Language Independent Skew Estimation Technique Based on Gaussian Mixture Models: A Case Study on South Indian Scripts

V.N. Manjunath Aradhya[1], Ashok Rao[2], and G. Hemantha Kumar[1]

[1] Dept of Studies in Computer Science,University of Mysore,
Mysore - 570 006, India
mukesh_mysore@rediffmail.com
[2] Dept of Electronics and Communication, S.J. College of Engineering
Mysore - India
ashokrao.mys@gmail.com

Abstract. During document scanning, skew is inevitably introduced into the incoming document image. Presence of additional modified characters, which get plugged in as extensions and remain as disjointed protrusions of a main character is really challenging in estimating inclination in skewed documents made up of texts in south Indian languages (Kannada, Telugu, Tamil and Malayalam). In this paper, we present a novel script independent (for south Indian) skew estimation technique based on Gaussian Mixture Models (GMM). The Expectation-Maximization (EM) algorithm is used to learn the mixture of Gaussians. Subsequently the cluster means are subjected to moments to estimate the skew angle. Experiments on printed and handwritten documents corrupted by noise is done. Our method shows significantly improved performance as compared to other existing methods.

1 Introduction

The volume of paper based documents continue to grow at a rapid rate in spite of the use of electronic version. As a result, both the transformation of a paper document to its electronic version, and its subsequent image processing and understanding have become an important application domain in computer vision and pattern recognition research. Document analysis and character recognition are usually performed through several phases: scanning, image enhancement, skew estimation and correction, segmentation, and character classification. The skew estimation of document images is particularly crucial among the document processing operations as it affects the subsequent understanding of the document. Several attempts have been made for skew detection and the methods can be mainly categorized into five groups: Hough transform, Cross Correlation, Projection profile, Fourier transformation and K nearest neighbor (K-NN) clustering.

Hough transform based technique for skew detection is presented in [11]. To reduce the computational burden, the bottom pixels of the candidate objects

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 487–494, 2007.

within a selected region are subjected to Hough transformation [10]. The hierarchical Hough transformation technique was also adapted for skew estimation[16]. The main idea of the these methods is to reduce the amount of input data which in turn reduces their computational complexities. An improved method to overcome the drawback of the method proposed in [10] is presented in [13]. The cross-correlation method proposed in [4] is based on the correlation between two vertical lines in a document image. The horizontal projection profile (HPP), proposed in [9], is a histogram of the number of black pixels along the horizontal lines of a scanned document. The method works based on text line profile peaks and troughs to estimate the skew angle. However the method works only for ideal cases and it is known to be time consuming. To alleviate this, modifications are done to this iterative approach for quick convergence [8]. An approach for skew estimation based on HPP is described in [12]. Here HPP's are calculated for each strip and from the correlation of the profiles of the neighboring strips the skew angle is determined. Although the proposed method is computationally inexpensive, it cannot not work well if the document is skewed beyond $\pm 10^0$. The Fourier transform based algorithm for skew estimation is presented in [14]. According to the method, skew angle of a document image corresponds to the direction where the density of Fourier space becomes the largest. However its computational complexity is very high. Nearest neighbor chain based approach for skew estimation in document images is proposed in [6]. Cao et al [15] proposed skew detection and correction in document images based on straight-line fitting. A skew detection and correction technique using Radon transform projection profile technique is described in [5]. An algorithmic technique that performs skew angle correction to handwritten Bengali text is reported in [1].

Aforementioned methods are script dependent and also they perform poorly if documents contain noise, degraded texts and varying font size of texts. More importantly, these methods may not obtain accurate results for south Indian scripts. This is due to additional modifying characters that remain as disconnected protrusions of a main character, which is one of the dominant feature of south Indian language particularly Kannada and Telugu. Hence, in this paper we present an improved technique of skew estimation for documents containing south Indian scripts. In addition, the proposed technique handles handwritten documents.

The organization of the paper is as follows: Proposed skew estimation technique is presented section 2. In section 3, Experiment and Comparative study are reported. Discussion and conclusion are drawn in section 4.

