2012 International Conference on Frontiers in Handwriting Recognition

Modeling Writing Styles for Online Writer Identification: A Hierarchical Bayesian Approach

Arti Shivram, Chetan Ramaiah, Utkarsh Porwal and Venu Govindaraju
Dept. of Computer Science and Engineering
University at Buffalo, Amherst, NY
{*ashivram,chetanra,utkarshp,govind*}@*buffalo.edu*

Abstract

With the explosive growth of the tablet form factor and greater availability of pen-based direct input, writer identification in online environments is increasingly becoming critical for a variety of downstream applications such as intelligent and adaptive user environments, search, retrieval, indexing and digital forensics. Extant research has approached writer identification by using writing styles as a discriminative function between writers. In contrast, we model writing styles as a shared component of an individual's handwriting. We develop a theoretical framework for this conceptualization and model this using a three level hierarchical Bayesian model (Latent Dirichlet Allocation). In this text-independent, unsupervised model each writer's handwriting is modeled as a distribution over finite writing styles that are shared amongst writers. We test our model on a novel online/offline handwriting dataset IBM_UB_1 which is being made available to the public. Our experiments show comparable results to current benchmarks and demonstrate the efficacy of explicitly modeling shared writing styles.

1. Introduction

The origins of writer identification and verification can be traced back to the field of automatic handwriting recognition [8]. Historically, handwriting recognition has focused on building invariant representations of content by minimizing inter-writer variation. This approach broadly treats writer-specific variation(s) as noise to be minimized or eliminated. In contrast, writer identification exploits writer-specific variation in order to discriminate between writers. This overarching objective to distinguish between writers has, in our observation, led to most research modeling writing styles as being intrinsic and unique to individual writers. This has been achieved by building distinct feature-space representations particular to each writer. The underlying assumption of this approach is that each writer has his/her unique handwriting style that is not shared with other writers and that the feature-space fully and completely defines each style and consequently, each writer.

We depart from this approach in the assumption that although each writer's *handwriting* is unique, their *writing styles* are not. That is, there are stylistic commonalities across writers, for instance, the degree of slant, loopiness and so on [4]. To this end, we seek to automatically learn different handwriting *styles* (shared among writers) and generate each writer's handwriting by sampling from a distribution of these finite writing styles using a hierarchical Bayesian model. We have developed and applied our modeling framework on a novel dataset - IBM_UB_1 - which is part of a larger, multi-lingual dataset, currently being made available to the research community in a phased manner [1].

1.1 Problem Relevance

The objective of the writer identification problem is to automatically determine the identity of the writer of a handwritten document from among a set of possible writers. Handwritten documents may come from scanned images of written pages *(offline data)* or may be captured as a series of pen-tip coordinates when written on digital media *(online data)*.

Research in this field has found application in a variety of domains ranging from forensics and security to intelligent, adaptive systems. Automated writer identification using statistical models provides a scientific basis for forensic document analysis and is also used as a soft biometric for person identification [12]. In the context of digital libraries it provides tools to index and retrieve historical handwritten documents. Recently, it has found application in validating authorship

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.235

of retrieved handwritten documents.

In intelligent environments such as smart meeting rooms, writer identification systems may be used to label handwritten notes, say, text written on whiteboard, with the writer's identity [13]. This information can help cross-validate results obtained from other modes such as video and/or audio. In addition, writer identification systems may help the automatic generation of tags and metadata that are used in indexing, searching and retrieval.

Moreover, with the explosive growth in the tablet form factor and greater availability of pen-based direct input platforms, writer identification in online environments offers the possibility of a more natural user experience via system adaptation. A related challenge caused by the ubiquity of this technology is the sheer volume of digital data generated. Current methods used for online writer identification may not be scalable to this increased volume, an issue we attempt to address by using a new model in this paper.

1.2 Extant Research in Online Methods

Past approaches pursued in addressing the problem of writer identification may be broadly classified into two categories - text-dependent and text-independent. In the former, the same text is written by different writers and the variation between writers is explicitly modeled. In contrast, the latter method identifies and uses features unrelated to the content written in order to distinguish between writers.

