correct is angular skewness that exist in each of the hand written words. In this paper we propose and test algorithms that can be used for angular skew detection and correction of hand written Hindi text.***

## Keywords

***Document pre-processing, skew correction, skew detection, angular skew.***

## 1. Introduction

In a country like India where a number of states use Hindi as their official language, a large number of documents are written in Hindi. According to the 2001 Indian census [1], 258 million people in India reported their native language to be "Hindi". In modern world, digitalization of these documents is not only needed but necessary. A lot of government and court records that are written in Hindi can be made available to millions of people if they are successfully digitalised. However, many problems still exist in large scale digitalization of documents, one such problem is angular skew that is present in hand written text. We propose and test some algorithms for angular skew detection and correction of Hindi text. Angular skew detection and correction is a part of pre-processing of documents before they can be used by OCR tool for digitalization. Angular skew correction is raised as a problem not yet solved [2]. To the best of our knowledge no algorithm exists for angular skew detection and correction of Hindi text.

**Rohit Sharma**, 4th Year Undergraduate Student, Department of CSE & IT, JIIT, Noida, India.
**Utkarsh Mathur**, 2nd Year Undergraduate Student, Department of ECE, JIIT, Noida, India.
**Naveen Srivastava**, 3rd Year Undergraduate Student, Department of ECE, JIIT, Noida, India.

## 2. Properties of Hindi

Hindi is written in Devnagari script. It is noted that many characters of these alphabets have a horizontal line in Devnagari called “sirorekha”. Here we call it as “head-line”. When two or more characters sit side by side to form a word in the language, the head-line portions touch one another and generate a long head-line.

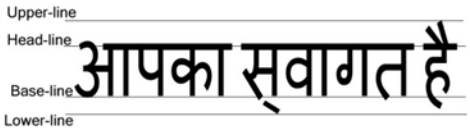



**Fig.1: Devnagari text lines**.

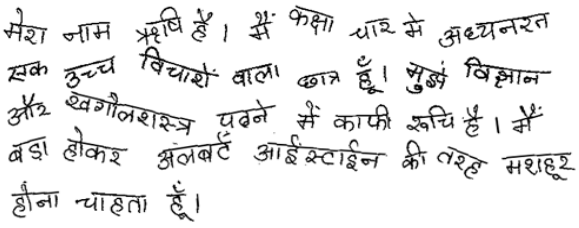
In most Indian languages, a text line may be partitioned into three zones. The upper-zone denotes the portion above the head-line, the middle zone covers the portion of basic (and compound) characters below head-line and the lower zone is the portion below base-line.

Many skew detection and correction algorithms exist which uses Hough transform [3, 4, 5, 6, 7, 8], Fourier transformation [11], Projection Profile [12, 13] methods for skew detection and correction but all of these methods are used for local skew correction. In Hough transform [3, 4], the points coordinate system are described as a sum of sinusoidal distribution: $p = x\cos\theta + y\sin\theta$, the skew angle is calculated on the basis that at the skew angle the density of Transform spaces is maximum. Horizontal projection profile [12, 13] is a histogram of the number of dark pixels in horizontal lines of the given text. Fourier Transformation [11] works on the principle that skew angle is the one at which concentration of spectrum is highest for the document. In Clustering [14], the skew angle for all the words in the document is found and a histogram is made for the determined skew angle. The maximum clustered skew angle in histogram is the skew angle of the document. In another method [9], the centers of the nearest neighbors of the adjoining words in the document are vectored and are linked to determine the skew angle. There have been quite a number of successes in determination and correction of small skew angles of large text fields. However, no work has been done in

angular skew detection and correction of Hindi hand written text. We try to use existing methods and develop new ways for detection and correction of angular skew of handwritten Hindi text using some properties of Hindi language.

## 3. Word Segmentation

The first step towards angular skew removal is identification of individual words from a given text. Only successful segmentation of individual words can ensure that the skew removal is performed accurately. To separate individual words we first pass the entire image through a canny edge detector [15] to make a logical image for further processing. Purpose of this filter is to reduce noise and recovery of information in extreme lighting conditions where normal thresholding fails. We then remove holes and small area noise. The image is dilated to obtain clusters that could be declared as a word. Then we use these clusters as base values, to compute bounding box [16] of each word and the words thus identified are separated for skew correction. This way we ensure that the words are separated.