2 Proposed Methodology

This section presents the proposed methodology that is based on GMM and moments. The proposed method first extracts individual text lines present in the document image using the method described in [7]. This technique is based on boundary growing algorithm, which helps us in extracting text line present in document page. The resultant text line obtained from this algorithm is then passed on to GMM process to extract mean vector points. The resultant mean

vector points are then used to estimate skew angle using moments. To implement this, we first explain the concept of GMM described in [2].

2.1 Gaussian Mixture Models (GMM)

GMM is a simple linear superposition of Gaussian components, aimed at providing a rich class of density than a single Gaussian. The formulation of Gaussian mixtures will provide us with a deeper insight into this important distribution and will serve to motivate the expectation-maximization algorithm. A distribution can be written as a linear superposition[1] of Gaussian in the form:

$$P(x) = \sum_{k=1}^{K} \pi_k \eta(x/\mu_k, \Sigma_k) \tag{1}$$

which is called a *mixture-of-Gaussians*. Where $\eta(x/\mu_k, \Sigma_k)$ is the multivariate Gaussian distribution of the form:

$$\eta(x/\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2} \left|\Sigma\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)} \tag{2}$$

Each Gaussian density $\eta(x/\mu_k, \Sigma_k)$ is called a component of the mixture and has its own mean μ_k and covariance Σ_k. The parameter π_k in Eq.(1) is called mixing coefficient. If we integrate both sides of Eq.(1) w.r.t x, both p(x) and the indidividual Gaussian components are normalized, we obtain $\sum_{k=1}^{K} \pi_k = 1$. Also, the requirement that $p(x) \geq 0$, together with $\eta(x/\mu_k, \Sigma_k) \geq 0$, implies $\pi_k \geq 0\ \forall k$. Combining this with Eq.(1) we obtain $0 \leq \pi_k \leq 1$.

From the sum and product rules, the marginal density is given by

$$p(x) = \sum_{k=1}^{K} p(K) p(x/K) \tag{3}$$

which is equivalent to Eq.(1) in which we can view $\pi_k = p(k)$ as the prior probability of picking the k^{th} component, and the density $\eta(x/\mu_k, \sum_k) = p(x/K)$ as the probability of x conditioned on k.

From Baye's theorem the posterior probabilities $p(K/x)$, which is also known as responsibilities , are given by:

$$\gamma(z_k) \equiv \quad p(K/x) \tag{4}$$

$$= \frac{p(K)p(x/K)}{\sum_l p(l)p(x/l)} \tag{5}$$

$$= \frac{\pi_k \eta(x/\mu_k, \Sigma_k)}{\sum_l \eta(x/\mu_l, \Sigma_l)} \tag{6}$$

The form of the Gaussian mixture distribution is governed by the parameters π,μ and Σ, where we have used the notation $\pi \equiv \pi_1, \pi_2, \ldots, \pi_K$, $\mu \equiv \mu_1, \mu_2, \ldots, \mu_K$ and $\Sigma \equiv \Sigma_1, \Sigma_2, \ldots, \Sigma_K$. We now adapt an iterative algorithm, known as Expectation Maximization (EM) algorithm, to extimate the values of μ,Σ and π.

[1] Please note that mixture of Gaussian need not be a Gaussian.

We first choose some initial values for these parameters by running K-means clustering algorithm. Then we alternate between two steps known as Expectation(E) Step and the Maximization(M) step to update the values of these parameters until convergence criteria is reached. The EM algorithm for GMM[2] can be summarized as follows:

1. Initialize the parameters μ_k, Σ_k amd π_k by running K-means clustering algorithm and evaluate the log of the likelihood function using Eq.(7)
2. **E Step** Evaluate the responsibilities using Eq.(6) with current parameter values.
3. **M Step** Re-estimate the parameters using the current responsibilities:

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, u_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, (x_n - \mu_k^{new})(x_n - \mu_k^{new})^T$$

$$\pi_k^{new} = \frac{N_k}{N}$$

4. Evaluate the log likelihood:

$$ln\ p(X/\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_k \eta(x_n/\mu_k, \Sigma_k) \tag{7}$$

   and check for convergence of log likelihood. If the convergence criterion is not satisfied iterate from step 2.

