2012 10th IAPR International Workshop on Document Analysis Systems

# Accent Detection in Handwriting based on Writing Styles

Chetan Ramaiah, Utkarsh Porwal and Venu Govindaraju
*Department of Computer Science and Engineering*
*University at Buffalo*
*Buffalo, NY*
*chetanra, utkarshp, govind@buffalo.edu*

***Abstract*—Accent in handwriting can be defined as the influence of a writer's native script on his/her writing style in another script. In this paper, we approach the problem of detecting the existence of accents in handwriting. We approach this problem using two sets of writers, those who can write only in English, and the other set being multilingual writers who can also write in English. We learn the writing styles that are predominant in each set and use it as features in classification. Latent Dirichlet Allocation is used to learn the distribution over writing styles. Experimental results suggest the existence of accents in handwriting.**

***Keywords*-Accents in Handwriting;Handwriting Styles Modeling;Topic Models**

## I. Introduction

Accent in speech can be defined as the differences in the articulation habits of non native speakers as compared to that of native speakers. These differences arise because the intonation, rhythm and pronunciation styles of languages differ from one another. Automated speech recognition is the process of converting spoken words to text, generally using machine learning approaches. There are two primary branches of automated speech recognition: Speaker dependent and the speaker independent model [1], [2], [3]. In the speaker dependent model, a model is trained that is specific to a particular user. Clearly, this approach needs a training corpus from each user. The other alternative, that is the speaker independent approach, consists of two widely used approaches: accent dependent and the accent independent approach [4]. The accent dependent approach entails identifying the accent and then training a model for each accent. In the accent independent approach, a model is trained which encompasses as many different accents as possible. In both cases, accents play a huge role in speech recognition. Like in speech recognition, we can use accents in handwriting to help improve the recognition accuracy in OCR systems.

Research in the field of accents in handwriting has been fairly limited. Accent detection and identification can be a useful tool in handwriting analysis, forensics and as a soft biometric. Analysis of handwriting styles is beginning to recieve considerable attention recently. Bharadwaj et al. [5] used a topic model based approach to learn the latent handwriting styles and used the handwriting styles to perform writer identification. Bharadwaj et al. [6] also used handwriting styles for handwritten document retrieval. Brink et al. [7] introduced the concept of vantage writers, which is the idea that every handwriting sample can be represented by comparing it with a selected collection of handwriting samples picked randomly. Demonstrating the utility of handwriting styles for a variety of tasks ranging from handwriting recognition [8] to historical document dating [9] clearly illustrates that there is useful latent information in handwriting.

Farooq et al. [10] proposed a method for detecting accents in Arabic writing by using Gabor filters for feature extraction and Support Vector Machines (SVM) for classification. They postulated that the shape, smoothness and sharpness of characters were the distinctive features distinguishing the two categories. The primary focus here was to distinguish between native and non-native writers of the Arabic script. To the best of our knowledge, there has been no focus on learning the writing styles that native writers tend to use and utilize that attribute to distinguish between writers. A generative model that learns writing styles of native and non native writers can be used to study the differences between writing styles and the effect on writing one script on learning to write another script.

In this work, we approach the problem of accent detection in handwriting. The problem is of classifying a handwriting sample into one of two categories: native or non-native writer. Feature extraction techniques such as fractal features [11], contour direction [12], structure and concavity [13] are applied. Latent Dirichlet Allocation(LDA) [14] is used to learn the distribution over writing styles. Finally, Support Vector Machine(SVM) [15] is the classifier used to distinguish between the two classes. Experiments are performed at the feature level and at the writing styles level. Figure 1 illustrates the architecture of the system. The data used in our experiments were relatively clean, hence most standard preprocessing and line segmentation techniques work.

The remainder of this paper is organized as follows: Section II illustrates how LDA was applied to handwriting styles, section III covers the various feature extraction techniques, section IV covers experiments and results, and section V is the conclusion.

