Multiple Skew Estimation In Multilingual Handwritten Documents

D. S. Guru[1], M. Ravikumar[1]and S. Manjunath[2]

**[1]Department of Studies in Computer Science, University of Mysore,
Mysore-570016, Karnataka, India**
**[2]PG Department of Computer Science, JSS College of Arts Commerce and Science,
Mysore-570025, Karnataka, India**

Abstract

In Indian scenario, most of the time the office documents such as forwarded notices and other documents are multilingual with multiple skews. This poses new challenges in the field of document image analysis. In this direction we are presenting a method of estimating the skew in handwritten multilingual documents. From a handwritten multilingual document image each word is segmented using morphological operations and connected component analysis. Skew of each word is estimated by fitting a minimum circumscribing ellipse. The orientation of each word is estimated and then words are clustered using adaptive k-means clustering to identify the multiple blocks present in the document and average orientation of each block is estimated. In order to corroborate the efficacy of the proposed model experimentation on our own dataset is carried out.

Keywords: *Multilingual Handwritten Document; Multiple Skew; Connected Component Analysis; Adaptive Clustering*

1. Introduction

There is a growing trend to share and exchange information electronically. The need to convert existing paper documents into electronic ones for better archival, retrieval and maintenance is therefore growing. Much knowledge is acquired from documents such as technical reports, government files, news papers, books, journals, magazines, letters, bank cheques, to name a few. The acquisition of knowledge from such documents by an information system can involve an extensive amount of handcrafting, which is generally time consuming and can severely limit the application of information systems. Thus, automatic knowledge acquisition from documents has become an important subject [1]. Conversion of paper documents into electronic forms essentially requires scanning and digitization. Many of the existing approaches of skew detection can only process pure printed document images successfully. But it is a challenging problem to process handwritten documents [2]. One of the most challenging tasks in analyzing handwritten documents is to tackle the inherent skew that is introduced due to writer's handwriting, segment the handwritten lines and estimate the skew angle and its direction [3]. Even though some works are reported in the literature on skew estimation of handwritten documents they are monolingual and generally with a single skew. In country like India with multi languages, officers generally write forwarded notes or observations in different languages with different orientations. This imposes a greater challenge in estimation of multiple skews in a document with multi linguistic. To the best of our knowledge, no work has been reported towards estimation of multiple skews in handwritten multilingual document images. This poses new challenges in the field of document image analysis and has motivated us to take up this research work.

In this work, we consider multilingual handwritten documents with multiple skews. We designed an algorithmic model by the use of morphological operations and clustering of data points. Initially different words are segmented out and then slope of each word is calculated. Words are later clustered based on their likelihood with respect to their spatial coordinates and slopes. The blocks or paragraphs belonging to each cluster are identified, the skew of which is estimated to be the average of the slope of the words belonging to that cluster. It is interesting to note that the proposed method is independent of the scripts and also the writers.

The paper is organized as follows; Section 2 presents an overview of works related to skew estimation in printed and handwritten documents. The model proposed to estimate multiple skews in multilingual document is presented in Section 3. Experimental results on our own data set are presented in the Section 4 and the conclusions are drawn in Section 5

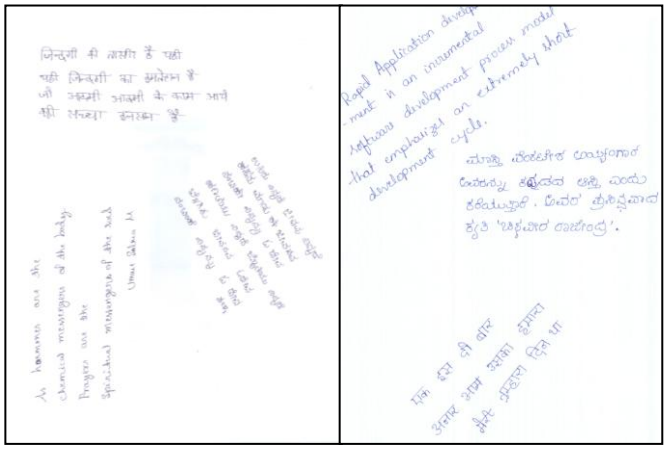
2. Related work

There have been several techniques proposed for accurate estimation of skew angle in document images. The proposed approaches are based on Hough transform [17][18][16][15][9] projection profiles [20][7][19], connected component analysis [22] [28][23][24], cross correlation [25][26], Gradient analysis [4], Fourier spectrum [5], Morphological transforms [6], connected component blotching and linear regression [7], linear regression analysis[8], piecewise covering by

parallelograms [9], fuzzy run length [10], averaged block directional spectrum [11], static and dynamic thresholds [12], minimum area bounding rectangle [13] and Cohen's Class Distributions [14].

