Writer Identification in Offline Handwriting Using Topic Models

Anurag Bhardwaj
Department of Computer Science
University at Buffalo
Amherst, NY 14228
ab94@buffalo.edu

Manavender Malgireddy
Department of Computer Science
University at Buffalo
Amherst, NY 14228
mrm42@buffalo.edu

Srirangaraj Setlur
University at Buffalo
Amherst, NY 14228
setlur@buffalo.edu

Venu Govindaraju
Department of Computer Science
University at Buffalo
Amherst, NY 14228
govind@buffalo.edu

Ramachandrula Sitaram
HP Labs India
Bangalore, INDIA 560030
sitaram@hp.com

Abstract

In this paper, we describe a novel application of Topic Models for the task of writer identification from offline handwriting. State-of-the-art methods for writer identification employ the traditional feature-classification paradigm which does not provide enough information about the handwriting attributes such as writing style. We propose to address this issue by using a generative model in form of Latent Dirichlet Allocation(LDA) that automatically infers writing styles from handwritten document collection. This information is then used to represent each writer as a distribution over multiple writing style for classifying any unknown writer sample. Our experimental results show comparable performance with baseline systems and also demonstrate the efficacy of LDA for learning multiple handwriting styles.

1 Introduction

Handwriting can be understood as a generative process where the observable data (handwriting) is generated through a process that depends on the writer as well as the content being written. This leads us to two simple conclusions:(i)Same content written by different writers should be different and (ii) Different content written by the same writer should be different. The first conclusion is an automatic choice for any writer identification technique since the same content normalizes the issues related to document content and any analysis (comparison on image feature or model space) of such content for different writers would only model the writer style and characteristics which are of primary interest. This direction of research (text-dependent writer identification) has received considerable attention in the recent past [9], where researchers have analyzed a set of known character or word images for all writers and compared them in the feature or model space to understand writer specific attributes.

However, obtaining known character samples from each writer is impractical for a large document collection which makes text-dependent writer identification infeasible in many scenarios. This issue leads us to a new research question: how do we normalize effects of different content written by writers and obtain a text-independent representation of writer attributes. Few researchers have attempted to solve this problem by focusing on text-independent image features [8] which only

extracts writer specific information from the text regardless of its content. However, this method fails to answer a number of questions (e.g. what writing styles are involved with each writer or which two authors share a specific writing style) which can provide a greater insight in the forensic analysis of handwriting. The failure to explicitly model the writer style as a function of text independent features provide us a strong motivation to use a new model for writer identification which can efficiently model the writing style of each writer irrespective of text content of handwriting.

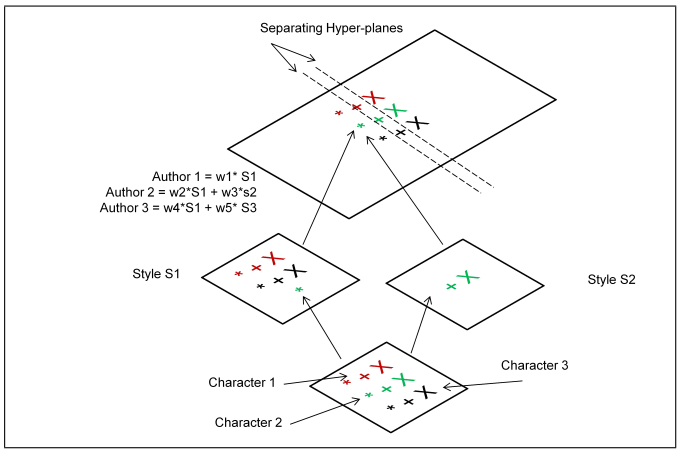

Figure 1: Proposed Writer Identification Model.