Since our model represents an advancement to online writer identification, we situate our review in the online paradigm. A signal-processing approach is used by Matsuura and Thumwarin [11] wherein a finite impulse response (FIR) system is modeled using the discrete cosine transform (DCT) of pen-tip coordinates. Identification is accomplished by comparing the impulse response pattern of test data with the referenced FIR system [11]. Li and Tan [10] propose a text-independent system which uses a codebook of features based on temporal sequence and shape codes related to the writing speed, pressure and trajectory. Subsequently, an artificial neural network classifier identifies the writer for an arbitrary input based upon several distance measures. Their system was tested on both English and Chinese scripts.

Tsai and Lan [17] attempt to solve the problem of writer identification by using the point distribution model (PDM). This model learns the eigenstructure for individual writers which represents their unique writing styles. Discrimination between writers is accomplished using the sum of strengths of major eigenmodes as a similarity metric. Tan et al [16] use a text independent three stage algorithm involving character-level prototypes from the IRONOFF database. These prototypes serve as the basis for creating individual distributions of handwriting styles for documents. Specifically, each document is mapped to the character prototype and transformed into a frequency vector using a fuzzy c-means algorithm. Subsequently, these frequency vectors are used in the classification stage to identify the writer corresponding to the test document.

Schlapbach et al [13] employ Gaussian mixture models to model individual writers. As a first step, they use all the training data from all the writers to train a single universal background model (UBM). In the second step, they build a model for each writer by adapting the universal model (UBM) to each writer's training data. During testing, a text of unknown identity is presented to each model whereupon the model returns a log-likelihood score. These scores are sorted and ranked and the text is assigned to the writer whose model produces the highest score.

In most of the above outlined work, the term 'writer' and 'writing style' may be used interchangeably since the implicit assumption is that a writing style is embedded in a writer. That is, each writer completely and uniquely defines his/her writing style. Different methods of discrimination are then employed to distinguish between writing styles and thus, writers. We depart from past research with respect to this fundamental assumption. Our starting point is that although *handwriting* is unique to writers, *writing style* represents a shared component of individual handwriting. Thus a person's handwriting can be *a priori* conceptualized as an individual-specific combination (determined by a person's physiology - *genetic factors*) of a shared pool of writing styles (often determined culturally - *memetic factors*) [14]. We explicitly model this theoretical framework by adapting the Latent Dirichlet Allocation model employed in Bhardwaj et al. [4][5] to the task of online writer identification.

A second limitation of directly modeling individual writers (without taking into account shared writing styles) relates to scalability. Specifically, as new writers are added to a corpus, the model necessarily needs to be recalibrated and reestimated. Thirdly, it is not possible to identify new writers without their pre-existing models. By using LDA we can efficiently model a large superset of writers by using a significantly smaller subset of writing styles. Moreover, as LDA is a generative model, writers who are not in the original corpus may also be identified from the existing, learned distribution of writing styles. The LDA-based model we outline below thus overcomes both these limitations (scalability and extensibility) which is critical in the online domain.

The rest of the paper is organized as follows. We first outline the LDA framework as applied to online writer identification. Subsequently, we provide an overview and description of the new dataset used in this research. Finally, we report results from experiments conducted on this dataset and summarize our core findings, contribution and offer suggestions for future research.

2. The LDA model

Latent Dirichlet allocation (LDA) is a generative probabilistic model first presented in the context of topic models [6]. Similar to the original model [6] we use a three level hierarchical Bayesian structure where each handwritten document is modeled as a random mixture over a set of finite handwriting styles, which in turn is modeled as a mixture over an underlying set of text-independent feature probabilities. Thus, in a sense, each writer's handwriting (handwritten document) is represented as a distribution over *latent* handwriting styles that is automatically learned from an underlying distribution over text-independent features extracted for each document.