From the literature survey, we found that most of the existing methods are to estimate skew in printed documents with text as well as non-text. Though they perform well on printed documents, they cannot be employed for handwritten documents. Only few works [29][30] are reported in the literature to estimate skew in handwritten documents that too written in a single script.

The most challenging task involved in the handwritten documents with multiple skew is identifying different blocks written with different skews and correcting it. Figure 1 shows sample multilingual documents containing multiple skew. To the best of our knowledge there is no work found in the literature which can estimate skew of a multilingual document containing multiple skews and correcting it. Hence in this work we have taken up this work as the starting step in handwritten document analysis and recognition.



(a) (b)

Figure 1. Samples of multilingual handwritten documents with multiple skew

3. Proposed Model

The method works on the entire handwritten multilingual document image with multiple skews. Segmenting the handwritten document is necessary in order to identify the different blocks written in different orientations. After segmenting the blocks it would be easy to estimate the multiple skewed blocks.

To segment the blocks we carry out morphological operations (dilation followed by erosion) on the document in order to merge the characters of each word as a single component. Based on the connected component analysis, we segment each word present in the handwritten documents. The reason for identifying each word in the document is that the skew of a block depends on the skew of each word present it. In order to estimate the skew angle of each word we fit a minimum circumscribing ellipse by treating pixels of each word as data points and the slope of the major axis of the ellipse will be considered as the skew angle of the word. To select the pixels belonging to a word, we use the boundary of the connected component (word) and the pixels falling inside the boundary will be treated as data points to fit the ellipse and the angle of major axis will be the skew angle and length of the major axis will be the length of the word [21]. Instead of using directly the angle of major axis we recommend to discretize the angle into bins where each bin will span with a range of α to avoid overhead of calculations in grouping the words into blocks.

To identify the blocks of text written in different orientations we use the orientation of each word along with the coordinates of the end points of the major axis of each word. This is essential as there may be multiple blocks in the single document with different / same orientations at different locations of the document. Hence, the spatial coordinates of end points of the major axis help us in preserving the spatial location of the words. To cluster the data we suggest to store the obtained angle α and coordinates (x_1, y_1, x_2, y_2) of the end points of each axis in a two dimensional matrix where each row corresponds to a word as shown in Table 1.

Table 1. Data Structure used to preserve the orientation of words

	X_1	Y_1	X_2	Y_2	Bin Angle
W_1	X_{11}	Y_{11}	X_{12}	Y_{12}	α 1
W_2	X_{21}	Y_{21}	X_{22}	Y_{22}	α 2
W_3	X_{31}	Y_{31}	X_{32}	Y_{32}	α 3
.					
Wn	X_{n1}	Y_{n1}	X_{n2}	Y_{n2}	α n

In this work we recommend discritization of the skew angle α instead of having it in real domain. Now we suggest clustering of words based on their spatial coordinates and skew angle. Normally K-means clustering technique is used in the literature to cluster the data points. However it is quite difficult to fix up the appropriate value for the parameter k (number of clusters). This essentially requires supervised knowledge which is generally not available in a complete automation process. Hence, in this

work we recommend to use the adaptive k-means clustering [27] algorithm. This algorithm is used as there is no over head of identifying the number of clusters. The adaptive k-means clustering is based on spectral based clustering which automatically identifies the number of clusters. After clustering, all the words belonging to each cluster are together treated as a single block. The orientation of each block is estimated by calculating the average orientation of each word present in that block. The obtained average orientation is treated as the skew angle of that block. The block diagram of the proposed model is presented in Figure 2.