2 Writer Style modeling using LDA

We propose to use an Author-Style-Feature model (similar to Author-Topic Models[7]) for modeling writing style as well identifying writer class. As shown in figure 1, each author or writer is represented as a probability distribution over multiple writing styles (e.g. cursive, loopy, straight slant) which in turn is represented as a distribution over various text-independent features extracted from handwriting. Using text-independent features provides us the flexibility to represent a large number of writers with limited data and enables us to explicitly model their writing styles. We propose to use Latent Dirichlet Allocation (LDA) [1] for modeling writing style from handwriting features. There are a number of advantages of using this approach as compared to the current baseline systems for writer identification. Firstly, LDA provides us a generative model for writer style modeling which is flexible (as it can be used with most of the image based features) and provides a strong theoretical framework for writer style learning. Secondly, this approach also attempts to address the issue of large number of writers in the corpus. Usually, with a large number of authors, the task of classification becomes more complex with increase in number of classes. We hypothesize that the growth in the number of writers is limited as compared to the growth in writer style and each writer can be efficiently modeled as a distribution over multiple writing styles. Therefore, LDA enables us to efficiently model fewer writer style classes instead of larger writer classes with the same text-independent features. Thirdly, LDA based style modeling helps us in comparing two different writers on style space as well. Using this mechanism, we can easily answer why two writers are similar to each other or what writing style (or handwriting accent) are involved in the handwriting which can also be utilized in accent classification (native or non-native writer) of handwriting [3].

Firstly, the input image is binarized and connected component analysis is performed to extract components from the image. The extracted components are then passed through an edge detection scheme where the contour image of the component is obtained. Our feature extraction is then performed on the contour image. We adapt contour angle features described in [8] for our task. From the contour image, each foreground pixel (pixel set as black) is set at the center of a rectangular mask of width n pixels on left, right and above. For each periphery foreground pixel which forms an edge with center pixel[2], we compute the angle of the edge with respect to horizontal and update

an angle histogram. Finally, the angle histogram corresponding to the whole line image is taken as the feature value. In our experimental setup, we use three different masks of pixel width 3, 4 and 5. Angles computed from each of the masks are binned into 8, 12 and 16 bins, thereby accounting for 36 features for each line image.

Using a notation similar to LDA, we describe a generative model for each feature f in a handwritten document D as follows:

1. Select $N \sim$ Poisson(ξ).
2. Select $\theta \sim$ Dir(α).
3. For each of N features:
   (a) Select a writing style $s \sim$ Multinomial(θ).
   (b) Select a feature f_n from $P(f_n|s_n, \beta)$ which is also a multinomial probability distribution.

Exactly like an LDA model, the joint distribution of a writing style distribution θ, a set of N writing styles **s** and observed feature values **f** is given as:

$$P(\theta, \mathbf{s}, \mathbf{f}) = P(\theta|\alpha) \prod_{n=1}^{N} P(s_n|\theta) P(f_n|s_n, \beta) \tag{1}$$

Once LDA has generated a probability distribution over all the K writing styles for each line image, we use this distribution to identify the writer class. A simpler way is to use the whole distribution as a K dimensional feature vector and train a multi-class classifier for discriminating each writer based on his writing style distribution.

3 Experiments

Our dataset is a subset of publicly available IAM database [6] and consists of 4075 line images written by 93 different writers. We conduct 4-fold cross validation to benchmark the performance of various writer identification systems. Our baseline system uses the contour angle features directly to train a 93 class SVM using the LIBSVM [5] implementation. The LDA system uses the contour angle features to obtain a set of writing styles from LDA [4] and again uses the style distribution obtained from LDA to train a 93 class SVM. We also perform two similar experiments with different methods for generating writer style (K-means and Hierarchical clustering) to illustrate the efficacy of LDA. Each system is evaluated with two different feature sets, one consisting of only style features and other consisting of both style and contour angle features. Table 1 shows the relative performance of each system over each fold. As shown, LDA based method with only style features performs closer to the baseline method (only SVM) with an extra information in form of writing styles for each writer. On the other hand, in combination with angle features, LDA based method outperforms all other methods which demonstrates its strength. Figure 2 also shows the qualitative performance of LDA for generating writer styles. We randomly choose two writing styles generated and sample one image written by 10 different writers. As the figure suggests, there are two distinct writing styles generated by LDA (e.g. cursive), which are consistent even across samples from different writers.

4 Conclusion

In this paper, we present a novel application of using statistical topic models (LDA) for generating writing styles from offline handwriting. Our proposed technique extends the current state-of-the-art methods in writer identification by providing an explicit modeling of writer style using a generative model. Our current work focuses on extending the model for writer accent classification and evaluating the effect of various image features on LDA based writer style modeling.