In our implementation of LDA (*Figure 1*):

- A *document* is a sequence of N features denoted by $\boldsymbol{f} = (f_1, f_2, ...f_N)$.
- A *corpus* is a collection of M documents denoted by $D = \{\boldsymbol{f_1}, \boldsymbol{f_2}, ...\boldsymbol{f_M}\}$

The generative process in this model is:

1. Pick a k-dimensional style mixture θ with probability $p(\theta|\alpha) \sim$ Dir(α).
2. For each feature:
   (a) Pick a writing style w_n with probability $p(w_n|\theta) \sim$ Multinomial(θ).
   (b) Pick a feature f_n from $p(f_n|w_n, \beta)$, which is also a multinomial probability distribution conditioned on the writing style w_n.

Given the parameters α and β, the joint distribution of a style mixture θ, a set of writing styles $\boldsymbol{w}$ and a set of features $\boldsymbol{f}$ is given by:

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

The probability of observing a document $\boldsymbol{f}$ = $(f_1, f_2, ...f_N)$ is obtained by marginalizing over the style mixture θ and handwriting styles w:

$$p(\boldsymbol{f}|\alpha, \beta) = \int p(\theta|\alpha) \Bigg(\prod_{i=1}^{N} \sum_{w_i} p(w_i|\theta) \, p(f_i|w_i, \beta) \Bigg) d\theta \quad (2)$$

where p($\theta|\alpha$) is the Dirichlet Prior given by:

$$p(\theta|\alpha) = \frac{\Gamma\Bigg(\sum_{i=1}^{k} \alpha_i \Bigg) \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \quad (3)$$

There are two problems that are required to be solved in this model. One is the inference problem of calculating the posterior distribution of the hidden variables given a document i.e. $p(\theta, \boldsymbol{w}|\boldsymbol{f}, \alpha, \beta)$. The second is the parameter estimation problem of determining the corpus level parameters α and β that, given a corpus $D = \{\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_M}\}$, maximize the log likelihood of the data. Thus, the number of parameters to estimate in this model are *k* parameters $\alpha_1 ... \alpha_k$ for the Dirichlet distribution and $|V| - 1$ parameters for each of the k style models, where V is the vocabulary size of the text-independent features (feature vector dimension in this case). These parameters are estimated using the variational inference algorithm.

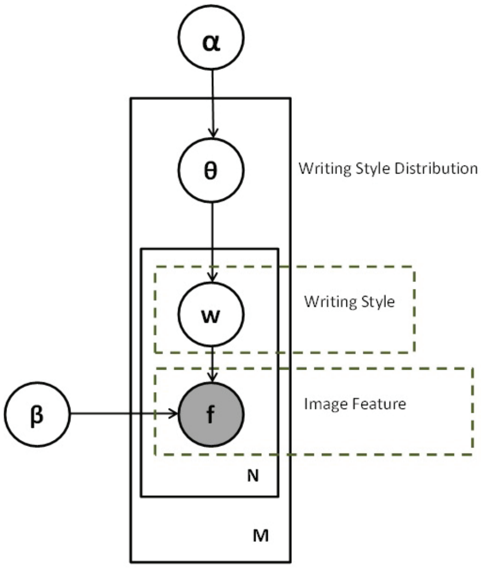
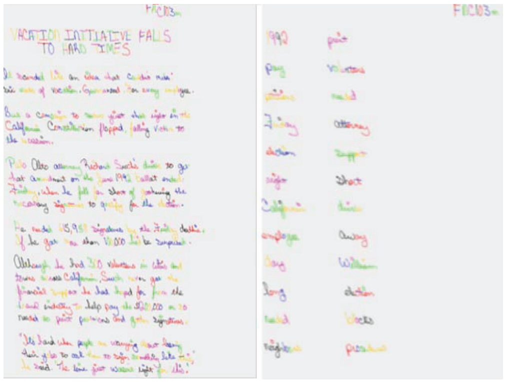


Figure 1. Latent Dirichlet Allocation Model for modeling handwriting styles [4]


FAC103m

VACATION INITIATIVE FALLS
TO HARD TIMES

He sounded like an idea that couldn't miss: six weeks of vacation, guaranteed, for every employee.