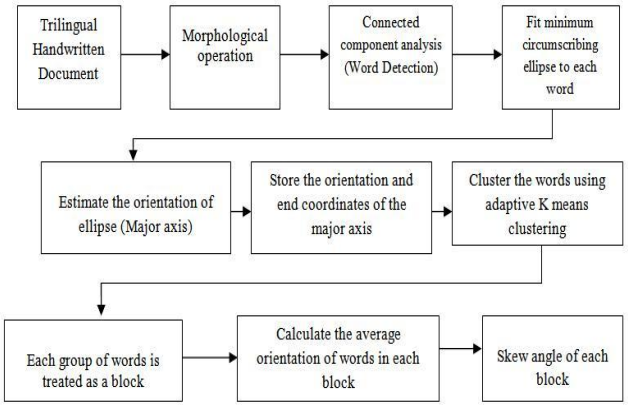


Figure 2. Block diagram of the proposed model

4. Experimentation

In order to conduct experimentation we have created our own data set containing tri lingual documents. We have created 114 handwritten documents using three languages: Kannada, English and Hindi. Also documents containing arbitrary words of their interest in different orientations were also created. The contents are written in an unconstrained manner. Table 2 shows the average number of paragraphs, lines and words containing in the entire data set and the Fig 3 shows the results obtained from the proposed model.

Table 2. Average number of paragraphs, lines and words in the data set considered for experimentation

Average number of	Kannada	Hindi	English
Paragraphs	2.35	1.95	2.25
Lines	5.2	4.2	4.05
Words	16.25	15.35	18.5

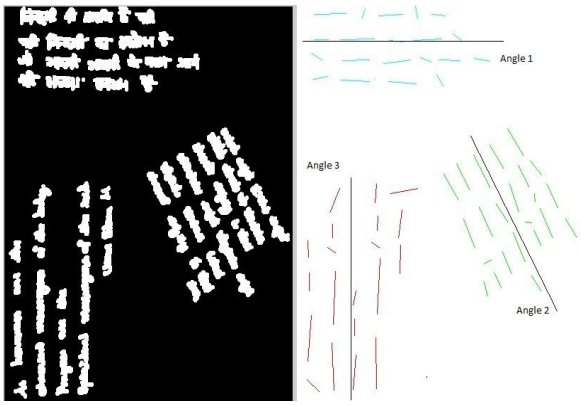


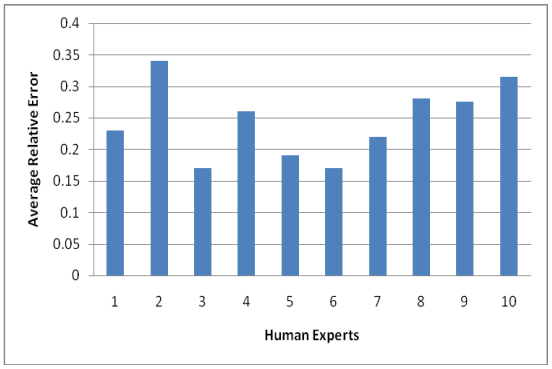
Figure 3. A sample result obtained by the proposed model

To evaluate the performance of the proposed model the skew of each block is estimated manually by drawing a line on each paragraph in the document and the orientation of each line is stored. The stored orientation of the line of each paragraph is compared with the skew obtained by the proposed model. Further to corroborate the efficacy of the proposed model ten human experts are asked to obtain the orientation of each block present in the document and the same has been used to compare the results obtained by the proposed method. Figure 4 shows average relative error in estimating the skew in multilingual handwritten documents with respect to human experts and results obtained by the proposed model on our data set

$$Average\ \mathrm{Re}\,lative\ Error = \left| \frac{\alpha_a - \alpha_o}{\alpha_a} \right| \qquad (1)$$

Where α_a = Actual skew and α_o = Obtained skew

Figure 4. Average relative error rate of the proposed method with respect to human experts.



In order to compare our proposed method with other state of the art techniques we have made a qualitative comparative analysis. For comparison we have considered several factors such as whether the model can handle documents written in multilingual with multiple skews. Also we have considered whether the model can work only on printed, handwritten or a document containing both handwritten as well as printed. These factors are specifically considered to show the superiority of the proposed model. Table 3 contains qualitative comparative analysis of the proposed model with that of state of the art techniques. From Table 3 it is observed that none of the state of the art techniques work on documents written in multilingual with multiple skews and also documents containing both handwritten and printed text.