**Original Image**

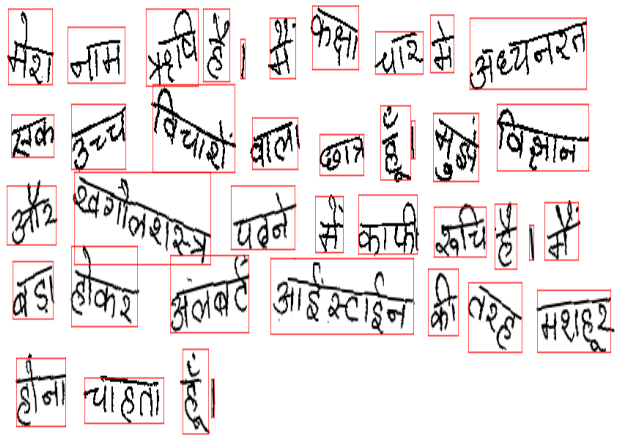


**Fig 2: Word detection**

## 4. Angular Skew Correction

Once we have detected the words from a given text, we can run our algorithm on the individual words to remove angular skewness of words.

In our method we use Hough transform to find straight lines in the words. In Hindi the head-line is the longest straight line, we use this property of Hindi language and select the longest straight line we obtained by using Hough's transform. We calculate the angle of inclination of the line with the x-axis. This gives the skew angle of the word, rotating the word by the skew angle removes the angular skew of the word.

**Problems encountered:**
1) Hindi punctuation marks like ("|" Full Stop), (";" Semicolon), ("?" Question Mark) and ("!" Exclamation Mark) give a vertical straight line by Hough transform. This line does not give us the skew angle.

**Proposed Solution:**
We make a logical assumption that hand written text cannot have skew more than 60˚ so if skew angle θ is greater than 60 (as angle for punctuation marks is near 90˚) replace θ by θ – 90 and then perform the rotation operation.

Algorithm to remove angular skewness by finding Hough lines:
**Step1:** Process the text to extract individual words.

**Step 2:** Use Hough line Transformation to find the head-line of the words selected. The longest line is selected.

**Step3:** Using Hough line Transformation as base values, computed the angle of inclination θ of these lines.

**Step4:** If θ>60 replace θ by θ – 90 (This step is performed to avoid orientation changing of Hindi punctuation marks like ("|" Full Stop), (";" Semicolon), ("?" Question Mark) and ("!" Exclamation Mark).

**Step5:** If θ<60 Rotate each word such that each word such by an angle equal θ such that θ = tan¬-1 (slope of Hough line).

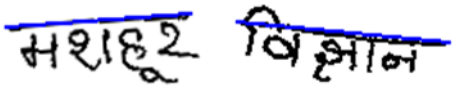


**Fig 3: Detected headlines from Hough lines**

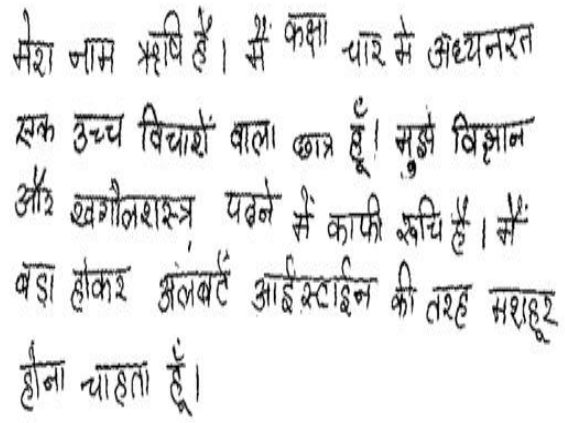


**Fig 4: Text after angular skew correction Hough lines method.**

## 5. Local Skew Correction

After rotation of words using our algorithms, the relative position of words may change, causing induction of linear skew in text. To correct the alignment of words, the mean of centres of Hough lines and regression lines which lie within a tolerance of 10% from first word is computed. The centres are then shifted such that they lie on the mean line thus normalizing the centres of every word. The lines now will be nearly straight and local skew is removed.

## 6. Experimental Results

We tested our algorithm on scanned texts for angular skew correction for around 1200 words. The angular skew was detected and corrected angular skew with an accuracy of 94.91%. The error was detected mainly for small words, were correct Hough line was not detected.