But a campaign to enshrine that right in the California Constitution flopped, falling victim to the recession.

Palo Alto attorney Richard Smith's drive to get that amendment on the June 1992 ballot ended Friday, when he fell far short of gathering the necessary signatures to qualify for the election.

He needed 615,958 signatures by the Friday deadline. If he got more than 10,000 he'd be surprised.

Although he had 300 volunteers in cities and towns across California, Smith never got the financial support he had hoped for from the travel industry to help pay the $200,000 he needed to print petitions and gather signatures.

"It's hard when people are worrying about losing their jobs to ask them to sign something like this," he said. "The time just wasn't right for this."

FAC103m

1992 point
pay volunteers
petitions needed
Friday attorney
election support
signer short
California drive
employee away
day William
long election
needed weeks
neighbors president

Figure 2. Example summary-query text document

3. IBM_UB_1 Handwriting Dataset

University at Buffalo (Center for Unified Biometrics and Sensors - *CUBS*) is releasing part of a dual (online + offline) handwriting dataset [1] that has been created from raw data that was originally collected by IBM and donated to the University at Buffalo. This corpus contains online handwriting data, collected on the Cross-Pad, along with their corresponding offline pages.

The online data, presented in a standardized XML format - InkML [2], contains the trajectory information of pen tip on paper as a sequence of x, y coordinates sampled over time. They also contain meta information of the data as XML annotations. The hardcopy of these handwritten documents are scanned into 300 dpi grayscale TIFF images forming their offline counterpart.

The dataset contains handwritten documents in English from 43 writers. A set of 10 topic scripts were generated at random and for each document written by a specific writer, there is a summary text and a corresponding query text (*Figure 2*). The summary text contains one or two pages of writing on a particular topic, while the query text contains approximately 25 words that encapsulate the summary text. Each summary-query pair is labeled with a unique ID in the top right corner that can be used to verify the correspondence between them. Additionally, ground truth information is available for the online query text documents at the word level. The current release will give a page level correspondence between the online and offline documents.

The summary text documents have been used for building and testing our model.

4. Experimental Setup and Evaluation

4.1 Feature

In this paper we start with the assumption that each writer's handwriting is an individual-specific combination of a finite set of handwriting styles. Thus, we looked for a feature that best captures the commonalities shared across writers (in terms of the writing styles - *slant, loopiness etc.*) without losing information about the writer-specific idiosyncratic combination of these styles (*degree of slant, amount of loopiness etc.*). The edge-hinge distribution feature proposed by Bulacu et al [8][9] tries to characterize the changes in the writing direction adopted by a writer. Transitions in writing directions on one hand, can model the *shared component* of an individual's writing (writing style) and on the other hand, be granular enough to capture individual specific variation.

In our paper we use the edge-hinge feature as a starting point and make modifications appropriate for online handwriting. Since, online data comprises of a series of points sampled over time it is neither necessary nor optimal to look at a neighborhood surrounding each point. Hence, instead of placing squares over each pixel, we construct hinges using the adjacent points for every pen-tip location. This feature - adjacent-point hinge - is calculated using three points. Specifically, for every point, we calculate two angles ϕ_1 and ϕ_2. ϕ_1 is the angle that the stroke connecting the current point p_i and the subsequent point p_{i+1} makes with the horizontal. ϕ_2 is the angle that the stroke connecting the current point p_i and the previous point p_{i-1} makes with the horizontal. These angles are then binned into a two dimensional array which is then normalized to give the joint probability distribution of the angles $p(\phi_1, \phi_2)$. The pen-tip locations that govern ϕ_1 and ϕ_2 are assumed to have Markovian properties i.e., the location of each point depends only on the previous point. Further, the size of these bins determines our feature vocabulary (feature vector dimension). For example, if we choose a bin size of 15 degrees the adjacent-point hinge feature has 300 components.

