Script Based Trilingual Handwritten Word Level Multiple Skew Estimation

M. Ravikumar, D.S. Guru, S. Manjunath and V.N. Manjunath Aradhya

Abstract Skew estimation and correction plays an important role in document analysis. In the present work, we proposed a model to estimate multiple skews present in trilingual such as Devanagari, English, and Kannada handwritten documents at word level with a priori knowledge about the corresponding scripts. The idea of using different skew estimation techniques for different scripts such as Hough transform (HT) for Devanagari words, Gaussian Mixture Models (GMM) and convex hull for Kannada and English words is proposed. The effectiveness of these approaches has been reported by testing on a dataset consisting of 1000 words in each script. Experimental results show that the proposed approaches are effective in estimating and correcting the handwritten skew words.

Keywords Handwritten skew estimation · Skew correction · Gaussian mixture models · Hough transform · Convex hull

M. Ravikumar (✉) · D.S. Guru
Department of Studies in Computer Science, University of Mysore,
Mysore, Karnataka, India
e-mail: ravi2142@yahoo.co.in

D.S. Guru
e-mail: dsg@compsci.uni-mysore.ac.in

S. Manjunath
Department of Computer Science, Central University of Kerala,
Kasaragod, Kerala, India
e-mail: manju_uom@yahoo.co.in

V.N. Manjunath Aradhya
Department of MCA, S.J. College of Engineering, Mysore, Karnataka, India
e-mail: aradhya.mysore@gmail.com

S.C. Satapathy et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 434,
DOI 10.1007/978-81-322-2752-6_53

1 Introduction

India is a multilingual country, written communication in government sectors documents and forwarded notices generally contain both printed as well as handwritten texts in multiple scripts. The technology is being used to set up a paperless office, in this regard a huge amount of documents need to be digitized. In order to digitize a document, it is very much essential to develop a multi script OCR for recognition of all the scripts present in the document. Research community has tackled the problem with respect to the design of OCR for printed documents. However, a greater attention has to be given for design and development of multilingual OCR for handwritten documents. Analysis of multilingual and multiple skewed handwritten documents pose two important challenges in development of multilingual OCR, i.e., identification of scripts and its estimation and correction of skew angle. Normally, to the given document, separating printed text from the handwritten text followed by word segmentation, skew correction and later script identification tasks are to be performed. In case of document containing words with multiple scripts having multiple skew, order of using skew correction and script identification can be tackled by using any of the two following ways. First and foremost task is that the script can be identified and script based skew estimation techniques can be employed. On the other hand, skew correction followed by script identification. In the later case skew estimation leads to a problem of selection of appropriate skew estimation technique as it depends on the script. Hence, the first case of performing script identification followed by skew estimation is suitable order for such application of developing OCR for a multilingual handwritten document containing multiple skews.

Script identification can be performed at three levels i.e., script identification at block level, textline level and word level. Specifically, for multilingual documents containing multiple skews, word level script recognition is recommended [1, 2]. If the script of a word is known, then selecting a particular algorithm for estimating the skew angle will be an easier task. From the literature, it is learnt that the efficiency of an OCR system mainly depends on the effectiveness of skew estimation and correction. Hence skew estimation in general and at word level in particular plays a vital role in the field of handwritten document analysis. In this paper, we explore the approaches to correct the skew of multilingual handwritten document.

The organization of the paper is as follows, in Sect. 2, we present a brief review only on handwritten skew estimation techniques. In Sect. 3, we present the description the proposed model and experimental analysis is presented in Sect. 4. Finally, conclusion is presented in Sect. 5.

2 Related Work

Analysis of any handwritten document image requires that there shall not be any skew. Producing a document without any skew seems to be inevitable. Hence Skew estimation and correction are required before the actual document image analysis.

Inaccurate de-skew will significantly deteriorate the subsequent processing stages in document analysis and may lead to incorrect layout analysis, erroneous word or character segmentation, and hence results in misclassification. The overall performance of a document analysis system will subsequently decrease due to the presence of a skew [3, 4].

On the other hand in real time applications, a document may contain multiple skews because, there may be different number of components annotated using different scripts in different direction. Hence an effective algorithm is required for multiple skew detection of multilingual document, which makes OCR system capable of processing. Already a good number of skew estimation algorithms are available in the literature and most of the existing algorithms will work for single language. Hence, estimation of the skew angle for multilingual handwritten documents with multiple skews is still a challenging issue.

Most of the existing skew estimation techniques can be broadly divided into the following categories according to the basic approach they adopt: Projection Profile Analysis, Hough transform, Nearest neighbors clustering, Cross-correlation [5].