Using the obtained means of k clusters μ_k, $\forall k = 1, \ldots, K$, first and second order moments are calculated using the Eq. 10. This is used for finding the inclination of the given skewed text line. Figure 1 depicts the mean points obtained for the input skewed document using mixture-of-Gaussians.

2.2 Moments for Skew Estimation

In this section, we present moments based method for the estimation of skew angle. The moments are computed using Eq.(8) and Eq.(9), x and y is the cluster mean points obtained from the GMM, p and q define the order of moments. Angle of each text line present in the document is estimated using Eq.(10). For detailed mathematical derivations, see Ref.[3].

$$m_{pq} = \sum_1^n \sum_1^n x^p\, x^q \tag{8}$$

$$\mu_{pq} = \sum_1^n \sum_1^n (x - \bar{x})^p (y - \bar{y})^q \tag{9}$$

$$\theta = \frac{1}{2} tan^{-1} \left[\frac{2\mu_{11}}{(\mu_{20} - \mu_{02})} \right] \tag{10}$$

where θ is the estimated skew angle of the segmented text line.

[2] Given a GMM, the objective is to maximize the likelihood function with respect to the parameters comprising μ, Σ and π_k.

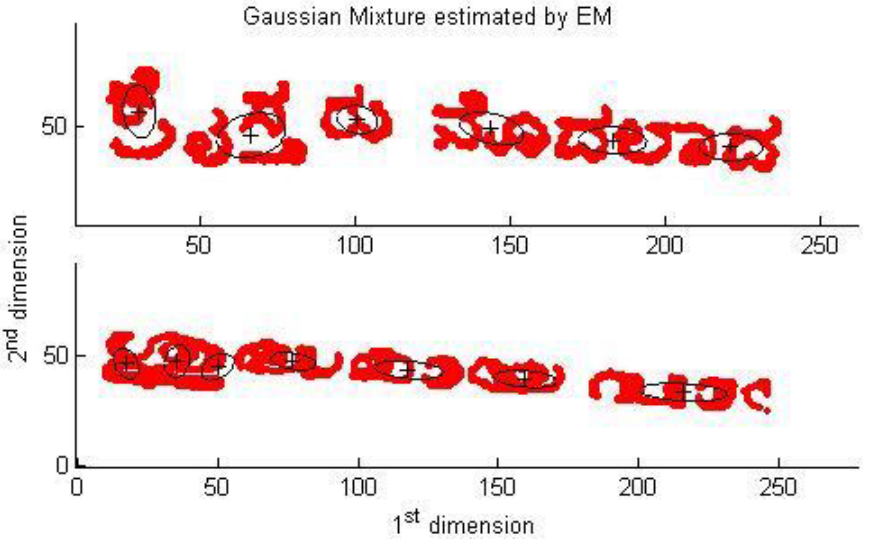

Fig. 1. Illustration of Gaussian Mixture Models for a skewed text line

3 Experimental Results and Comparative Study

This section presents the results of the experiments conducted to study the performance of the proposed method. The method has been implemented in MATLAB 7.0 on a Pentium IV 3.0 GHz with 1GB RAM. For experiment purpose, 20 documents are considered from different sources such as Kannada, Tamil, Telugu, Malayalam, and English. Each document is rotated with four skew angles (3,5,10 and 15). Further to show the superior performance of the proposed method, handwritten English documents and documents with noise are also considered. We have taken two decision parameters such as Mean Skew Angle (M) and Standard Deviation (SD) which are reported in Table 1. In addition, finding optimal number of mixtures to yield best recognition accuracy is highly subjective in nature. Hence, it is empirically fixed to nine mixtures for optimal performance. From Table 1, it is evident that the skew angle obtained by proposed method for English document is better when compared to other south Indian languages.