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.13

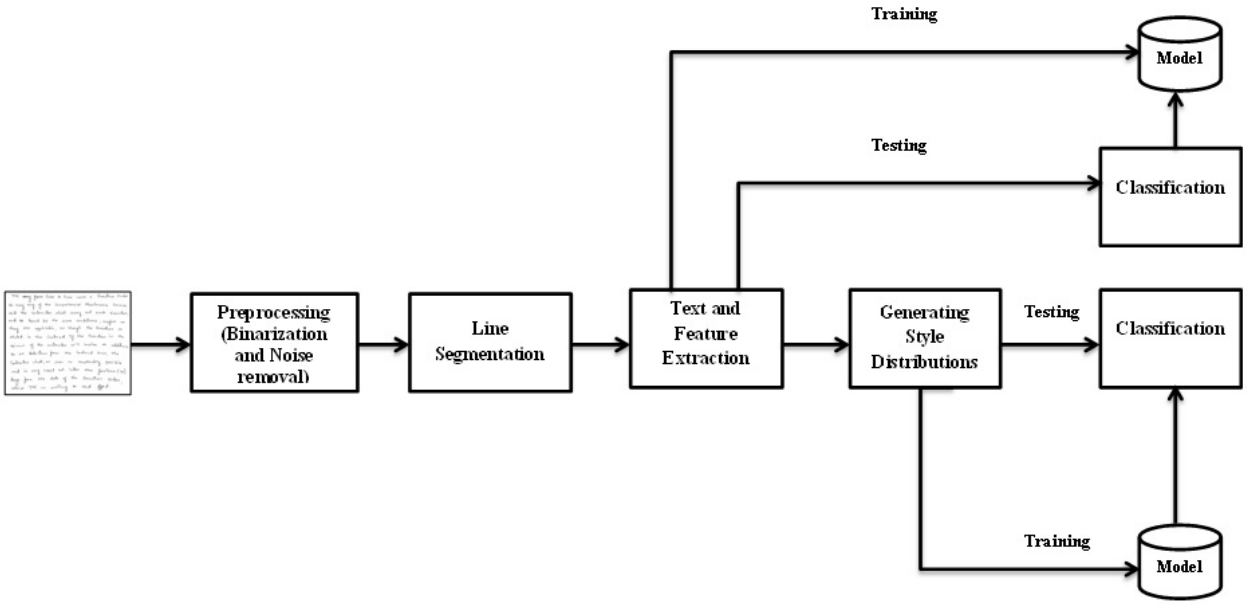


Figure 1. Architecture of the system

## II. Modeling handwriting styles using Latent Dirichlet Allocation

LDA was originally a generative unsupervised model for topic modeling in text documents [14], it is adaptable to learn handwriting styles distribution in handwriting documents [6]. LDA is an approach that was proposed to identify latent topics, which are a distribution over words, in text documents, and to learn a distribution over the topics for each document. In a similar vein, we attempt to learn the latent handwriting styles present in handwriting documents and learn a distribution over handwriting styles for each document. We also postulate that this distribution is a distinguishing attribute between documents. Without loss of generality, if we assume that a script has *k* unique writing styles (slanting, straight, loopy etc), then each handwritten document will be a distribution over these writing styles. This distribution can be used to draw a correlation between writing styles and the nativity of the writer.

The generative process for modeling handwriting styles is:

(1) Select $\theta \sim$ Dir($\alpha$).

(2) For each of the N features,

- Select the distribution over writing styles $\omega \sim$ Multinomial($\theta$).
- Select features f from the multinomial distribution P(f|$\omega$,$\beta$).

The joint distribution of features f , a writing style distribution $\theta$ and a set N of writing styles $\omega$ is given by

$$p(\theta, w, f) = P(\theta|\alpha) \prod_{i=1}^{N} P(w_i|\theta) P(f_i|w_i, \beta) \qquad (1)$$

where P($\theta|\alpha$) is the Dirichlet Prior. Marginalizing over $\theta$ and styles $\omega$ , the distribution of each image is given by:

$$P(f|\theta, \beta) = \int p(\theta|\alpha)[\prod_{i=1}^{N} \sum_{\omega_i} p(\omega_i|\theta) P(f_i, |\omega_i, \beta)]d\theta \quad (2)$$

Variational approximation is used to learn the posterior distribution, due to the intractability of the distribution for exact inference [14].

LDA not only enables us to study how nativity affects writing style, but it is also a dimensionalty reduction technique as representing a document by its distribution over writing styles greatly reduces the feature length.