A skew angle estimation approach based on the application of a fuzzy directional run length is proposed for complex address images which use a new concept of fuzzy run length, which imitates an extended run length [6]. Skew angle detection of a cursive handwritten Devanagari script at word level is proposed in [7] where each word is fit into a standard frame work and Wigner-Ville distribution is applied to each rotation of a word ranging from −89° to +89° and the skew angle is selected as maximum angle of distribution. In [8], a method for skew detection and correction of entire document using HT and contour detection is proposed.

A method based on mixture models in which expectation maximization algorithm is used for estimation of skew angle in unconstrained Kannada handwritten documents is discussed in [9]. In this work, gaps between the characters are filled using a morphological operation like dilation. Textlines and words are extracted using component extension method and vertical projection profile techniques respectively. Further these words are passed to the GMM to extract the mean vector points. With these mean vector points skew angle is estimated. Some more works related to handwritten skew estimation are reported in the literature [10–12].

Multiple skew estimation in multilingual handwritten documents is discussed in [13]. Each word in a document 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. Estimation and correction of skew angle at word level for Devanagari script is discussed in [14], in which HT is applied on words for skew estimation and rotation transform is used for correction of skew angle.

The methods discussed above concentrates only on skew estimation of handwritten documents containing words written in one script and single skew. In the proposed work, we concentrate mainly on skew estimation of a multilingual document containing multiple skew present in a forwarded official documents comprising three scripts: Devnagari, English and Kannada.

3 Proposed Methodology

The block diagram of the proposed system is shown in Fig. 1. A trilingual document containing printed and handwritten text is subjected to the system. Using connected component analysis the words are segmented. After word segmentation, the words are classified either as printed or handwritten. In order to carry out this we use the work carried out in [15]. Statistical texture features of words such as mean, standard deviation, smoothness, moment, uniformity, entropy and local range including local entropy is extracted. The K-nearest neighbor classifier has been used to classify the words into printed or handwritten.

Normally skew estimation technique depends on the script specifically in case of handwritten. Indeed the knowledge about the script helps us in identifying the skew more accurately. For example, if the document contains Devanagari script we can directly use skew estimation model which works based on line estimation techniques as Devanagari script has Shirorekha feature. In case of document containing Kannada/English then we can use either the application of Gaussian Mixture Models (GMM) or convex hull to estimate the same. Hence, prior to skew detection we apply script identification at word level and we use the work of script identification described in [1]. Once the script of the word is recognized further three different skew estimation techniques are carried out, which are as follows:

1. Hough transform (HT)
2. Gaussian Mixture Models (GMM)
3. Convex hull

From the above three approaches, we estimate candidate points for skew estimation. Once the candidate points are obtained, we use linear regression based line fitting and moments for skew estimation. The brief description about each technique is presented in the following subsections.

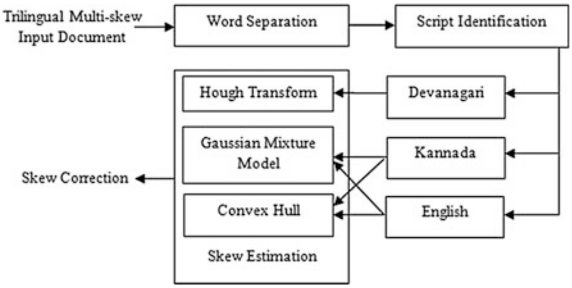


Fig. 1 Block diagram of the proposed model

3.1 Hough Transform (HT)

Hough transform is a technique to find lines, circles or any other parametric curves. In the present approach, we use Hough transform to estimate straight line present in the word image and use the same straight line to estimate the skew of the word. Each point in the image plane is transformed to a sinusoidal curve in the parametric space, i.e., $\rho = x \cos \theta + y \sin \theta$ where ρ is the distance of line from the origin, and θ is the angle of ρ with respect to x axis. The parametric space is divided into a number of accumulator cells. Straight line in the image plane will be converted into a number of sinusoidal curve and each sinusoidal curve will crossover at a particular point in the parametric space forming a peak. Once peak points are detected we can find end points of line segments corresponding to peak values. Shirorekha is a longest line present in a Devanagari word. Using Hough transform, we estimate the longest line in the Devanagari word. The problem with direct use of Hough transform as mentioned in case of [14] for our application is that the line will be partitioned and multiple lines will be identified. The estimated lines are treated as candidate lines and in order to identify true Shirorekha we use technique of boundary growing of candidate lines in the direction of the angle obtained by the line. After boundary growing the line which forms longest line will be treated as Shirorekha line. The end points of Shirorekha are candidate points for skew estimation.

