2014 22nd International Conference on Pattern Recognition

Data Sufficiency for Online Writer Identification: A Comparative Study of Writer-Style Space vs. Feature Space Models

Arti Shivram, Chetan Ramaiah and Venu Govindaraju
Department of Computer Science and Engineering
University at Buffalo - SUNY
Buffalo, New York 14260
Email: {ashivram, chetanra, govind} @ buffalo.edu

***Abstract*—A key factor in building effective writer identification/verification systems is the amount of data required to build the underlying models. In this research we systematically examine data sufficiency bounds for two broad approaches to online writer identification - feature space models vs. writer-style space models. We report results from 40 experiments conducted on two publicly available datasets and also test identification performance for the target models using two different feature functions. Our findings show that the writer-style space model gives higher identification performance for a given level of data and further, achieves high performance levels with lesser data costs. This model appears to require as less as 20 words per page to achieve identification performance close to 80% and reaches more than 90% accuracy with higher levels of data enrollment.**

I. Introduction

Recent years have seen widespread growth in touchscreen form factors such as tablets and smartphones. These technologies are now well integrated into the fabric of everyday life, whether be it in the workplace or personal-userspace, making person identification in the online domain a pertinent problem to investigate. Further, person identification also forms the first step towards building a number of downstream applications such as added security layers or user-based adaptations in collaborative or multi-user spaces. Since handwriting is a natural way of text input in these environments, writer identification/verification can be a viable method for establishing or verifying a user's identity.

A key factor in building effective identification/verification systems relates to the amount of data required to build the underlying models [1]. Determining data sufficiency bounds therefore, provides a theoretical basis for systematically investigating and comparing different models. Further, from an application standpoint, it can help guide efforts towards building better feature functions and/or efficient baseline models that adapt to the task at hand.

Prior works on the data sufficiency question investigate this issue in the context of offline (scanned images) data [1] [2]. Nonetheless, the findings from these studies need not translate to the online context (pen-tip trajectories) as the nature of the data and consequently the features or approaches used in both contexts vary substantially. Though a large body of literature exists for online writer identification and verification methods, to the best of our knowledge no formal study has been conducted on the amount of online data required to build them. In order to bridge this gap, we undertake an empirical data enrollment study comparing two broad approaches to online writer identification/verification - feature-space vs. writer style-space approaches.

In subsequent sections we first provide a description of the two broad approaches outlined above, following which we lay out the the enrollment study design and procedure. We report results from 40 experiments across 2 datasets and conclude with our findings and recommendations for future research and practice.

II. Feature Space vs. Writer Style Space Approaches

A. Feature Space Models

The main objective of writer identification is to distinguish between writers using their handwriting as the discriminating characteristic. Since handwriting forms the basis of discrimination, in the feature space model each individual's handwriting is taken to represent a unique writing style intrinsic to that writer. Thus, distinct feature-space representations are built for each writer with the underlying assumption that each writer possesses his/her singular handwriting style that is not shared with other writers and that the feature-space fully and completely defines each style and consequently, each writer [3] (Figure 1).

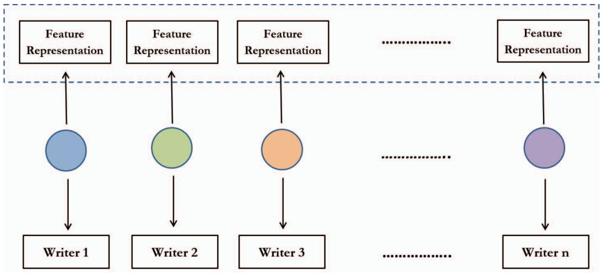


Fig. 1. Conceptual Model: Feature Space Approach

B. Writer Style Space Models

The writer-style space model departs from the above approach in the core assumption that although *handwriting* is

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.538

unique among writers, writing styles are not, in that, there are stylistic commonalities across writers (degree of slant, loopiness and so on).

Handwriting theory lays out two major factors that influence the way a person writes - genetic factors (unique to an individual) and memetic or cultural factors (shared amongst individuals) [4]. Thus, in the writer-style framework, writing styles represent a *shared* component of an individual's handwriting. In other words, a person's handwriting can be *a priori* conceptualized as an individual-specific combination (determined by a persons physiology - genetic factors) of a shared pool of writing styles (often determined culturally - memetic factors) [3] (Figure 2).