4.2 Analysis and Results

We conducted three experiments in which we systematically examine (a) the effect of feature vocabulary length on model performance, (b) the effect of writing style specification on model performance, and, (c) our model's performance on a different task, namely, writer verification. We present the procedure and results of our studies below.

With regard to the two writer identification studies (Study 1 and 2), following the procedure outlined earlier, we construct the text-independent feature probability distribution for all of the documents in our dataset. Subsequently, we apply the LDA model [3] to a training set (subset of the data) to learn the latent handwriting styles. At this stage, we *a priori* fix the number of handwriting styles to learn. Thereafter, the model generates a distribution of these latent styles for individual writers in the training set, which is then used to train an *n*-class SVM (where *n* is the number of writers). Using these learned latent styles, we generate the style distributions for writers in the test set and use the trained SVM to identify individual writers. Our primary criterion for model evaluation is the percentage of correctly identified writers.

4.2.1 Study 1: Feature Vocabulary Length

In this study we vary the feature vocabulary length across three levels. This is operationalized by changing the bin width for our feature array. Specifically, we fix bin width at 12, 15 and 18 degree angles. For each bin width level we evaluate the performance of LDA/SVM model against the performance of a baseline, direct feature/SVM model. Results are presented in Table 1. In essence, we find that decreasing the feature vocabulary length i.e., increasing the bin width (within the range reported) appears to have little effect in the baseline model but improves the LDA model's performance steadily (*Figure 3*). In fact, at the highest bin width in our range, the LDA model outperforms the baseline model. This pattern is consistent with the idea of shared handwriting styles. As the granularity of the feature vocabulary increases, information regarding the shared component of handwriting may be lost.

Table 1. Feature Vocabulary Length Study

Model	Bin width 12	Bin width 15	Bin width 18
Baseline	84.39%	84.92%	84.66%
LDA + SVM	82.83%	84.13%	86.22%

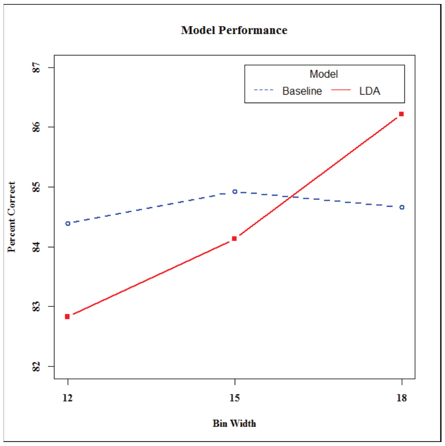


Figure 3. Study 1: Feature Vocabulary Length

4.2.2 Study 2: Handwriting Styles

In our approach we model writers using a finite subset of shared handwriting styles. In the absence of a theoretical basis for predicting the number of handwriting styles, we follow an empirical approach in investigating this issue. For our data, we fix the number of handwriting styles at six different levels and evaluate model performance at each level. We tabulate our findings in Table 2. We find that model performance appears to asymptote at higher levels of handwriting styles (*Figure 4*). This supports our conjecture that a limited set of handwriting styles may adequately describe a larger superset of writers. Increasing the number of handwriting styles beyond a point (e.g., from 20 to 30 in our test range) leaves model performance relatively unchanged.

Table 2. Handwriting Styles Study

Number of styles	Accuracy (%)
10	86.22
13	88.16
16	86.34
20	89.47
26	89.33
30	89.33