3.2 Gaussian Mixture Models (GMM)

In order to provide good approximation to multimodal distributions, it is failed due to its intrinsic unimodal property of Gaussian distributions and also very limited in representing adequate number of distributions. Latent variables also called as hidden or unobserved variables can be used in mixture distribution which allows us to solve the aforesaid issues. Discrete latent variables can also be interpreted as assigning data points towards specific components of mixtures. Expectation Maximization (EM) algorithm is one of the techniques for finding maximum likelihood estimation in latent variable [16]. During learning process, we discover mean centers μ_k and variances Σ of the Gaussian components associated with mixing coefficients which are employed as regulating parameters. Gaussian mixtures used to analyze complex probability distribution and are formulated in terms of discrete latent variables and further defined as the weighted sum of '*K*' Gaussian components.

In the present work, the obtained means of *k* clusters $\mu_k, \forall k = 1, \ldots, K$ is then used for estimating the skew of a word. Selecting K value is a highly subjective in nature. Therefore, we fix the value of *K* equal to number of connected components in the word image. The centroid of each cluster is treated as candidate points to estimate the skew angle of the word. Detail information regarding GMM can be seen in [9].

3.3 Convex Hull

As words are presented with few connected components, the pixels of connected components are treated as points in two dimensional spaces. A convex polygon is fit to the set of word pixels such that all points of the connected component lie in a polygon. The centroids of each convex hull of the connected components are treated as candidate points for skew estimation [17].

3.4 Skew Estimation

In case of HT based approach, it is recommended to apply only to the script of Devanagari words as Shirorekha. The end points of the line are considered and angle obtained by the end points of the line is used to estimate the skew angle. In case of Kannada and English we recommend either to use convex hull based approach or GMM. Once the candidate points are estimated we use two different techniques to estimate the skew viz., linear regression method and second order moments based approach. In case of linear regression based approach for the candidate points we fit straight line using linear regression and the skew angle of the estimated line is treated as skew angle of the word image. In case of moments based approach, we estimate the skew angle of the candidates based on the statistical properties of the points as mentioned in [14].

4 Experimentation

In order to carry out the experiment, we have created a handwritten word database of trilingual (Devanagari, English and Kannada) scripts. Each script is treated as a class and the each class contains 1000 words. Handwritten documents were collected using flatbed HP scanner at 300 resolution dpi and preprocessing techniques such as binarization and noise removal are also carried out. The words are segmented and used for the experimentation. Sample words of each class are given in Table 1. Experiment was conducted on a desktop computer using Matlab (version 2013), with Windows operating system having 2 GB RAM capacity.

To evaluate the performance of the proposed model, the skew of each word is estimated manually by drawing a line on each word and the orientation of each line is stored. The stored orientation of the line of each word is compared using average relative error obtained as per Eq. 1, with the skew obtained by the proposed model. Class-wise average relative error is reported in Table 2.

$$Average\,Relative\,Error = mean\left(\left|\frac{\theta_{acutal} - \theta_{obtained}}{\theta_{actual}}\right|\right) \tag{1}$$

Table 1 Sample images of trilingual words

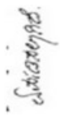
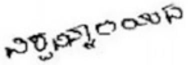
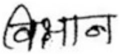
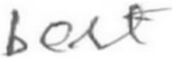
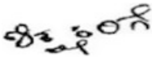
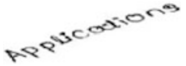
Script	Kannada	Devanagari	English
Sample images			
			
			

Table 2 Average relative error of the proposed model for different scripts using different methods

Class	Method	Average relative error	
		Linear regression analysis	Second order moments
Devanagari	Hough transform	0.35	0.25
Kannada	Gaussian mixture	0.48	0.43
	Convex hull	0.40	0.38
English	Gaussian mixture	0.21	0.19
	Convex hull	0.20	0.16

From Table 2, it is clear that the convex hull based approach performs better for both Kannada and English words. The reason for high relative error for Gaussian Mixture based model in case of Kannada words is due to the presence of modifiers. When Gaussian mixtures are obtained based on the connected component analysis, the modifiers will also get formed to a cluster and the centroids will make the slope of the line to deviate from its actual slope. Whereas, in case of English it is observed that the Gaussian Mixture and the convex hull based approaches have resulted almost same relative error because in English script there is such modifiers present.

However, most of the researchers have recommended to use accuracy as a performance measure to compare different skew estimation techniques. Hence, we have calculated the accuracy of the proposed model and are tabulated in Table 3.