## 7. Conclusion

To sum up briefly the paper exposes a successful method to remove angular skewness in hand written Hindi text. This method was made mostly for Hindi language as it is deeply dependent on presence of head line for skew detection but can be used for all scripts with a headline for example: Bangla and Gurumukhi. This method solves a major problem of angular skew in optical character recognition of scanned Hindi hand-written documents.

## References

[1] Office of the Registrar General & Census Commissioner, India (2001). [Online].Available:http://censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement1.html.

[2] Bhupendra Kumar, Aniket Bera and Tushar Patnaik "Line Based Robust Script Identification for Indian Languages". International Journal of Information and Electronics Engineering, Vol. 2, No. 2, March 2012.

[3] Srihari, S.N. and V. Govindaraju. "Analysis of textual images using the Hough transforms". Machine Vision Applications, 2: 141-153. DOI: 10.1007/BF0121245, 5, March 2012.

[4] Duda, Richard O., and Peter E. Hart. "Use of the Hough transformation to detect lines and curves in pictures." Communications of the ACM 15, no. 1 (1972): 11-15.

[5] Le, D.S., G.R. Thoma and H. Wechsler. Automatic page orientation and skew angle detection for binary document images. Pattern Recognition, 1994.

[6] Pal, U. and B.B. Chaudhuri, 1996. An improved document skew angle estimation technique. Pattern Recognition Lett. , 17: 899-904. DOI: 10.1016/0167-8655(96)00042-6.

[7] Yu, B. and A.K. Jain, 1996. A robust and fast skew detection algorithm for generic documents. Patt. Recog., 29: 1599-1629. DOI: 10.1016/0031-3203(96)00020-9.

[8] Jipeng, Tian, G. Hemantha Kumar, and H. K. Chethan. "Skew correction for Chinese character using Hough transform." International Journal of Advanced Computer Science and Applications-IJACSA,(Special Issue) (2011): 45-48.

[9] O'Gorman, L., 1993. The document spectrum for page layout analysis. IEEE Trans. Patt. Anal. Mach. Intell., 11: 1162-1173. DOI: 10.1109/34.244677.

[10] A.F.R. Rahman and M. Kaykobad, A Complete Bengali OCR: A Novel Hybrid Approach to Handwritten Bengali Character Recognition, Journal of Computing and Information Technology, Vol. 6(4), 1998, pp. 395-413.

[11] Omar, K., A. Ramli, R. Mahmod and M. Sulaiman, 2002. Skew detection and correction of jawi images using gradient direction. Journal of Tech., 37: 117-126.

[12] Akiyama, T. and N. Hagita, 1990. Automated entry system for printed documents. Pattern Recognition, 23: 1141-1158. DOI: 10.1016/0031-3203(90)90112-X.
[13] Chaudhuri, Bidyut B. Digital document processing. Springer-Verlag London Limited, 2007.
[14] Hashizume, Akihide, Pen-Shu Yeh, and Azriel Rosenfeld. "A method of detecting the orientation of aligned components." Pattern Recognition Letters 4, no. 2 (1986): 125-132.
[15] J. Canny, &ldquo,A Computational Approach to Edge Detection,&rdquo, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, pp. 679-698, 1986.
[16] J. Ha, IT. Phillips and R.M. Haralick, "Document Page Decomposition using Bounding Boxes of Connected Components of Black Pixels," ISL report, Dept. Electrical Eng., University of Washington, 1994 Delivering IT Services as Computing Utilities", Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications. Vol.00, pp, 5-13, 2008.

**Rohit Sharma** is a 4th year graduate student of department of computer science and engineering from Jaypee Institute of Information Technology. He is a developer and an avid programmer. His interest fields are in Computer Vision, Machine Learning, Artificial intelligence and image processing.

**Utkarsh Mathur** is a 2nd year graduate student of the Department of Electronics and Communication from Jaypee Institute of Information Technology. He is a developer and an avid programmer. His interest fields are Embedded Systems, Image Processing, Audio Processing, Web Development and Shell Programming.

**Naveen Srivastava** is a 3rd year student of the Department of Electronics and Communication from Jaypee Institute of Information Technology, Noida, India. His interest fields are Embedded Systems, Image Processing and, Newtonian and Quantum Physics.