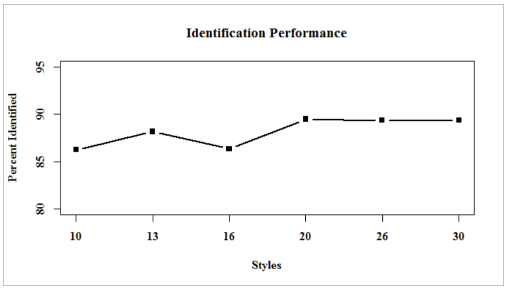


Figure 4. Study 2: Handwriting Styles

Table 3. Writer Verification Study

Model	10 styles	20 styles	30 styles
LDA + SVM	84.87%	87.13%	86.13%

4.2.3 Study 3: Writer Verification

In this study we apply our model to the task of writer verification. Our experimental setup follows the method outlined in [15] and [7]. We evaluate model performance in terms of writer verification across three levels of handwriting styles. The number of handwriting styles were fixed at levels identical to those undertaken in Study 2 in order to provide an appropriate platform for comparison. Results from Study 3 are presented in Table 3. We find a pattern similar to that seen in Study 2, i.e., we get the best performance at the middle of our range of handwriting styles.

As can be seen from tables 1, 2 and 3, our model achieves a peak identification performance of 89.47% and a peak verification performance of 87.13%. Both these levels of performance are achieved with 20 handwriting styles and a bin width of 18. It must be highlighted that this performance - comparable to existing benchmarks - was achieved by relying on the significantly simpler, adjacent-point hinge feature distribution.

5 Conclusion

Our core theoretical contribution in this paper is twofold - (a) we develop a new conceptualization for modeling an individual's handwriting, and, (b) we map the LDA writer-style model to our conceptual framework and extend it to the task of online writer identification and verification. Our studies present preliminary evidence supporting the theory of shared handwriting styles and underscore the efficacy of explicitly modeling them. In addition, this paper also makes a substantive contribution to the field by presenting a novel dataset - IBM_UB_1 - on which our model has been applied and tested. That said, a promising line of inquiry for future research could be in analytically determining the relationship between handwriting styles and number of writers. While we approached this problem empirically in the current investigation, future research may delve into this in more detail.

References

[1] IBM_UB_1Dataset.http://cubs.buffalo.edu/HWdata/.

[2] InkMarkupLanguage.http://www.w3.org/TR/InkML/.

[3] LatentDirichletAllocation.http://www.cs.princeton.edu/~blei/lda-c/.

[4] 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*, pages 265 –270, nov. 2010.

[5] A. Bhardwaj, M. Reddy, S. Setlur, V. Govindaraju, and R. Sitaram. Latent dirichlet allocation based writer identification in offline handwriting. In *Document Analysis Systems*, pages 357–362, 2010.

[6] D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[7] A. Brink, L. Schomaker, and M. Bulacu. Towards explainable writer verification and identification using vantage writers. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 824 –828, sept. 2007.

[8] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):701 –717, april 2007.

[9] M. Bulacu, L. Schomaker, and et al. Writer identification using edge-based directional features. In *IN PROC. OF ICDAR 2003 [SUBMITTED], 2003*, pages 937–941. IEEE Computer Society, 2003.

[10] B. Li and T. Tan. Online text-independent writer identification based on temporal sequence and shape codes. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pages 931 –935, july 2009.

[11] T. Matsuura and P. Thumwarin. On-line writer identification method based on fir system characterizing pentip movement. In *Signals and Electronic Systems, 2008. ICSES '08. International Conference on*, pages 201 –204, sept. 2008.

[12] A. Schlapbach. *Writer Identification and Verification*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 2008.

[13] A. Schlapbach, M. Liwicki, and H. Bunke. A writer identification system for on-line whiteboard data. *Pattern Recognition*, 41(7):2381 – 2397, 2008.

[14] L. Schomaker. Advances in writer identification and verification. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 1268 –1273, sept. 2007.

[15] S. Srihari, M. Beal, K. Bandi, V. Shah, and P. Krishnamurthy. A statistical model for writer verification. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 1105 – 1109 Vol. 2, aug.-1 sept. 2005.

[16] G. Tan, C. Viard-Gaudin, and A. Kot. Online writer identification using fuzzy c-means clustering of character prototypes. In *International Conference on Frontiers in Handwriting Recognition*, page 6, 2008.

[17] M.-Y. Tsai and L.-S. Lan. Online writer identification using the point distribution model. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, pages 1264 –1268, oct. 2005.