Table 1. Mean and Standard Deviation obtained by the proposed method

	Kannada		Telugu		Tamil		Malayalam		English	
True Angle	M	SD	M	SD	M	SD	M	SD	M	SD
3	3.12	0.432	2.5	0.3	2.86	0.5	2.86	0.58	3.16	0.18
5	5.05	0.436	5.06	0.46	5.20	0.64	5.80	0.62	5.10	0.28
10	10.5	0.364	9.75	0.61	10.4	0.52	10.5	0.48	10.23	0.15
15	15.02	0.36	14.90	0.37	15.60	0.32	14.62	0.28	15.30	0.17

Table 2. A Comparative study with existing methods for English documents

	Method-1[13]		Method-2[4]		Method-3[6]		Method-4[15]		Proposed Method	
True Angle	M	SD	M	SD	M	SD	M	SD	M	SD
3	3.12	0.34	3.43	0.94	3.86	0.81	3.21	0.51	3.16	0.18
5	5.68	0.456	5.09	1.04	5.71	0.95	5.52	0.68	5.10	0.28
10	10.17	0.42	10.19	0.82	10.73	0.79	9.86	0.54	10.23	0.15
15	15.72	0.51	15.35	0.93	15.53	0.51	15.02	0.32	15.30	0.17

Table 3. Mean and Standard Deviation for clean and noisy handwritten English document

	Handwritten		Noisy Handwritten	
True Angle	M	SD	M	SD
3	3.10	0.20	2.80	0.35
5	4.94	0.32	4.57	0.47
10	10.10	0.10	8.94	0.89
15	15.46	0.37	13.97	0.85

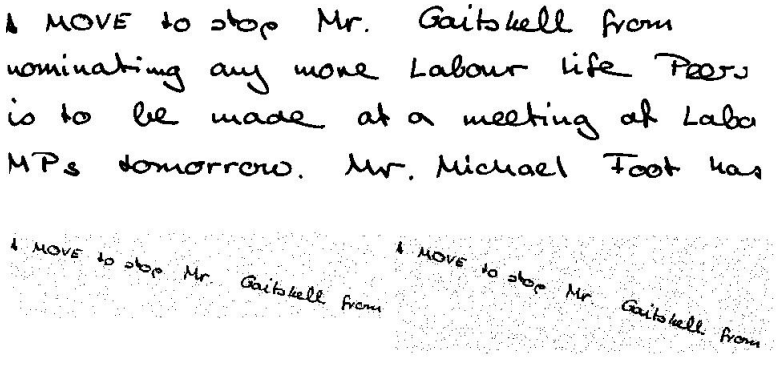

Fig. 2. Clean(top) and noisy(bottom) English handwritten documents

Moreover, amongst four south Indian scripts, our method obtained better results in terms of M and SD for Malayalam documents. A comparative study with other existing methods is carried out to show the performance of our method in terms of accuracy and efficiency. The mean and standard deviation obtained using the proposed method and the other methods are reported in Table 2 for printed English documents. From Table 2 it is clear that the proposed method performs better compared to other existing methods with respect to mean and standard deviation. We extended our experiment for handwritten English documents. For this experiment we considered 10 documents and each are rotated with four mentioned skew angles. Mean and standard deviation obtained for the handwritten English documents are reported in Table 3. It is clear that the proposed method

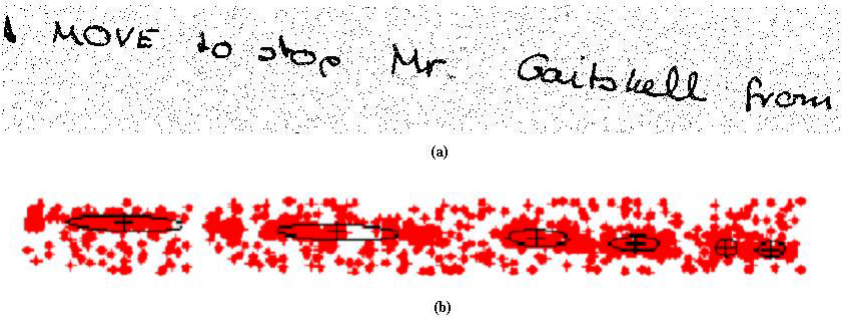

Fig. 3. (a): A Noisy handwritten document (b): Its illustration with mixture of six Gaussians

performs better even for handwritten documents. To check the robustness of the proposed algorithm, we tested our method on noisy documents also. For this, we considered five noisy handwritten documents[3] and each were rotated with four true angles. Sample handwritten documents contaminated by noise is as shown in Figure 2. Results obtained from the method are reported in Table 3. Figure 3 shows the illustration of Gaussian mixture for noisy handwritten document.