Table 3 Average accuracy of the proposed model for different scripts using different methods

Class	Method	Average accuracy	
		Linear regression analysis	Second order moment
Devanagari	Hough transform	90.12	91.25
Kannada	Gaussian mixture	89.36	90.14
	Convex hull	92.15	93.48
English	Gaussian mixture	92.81	95.66
	Convex hull	93.58	95.57

From Table 3, it is clear that for Kannada and English scripts, the convex hull based approach performs better than GMM model. Also it should be noted that the Hough transform based model is best suited for Devanagari words.

5 Conclusion

In this paper, a model to detect multiple skew in a multilingual handwritten document image is proposed. As a prerequisite, the input document is expected to be annotated with the associated scripts in the respective locations of the document image. For this purpose, script identification at word level is initially performed using [1] by extracting individual words from the document through the application of connected component analysis. To suggest a skew estimation technique for a word of particular script, experiments were conducted with different techniques such as Hough transform, convex hull and Gaussian mixture models. Through the experimental analysis it was observed that, Hough transform based approach best suits for the skew estimation in Devanagari scripts. Whereas, convex hull based approach gives better performance for skew estimation of both Kannada and English scripts.

References

1. Ravikumar, M., Manjunath, S., Guru, D.,S.: Analysis and Automation of Handwritten Word Level Script Recognition. Proceedings of IJCAISC, Vol. 369, pp. 213–225, (2015).
2. Manjunath, S., Guru, D.,S., Ravikumar, M.: Handwritten script identification: Fusion based approaches. ACEEE proc. of Int Conf on MPCIT, pp. 216–222, (2013).
3. Kasturi, R., Gorman, O., L., Govindaraju, V.: Document image analysis: a primer, Sadhana, Vol. 22, Part I, pp. 3–22, (2002).
4. Pavankumar, M.,N.,S.,S.,K., Jawahar, V.,V.: Information processing from document images. In information technology: principles and Applications(ED) pp. 522–547, (2004).
5. Lin, C., T., Fan, K, W., Yeh, C., M., Pu, C.,H., Wu, F., Y.: High-Accuracy Skew Estimation of Document Images. Int. J. of Fuzzy Systems, Vol. 8, No. 3, pp. 119–126, (2006).
6. Shi, Z., Govindaraju, V.: Skew detection for complex document images using fuzzy run length. ICDAR, Vol. 2. (2002).
7. Kapoor, R., Bagai, D., Kamal, T.,S.: Skew angle detection of a cursive handwritten Devanagariscript. Journal of Indian institute of science, Vol. 82, pp. 161–175.J, (2002).
8. Ramappa, M., H., Krishnamurthy, S.: Skew Detection, Correction and segmentation of Handwritten Kannada Document. Int. J. of Ad. Sci and Tech. Vol. 48, pp. 71–87, (2012).
9. Manjunath Aradhya, V. N., Naveen C., Niranjan S. K.,: Skew estimation for unconstrained handwritten documents. Proceedings of ICACC. pp. 1542–1548, (2011).
10. Mello, C. A. B., Ángel, S., Cavalcanti, G. D. C.: Multiple Line Skew Estimation of Handwritten images of Documents Based on a Visual Perception. Springer, LNCS 6855, pp. 138–145 (2011).
11. Brodić, D., Milivojević, Z.: Estimation of the Handwritten Text Skew Based on Binary Moments. Radio engineering. Vol. 21(1), pp. 162–169 (2012).

12. Kleber, F., Diem, M., Sablatnig, R.: Robust Skew Estimation of Handwritten and Printed Documents based on Grayvalue Images. 22nd ICPR, IEEE, pp. 3020–3025 (2014).
13. Guru. D., S., Ravikumar, M., Manjunath, S., Multiple Skew Estimation in Multilingual Handwritten Documents. IJCSI, Vol. 10, Issue 5, No. 2, pp. 65– 69, (2013).
14. Tripati, A., Jundale, Ravindra, S., Hegadi.: skew detection and correction of Devanagari script using hough transform, procedia computer science 45, pp. 305– 311, (2015).
15. Mallikarjun Hangarge, K.C. Santosh, Srikanth Doddamani and Rajmohan Pardeshi. Statistical Texture Features based Handwritten and Printed Text Classification in South Indian Documents. Proceedings of ICETECIT, Vol. 1, pp. 215–221, (2012).
16. Christopher, M., Bishop.: Pattern recognition and machine learning. Springer, (2006).
17. Manjunath Aradhya V N, Hemantha Kumar G, Shivakumara P. Skew Estimation Technique for Binary Document Images based on Thinning and Moments. Engineering Letters, 14:1, pp. 127–134, (2007).