Table 3. Qualitative comparative analysis of the proposed model with other state of the art techniques

	Multiple Skews	Multilingual	Handwritten	Printed
Kavallieratou et al., (2002)	NO	NO	**YES**	**YES**
Gatos et al., (1997)	NO	NO	NO	**YES**
Amin and Wu (2005)	NO	NO	**YES**	**YES**
Lu and Tan (2003)	NO	NO	**YES**	**YES**
Dey and Noushath (2010)	NO	NO	NO	**YES**
Proposed Model	**YES**	**YES**	**YES**	**YES**

5. Conclusions

In this paper we have presented a new model to estimate multiple skew of multilingual handwritten document .The proposed model is based on angle of major axis of each word segmented using connected component analysis. As the proposed model is based on morphological operations and connected component analysis, if the lines touch or close to each other then the estimation may not be proper. On the other hand it is observed that the proposed method works better for image blocks containing Kannada and English when compared to blocks written in Hindi. This is because the Hindi words contain elongated in the direction of minor axis. In the future we are going to use available models to identify the skews and apply our model to identify multiple skews.

References

[1]Tang, Y ., , S.,lee , Suen. C., : Automatic Document Processing: A Survey. Pattern recognition. Vol. 29. No. 12, pp 1931 – 1952 (1996)

[2]Babu, D,R., Kumat,P,M., Dhannawat,M,D.,: Skew angle estimation and correction of handwritten, textual and large areas of non- textual document images: A novel approach. International conference on image processing, computer vision and pattern recognition. pp 510-515. (2006)

[3]Kasiviswanathan, H., Ball, G, R., Srihari, S, N.,: Top down analysis of line structure in handwritten documents. 20th International conference on pattern recognition, pp 2025 – 2028 (2010)

[4]Sun, C., Si, D., : Skew and slant correction for document images using gradient direction. Fourth International Conference on Document Analysis and Recognition, pp. 142-146 (1997)

[5]Gorman, O, L.,: The document spectrum for Page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 11, pp. 1162 – 1173. (1993)

[6] Najman, L.,: Using mathematical morphology for document skew estimation. Proceedings of the SPIE, Vol. 5296, pp. 182-191. (2003)

[7]Guru, D, S., Punitha, P., Mahesha, S.,: Skew estimation in digitized documents: A novel approach. Proceedings of the Fourth Indian conference on Computer Vision, Graphics and Image processing, pp. 314 – 319. (2005)

[8]Shivakumara, P., Hemanth, Kumar, G., Guru, D, S., Nagbhushan, P.,: A novel technique for estimation of skew in binary text document images based on linear regression analysis. Sadhana, Vol.30, Part I, pp. 69-85. (2005)

[9]Chou, U, C., Chu, Y, S., Chang, F.,: Estimation of skew angles for scanned documents based on Piecewise Covering by Parallelograms. Pattern Recognition, Vol. 40, issue 2, pp. 443-455 (2007)

[10]Shi, Z., Govindaraju, V.,: Skew detection for complex document images using fuzzy runlength. Proceedings of the seventh International Conference on Document Analysis and Recognition (2003)

[11]Lowther, S., Chandran, V., Sridharan, S.,: An accurate method for skew determination in document images. Digital Image Computing Techniques and Applications, pp. 21-22. (2002)

[12]Shivakumara, P., Hemanth, Kumar. G., Guru, D, S., Nagbhushan, P.,: Skew estimation of binary document images using static and dynamic thresholds useful for