4 Discussion and Conclusion

Mixture Models are a type of density model which comprise a number of component functions, usually Gaussian. These component functions are combined to provide a multimodal density. GMM is widely used in data mining, pattern recognition, machine learning, and statistical analysis. In many applications, their parameters are determined by maximum likelihood, typically through iterative learning of the EM algorithm. In this paper, an efficient and robust methodology for skew estimation based on GMM is presented. The proposed method is

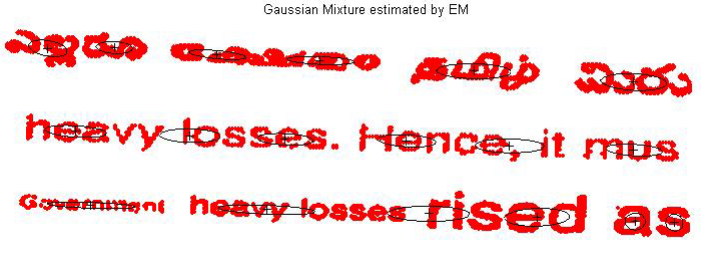

Fig. 4. Mixture of Gaussians for: (a) a multilingual text line (b) degraded text line and (c) text with varying size of words

[3] Here we used Salt-and-Pepper of noise density 0.02

independent of scripts, style and font size. The results for this is illustrated in Fig.4. Extensive experiments have been carried out to study the performance of the proposed method by considering the documents such as printed south Indian scripts, handwritten English and noisy handwritten documents. These experiments revealed the superiority of the proposed method. We plan to extend this work for other Indic scripts in future.

References

1. Basu, S., Chaudhuri, C., Kundu, M., Narsipuri, M., Basu, D.K.: Text line extraction from multi-skewed handwritten documents. Pattern Recognition 40(6), 1825–1839 (2007)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Pearson Education (2002)
4. Yan, H.: Skew correction of document images using interline cross-correlation. Computer Vision, Graphics, and Image Processing 55, 538–543 (1993)
5. Kapoor, R., Deepak, Kamal.: A new algorithm for skew detection and correction. Pattern Recognition Letters 25, 1215 (2004)
6. Lu, Y., Tan, C.L.: A nearest neighbor chain based approach to skew estimation in document images. Pattern Recognition Letters 24, 2315–2323 (2003)
7. Manjunath Aradhya, V.N., Hemantha Kumar, G., Shivakumara, P.: An accurate and efficient skew estimation technique for south indian documents. International Journal of Robotics and Automation (in press)
8. Baird, H.S.: The skew angle of printed documents. In: Proceedings of Conference Society of Photographic Scientists and Engineers, pp. 14–21 (1987)
9. Hou, H.S.: Digital Document Processing. Wisely, New York (1983)
10. Le, D.S., Thoma, G.R., Wechsler, H.: Automatic page orientation and skew angle detection for binary document images. Pattern Recognition 27, 1325–1344 (1994)
11. Srihari, S.N., Govindaraju, V.: Analysis of textual images using the hough transform. Machine Vision and Applications 2, 141–153 (1989)
12. Akiyama, T., Hagita, N.: Automated entry system for printed documents. Pattern Recognition 23(11), 1141–1158 (1990)
13. Pal, U., Chaudhuri, B.B.: An improved document skew angle estimation technique. Pattern Recognition Letters 17, 899–904 (1996)
14. Postl, W.: Detection of linear oblique structures and skew scan in digitized documents. In: Proceedings 8th International Conference on Pattern Recognition, pp. 687–689 (1986)
15. Yang, C., Wang, S., Heng, L.: Skew detection and correction in document images based on straight-line fitting. Pattern Recognition Letters 24, 1871 (2003)
16. Yu, B., Jain, A.K.: A robust and fast skew detection algorithm for generic documents. IEEE Transactions on Pattern Analalysis and Machine Intelligence 29(10), 1599–1629 (1996)