Table 1: Writer Identification Results for 4-folds

Method	Fold-1	Fold-2	Fold-3	Fold-4	Overall
SVM (36 contour angle features)	83.84%	83.15%	83.76%	82.96%	83.43%
k-Means+SVM (10 style features)	59.79%	61.63%	61.58%	60.04%	60.76%
H-Clustering+SVM (10 style features)	66.44%	62.02%	69.50%	67.44%	66.33%
LDA+SVM (10 style features)	**80.32%**	**80.42%**	**81.18%**	**77.18%**	**79.80%**
k-Means+SVM (36 contour angle + 10 style features)	83.84%	84.32%	83.66%	84.78%	84.14%
H-Clustering+SVM (36 contour angle + 10 style features)	84.03%	84.71%	84.45%	84.78%	84.49%
LDA+SVM (36 contour angle + 10 style features)	**84.88%**	**85.88%**	**85.34%**	**84.68%**	**85.20%**

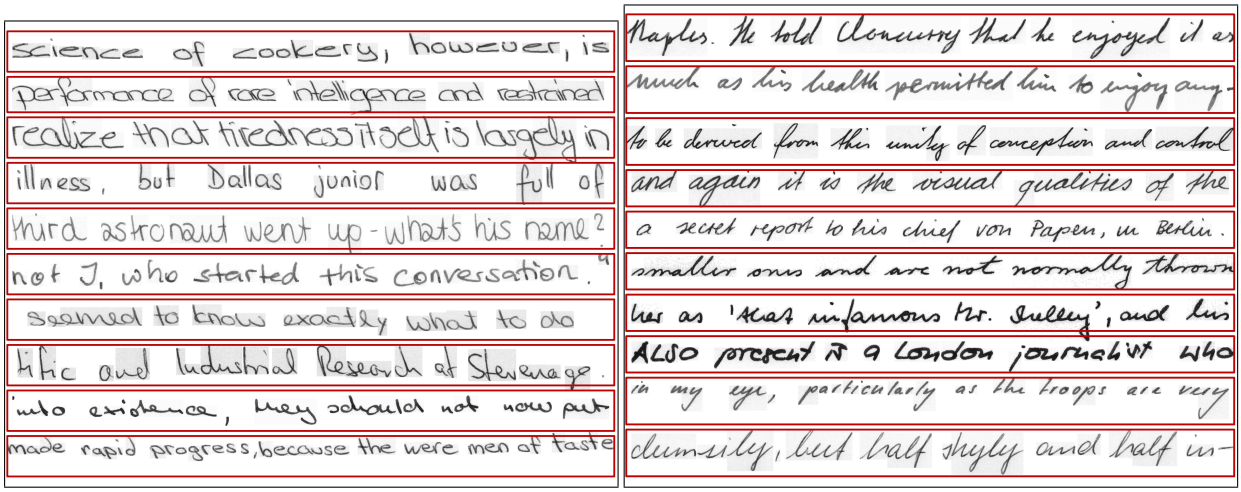

Figure 2: Two Writing Styles generated by LDA - Each containing samples from 10 different writers.

References

[1] Blei, David M. & Y. Ng, Andrew & Jordan, Michael I. (2003) Latent dirichlet allocation, In *The Journal of Machine Learning Research 3*, pp.993-1022.

[2] Bresenham Line Drawing Algorithm. http://en.wikipedia.org/wiki/Bresenhams_line_algorithm

[3] Farooq, Faisal & Lorigo, Liana & Govindaraju, Venu (2006) On the Accent in Handwriting of Individuals. In *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition*.

[4] Latent Dirichlet Allocation. http://www.cs.princeton.edu/~blei/lda-c/

[5] Chang, Chih-Chung & Lin, Chih-Jen (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[6] Marti, U. & Bunke, H. (2002) The iam-database: an english sentence database for off-line handwriting recognition. In *International Journal on Document Analysis and Recognition 5(1)*, pp. 39-46.

[7] Rosen-zvi, Michael & Griffiths, Thomas & Steyvers, Mark & Smyth, Padhraic (2004) The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 487-494.

[8] Schomaker, Lambert & Bulacu, Marius (2007) Text-Independent Writer Identification and Verification Using Textural and Allographic Features. In *IEEE transactions on pattern analysis and machine intelligence 29(4)*, pp. 701-717.

[9] Srihari, S. & Beal, M. & Bandi, K. & Shah, V. & Krishnamurthy, P. (2005) A Statistical Model for Writer Verification. In *Proceedings of 8th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1105-1109.