document image mosaicing. Journal of society of statistics, computer and applications, Vol. 1, No. 1, pp – 12 (2003)
[13]Safabakhsh, R., Khadiv, Si.,: Document skew detection using minimum area bounding rectangle. In Proceedings of the international conference on information technology, pp. 253-258. (2000)
[14]Kavallierautou, K., Fakotakis, N., Kokkinakis, G.,: Skew angle estimation in document processing using cohen's class distributions. Pattern Recognition letters, vol. 20, pp. 1305 – 1311 (1999)
[15] Kapoor, R., Bangi, D., Kamal, T, S.,: A new algorithm for skew detection and correction. Pattern Recognition Letters, vol. 25, pp. 1215-1229. (2004)
[16] Das, K., Chanda, B.,: A fast algorithm for skew detection of document images using morphology, International Journal on Document Analysis and Recognition, Vol. 4, pp. 109 – 114 (2001)
[17] Amin, A., Fischer, S.,: Robust skew detection method using the Hough transform, Pattern Analysis and Applications, Vol. 3, pp. 243-253. (2000)
[18]Chen, K, Y., Jwang, F, J.,: Skew detection and reconstruction based on maximization of variance of transition counts, Pattern Recognition, Vol. 33, pp. 195-208. (2000)
[19]Dey, P., Noushath, S.,: e- PCP: A robust skew detection for scanned document images. Pattern Recognition. Vol. 43, pp. 937-948 (2009)
[20]Kwag, K, H., Kim, H, S.,: Jeong, H, S., Lee, S. G.,: Efficient skew estimation and correction algorithm for document images. Image and Vision Computing, Vol. 19, pp. 25-35. (2002)
[21] Messelodi, S., Modena, M, C.,: Automatic identification and skew estimation of text lines in real scene images. Patter Recognition, pp. 791-810. (1999)
[22]Liolios, N., Fakotakis, N., Kokkinakis, G.,: On the generalization of the form identification and skew detection problem. Pattern Recognition, Vol. 35, pp. 253-264. (2002)
[23]Lu, Y., Tan, C, L.,: A nearest- neighbour chain based approach to skew estimation in document images, Pattern Recognition letters, Vol. 24, pp. 2315-2323. (2003)
[24]Amin, A., Wu, S.,: Robust skew detection in mixed text / graphics documents, International conference on document analysis and recognition, Vol.1, pp. 247-251. (2005)
[25]Yan, H.,: Skew correction of document images using interline cross-correlation .Computer Vision Graph. Image Process, 55(6), pp 538-543. (1993)
[26]Gatos, B., Papamarkos, N., Chmzas, C.,: Skew detection and text line position determining in digitized documents, Pattern Recognition, Vol.30, no. 9, pp. 1505 – 1519. (1997)
[27]Sanguinetti G., Laidler J and Lawrence N. D.,: Automatic determination of the number of clusters using spectral algorithms. Proceedings of IEEE Machine Learning for Signal Processing, pp. 28-30. (2005)
[28]Pal U., Sinha S and Chaudhuri B. B ., : Multi-oriented text lines detection and their skew estimation, Third Indian conference on computer vision, Graphics and image processing. (2002)
[29]Kavallieratou K., Fakotakis N., and Kokkinakis G.,: Skew angle estimation for printed and handwritten documents using wigner-ville distribution. Image and Vision Computing, pp. 813 – 824 (2002)
[30]Basu S., Chaudhuri C., Kundu M., Nasipuri M., and Basu D K.,:Text line extraction from multi-skewed handwritten documents, Pattern Recognition , Vol. 40, pp. 1825 – 1839.(2007)

D. S. Guru received his BSc, MSc and PhD degrees in Computer Science and Technology from the University of Mysore, Mysore, India, in 1991, 1993, and 2000, respectively. He is currently an associate professor in the Department of Studies in Computer Science, University of Mysore, India. He was a fellow of BOYSCAT. He was a visiting research scientist at Michigan State University. He is supervising a couple of major projects sponsored by UGC, DST, and Government of India. He is a life member of Indian professional bodies such as CSI, ISTE, and IUPRAI. He has authored 36 research papers for journals and 145 peer-reviewed conference papers at international and national levels. His area of research interest covers image retrieval, object recognition, shape analysis, sign language recognition, biometrics, and symbolic data analysis.

M. Ravikumar obtained his BE from Kuvempu University, Karnataka and M.Tech from Visveswaraya Technological University , Karnataka during 1996 and 2001 respectively. Currently he is pursuing PhD in Document Image Analysis in University of Mysore, Mysore, India. He is also working Assistant Professor, Department of Computer Science. Kuvempu University, Karnataka, India. He has authored a few peer-reviewed papers in journals and conferences. His areas of research cover Document Image Processing and Pattern Recognition.

S. Manjunath obtained his BSc and MS degrees in Computer Science from University of Mysore, Mysore, India, respectively, in the years 2003 and 2006. He obtained PhD in video processing from University of Mysore, Mysore, India in 2013. Currently he is working as Assistant Professor, PG Department of Computer Science, JSS College of Arts, Commerce and Science, Mysore. He has authored a few peer-reviewed papers in journals and conferences. His areas of research cover image and video processing, biometrics and text mining.