## III. Feature Extraction

In this section, we will describe the different feature extraction procedures considered and why they are relevant to our problem. Features which maximize the discriminability of writers have been chosen for our experiments. Slant is said to be the most distinguishing feature in writer identification [12]. Along with slant, we have also experimented with fractals, structure and concavity features at the line level. Combination of features also increases the accuracy of the system.

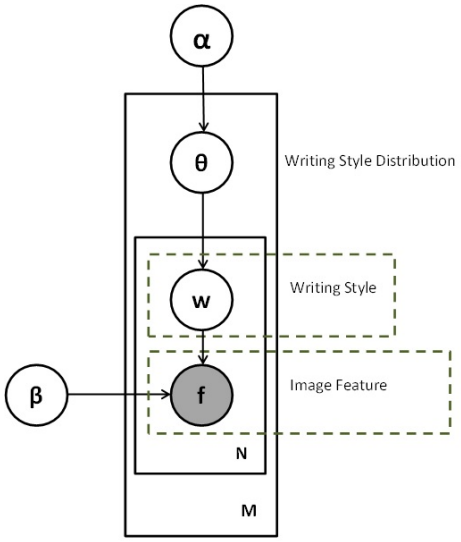


Figure 2. Latent Dirichlet Allocation Model for modeling handwriting styles [6]

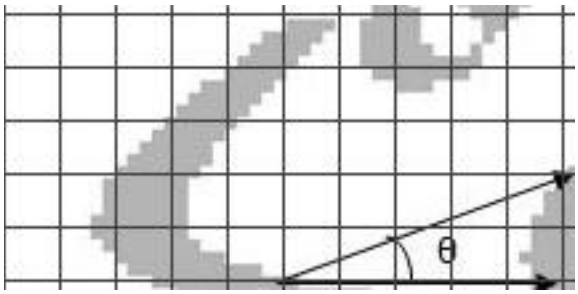


Figure 3. Schematic description for feature extraction. The angle $\theta$ is formed with the centre and the foreground extremity with the horizontal

### A. Contour Direction Distribution

The contour direction distribution (CDD) is extracted by evaluating the angle of inclination of fragments [12]. This angle of inclination can be considered to be the approximate slant with which a writer writes. The angle that two contour pixels that are a fixed distance away from each other makes with the horizontal is captured and used to generate a probability distribution function (PDF). The PDF is obtained by generating a histogram of angles. The distance between the central pixel and the pixel on the extremity is heuristically determined. Figure 3 illustrates the process. Pixels which lie at a distance of 3, 4 and 5 pixels from the centre pixel were taken for our experiments. They were binned into histograms containing 8, 12 and 36 bins giving us a feature vector length of 36.

### B. Fractal Features

Fractal features extracts information about how a handwritten region's pixels are affected at different resolutions [11]. This approach involves random selection of pixels, counting the number of on pixels around it for different areas, and normalizing the count based on the area of the region around the pixel. Several areas are chosen and normalized to get the count at different resolutions. Figure 4 illustrates the extraction of fractal features. Boxes of length 4,8,12...32 are drawn around randomly selected pixels, the number of on pixels are counted and averaged, and then normalized. This approach gives a feature length of 41.

### C. Structural and Concavity Features

Structural features are designed to capture small strokes on the gradient map that are directed horizontally, vertically or diagonally. These features are then combined to form larger scale features [13]. These features tend to capture the slight deviations that are unique to each writer. Concavity features capture information at a larger scale, they are designed to capture large horizontal and vertical strokes in the image. A feature length of 20 is obtained in this approach.

## IV. Experiments and Results

Experiments have been performed on various features and a unique dataset. Feature combination was performed by concatenating different features and using SVM for classification. For LDA, we heuristically chose 10 to be the number of unique writing styles, and each document was represented as a distribution over these 10 writing styles. In our experiments, great care was taken to ensure that the test dataset consisted of writers whose samples were present in the training set as well writers whose sample were not present. Upon examining the test cases where our approach failed, it was found that there was no major difference in accuracy in both the sets.

### A. Dataset

Around 200 non-native writers of English compromise the non-native corpora. Each writer wrote a page in his/her native script and a page in a non-native script, the contents of which are chosen at random. The non-native corpus consists of writers whose native scripts were various Indian scripts. No special significance is placed on the non-native writer's native script, that is, our dataset consists of non-native writers whose native scripts vary from writer to writer. Figure 5 is a sample document. Attributes such as age, handedness, sex, native script, familiarity with non-native scripts etc were also collected. These documents were scanned at 300 dpi and standard preprocessing and line segmentation algoritms were applied.

The native corpus was obtained from the Library of Congress collection [16], and consists of around the same number of writers. These documents also consist of about a page from each writer, which were preprocessed and then line segmentation was performed. There are a total of about 500 documents , which results in over 5000 lines of

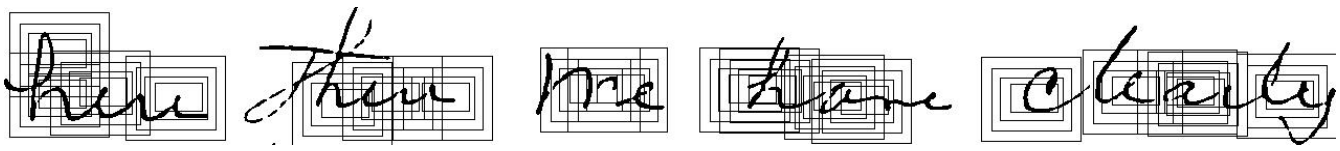

Figure 4. Illustration for Fractal feature extraction.

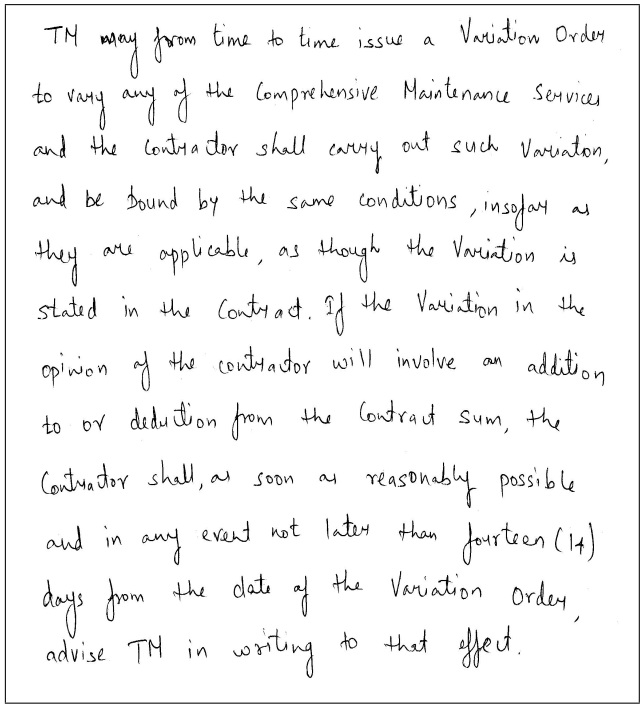


Figure 5. Sample document from a non-native writer

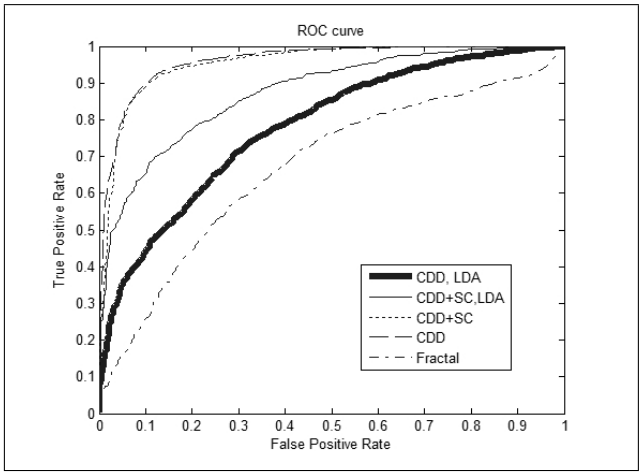


Figure 6. ROC curves of all experiments

Table I
RESULTS

| Features | Accuracy(%) |
|---|---|
| CDD | 94.14 |
| Fractal | 74.35 |
| CDD+ SC | 97.67 |
| CDD, LDA | 89.96 |
| CDD+SC, LDA | 93.59 |

handwriting, of which about 4000 were used in training and the remaining for testing.

### *B. Results*

Table I presents the results obtained. Here, CDD stands for Contour Direction Distribution, SC for Structural and Concavity features, and LDA for Latent Dirichlet Allocation. The best result was obtained by using the feature combination of CDD and SC. The accuracy drops by around 3% when LDA is applied over a feature extraction technique. This can be attributed to the loss of information due to dimensionality reduction. In the case of CDD+SC, LDA the feature vector length was reduced from 56 to 10. Although there is a drop in accuracy when using LDA, it provides the ability to analyze the distribution of writing styles in native and non-native writers. CDD features are naturally very good at increasing the discriminability of handwritings. They however, do not capture each writer's individuality in writing, such as the small trails after each character, the slant in dashing the t's etc. This kind of feature is captured in the structural features, as a result of which, combining the two feature extraction techniques gives improved results. Upon examining the confusion matrix for all the results, it was found that most of the failed cases were non-native writers being confused as native writers. This suggests that a very small percentage of non-native writers wrote naturally enough to pass off as native writers. Figure 6 gives the Receiver operating characteristic curves of all the experiments.

## V. CONCLUSION AND FUTURE WORK

The results strongly suggest the existence of accents in handwriting. This opens up a new area of research, where, like in speech recognition, we can use our results in Optical Character Recognition to improve recognition accuracy. Handwriting recognition is still a challenging problem due to

the large variability in writing styles, and our approach can help improve the accuracy. Our approach will also benefit the writer identification and handwritten document retrieval community. Although we have experimented on English only, the same approach can be used on other languages as well.

In the future, we can study accent identification, which would identify the non-native writer's native script. This would be beneficial in forensics and biometrics, as native script identification will be useful in uniquely identifying a person. Other work can include a scientific approach to studying and analyzing handwriting styles.

## REFERENCES


[1] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, pp. 141–153, 2004.

[2] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763 – 786, 2007.

[3] A. Faria, "Accent classification for speech recognition," in *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, ser. Lecture Notes in Computer Science, S. Renals and S. Bengio, Eds., vol. 3869. Springer, 2005, pp. 285–293.

[4] P. Fung and L. W. Kat, "Fast accent identification and accented speech recognition," in *In Proc. ICASSP*, 1999, pp. 221–224.

[5] A. Bhardwaj, M. Reddy, S. Setlur, V. Govindaraju, and S. Ramachandrula, "Latent dirichlet allocation based writer identification in offline handwriting," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. ACM, 2010, pp. 357–362.

[6] A. Bharadwaj, A. Thomas, Y. Fu, and V. Govindaraju, "Retrieving handwriting styles: A content based approach to handwritten document retrieval," in *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, nov. 2010, pp. 265 –270.

[7] A. Brink, L. Schomaker, and M. Bulacu, "Towards explainable writer verification and identification using vantage writers," in *ICDAR*, 2007, pp. 824–828.

[8] A. Bhardwaj, F. Farooq, H. Cao, and V. Govindaraju, "Topic based language models for ocr correction," in *Proceedings of the second workshop on Analytics for noisy unstructured text data*, ser. AND '08, 2008, pp. 107–112.

[9] C. Ramaiah, G. Kumar, and V. Govindaraju, "Handwritten document age classification based on handwriting styles," in *DRR*, 2012, to be published.

[10] F. Farooq, L. Lorigo, and V. Govindaraju, "On the Accent in Handwriting of Individuals," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, Oct. 2006.

[11] A. Chaabouni, H. Boubaker, M. Kherallah, A. M. Alimi, and H. E. Abed, "Fractal and multi-fractal for arabic offline writer identification," in *ICPR*, 2010, pp. 3793–3796.

[12] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 701 –717, april 2007.

[13] J. T. Favata, G. Srikantan, and S. N. Srihari, "Handprinted character/digit recognition using a multiple feature/resolution philosophy," in *Fourth International Workshop on Frontiers in Handwriting Recognition*, 1994, pp. 57–66.

[14] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[15] C. chung Chang and C.-J. Lin, "Libsvm: a library for support vector machines," 2001.

[16] Library of congress, manuscript division. [Online]. Available: http://www.loc.gov