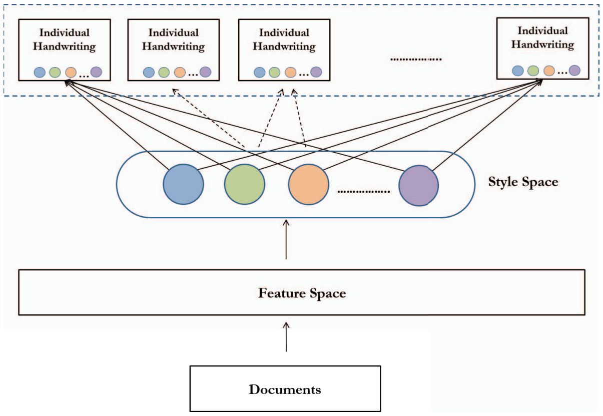


Fig. 2. Conceptual Model: Writer-Style Space Approach

Recent works explicitly model the above conceptualization by using a three level hierarchical Bayesian structure - Latent Dirichlet allocation (LDA) [3] [5]. In these frameworks each handwritten document is represented as a distribution over a set of finite (shared) writing styles, which in turn are modeled as a distribution over an underlying set of text-independent feature probabilities [3] [6] [7] (Figure 3).

Not only does the writer-style framework present a theoretically grounded representation of handwriting, recent empirical findings also suggest that this framework offers parsimony in parameterization [3]. Specifically, a relatively smaller set of parameters can successfully model a significantly larger superset of writers. Thus, explicitly modeling the shared component of human handwriting appears to yield greater discriminatory power to identification systems.

We aim to extend this emerging literature by examining whether such parsimony extends to the data sufficiency question as well. Since most online systems involve real-time user interaction, the lesser the amount of data required from a user, the more user-friendly the application is. Also, a robust system needs to achieve high identification performance even with less data. Thus, tighter data sufficiency bounds translate directly to more fluid user interactions as well as more effective systems in the online domain.

This provides the primary motivation for the data enrollment study reported in the next section. The study was conducted using two different publicly available online handwriting datasets - IBM_UB_1 [8] and IAM-OnDB [9].

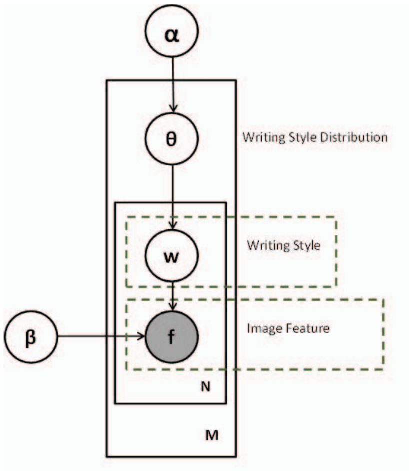


Fig. 3. Writer-Style LDA Model

III. Enrollment Study: Design and Procedure

A. Design Overview

The enrollment study was set up such that writer identification performance was assessed at different levels of handwriting data. Specifically, identification performance for the feature and writer-style models were repeatedly tested while gradually increasing the amount of text (*number of words*) per handwritten page. The number of words per page varied between 10 and 100 (in increments of 10) while keeping other model-related parameters constant. Specifically, the writer-style LDA model parameters - feature vocabulary length and number of writing styles - were fixed per prior empirical studies [3] at levels that maximized baseline identification performance. Similarly, for the feature-space models we chose feature sets that have been shown to provide state-of-the-art identification performance on the datasets used in this study [8] [10]. An n-class SVM trained using the LIBSVM [11] toolkit forms the classifier for both the style-space and feature-space models.

To summarize, in this overall data sufficiency investigation, we conducted experimental runs on two publicly available datasets, using two different sets of features each, across ten different levels of data availability. This yielded a total of 40 empirical studies comparing identification performance between the feature-space and writer-style space models.

This exhaustive set of comparisons enables us to construct factorial studies examining the interaction between (a) the type of dataset, (b) the type of feature and, (c) the type of model, in determining how much data is required for effective identification performance.

B. Feature Extraction

Since the domain of investigation involves online, temporal pen-tip trajectory data, we use two sets of point-based features for our experiments. First, we use the adjacent-point hinge feature [3] and second, we use what we term the DDC features (i.e., distance, direction and curvature). We describe these features briefly below:

1) Adjacent-Point Hinge Feature: Since, online data comprises of a series of points sampled over time, in the adjacent-point hinge feature, hinges are constructed using adjacent points for every pen-tip location [3]. Thus, this feature is calculated using three points. Specifically, for every point,

- Two angles ϕ_1 and ϕ_2 are calculated,
- 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 binned into a two dimensional array (bin-width is 18 degrees) 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.

2) DDC Features: In their work on writer identification Schlapbach et al. [10] describe a procedure to extract point- and stroke-based features. We follow their protocol to extract a collection of three point based features, described below:

- Distance: The distance between point $\boldsymbol{p_i}$ and $\boldsymbol{p_{i+1}}$.

$$\boldsymbol{l_i = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}} \quad (1)$$

- Writing direction: The cosine and sine of the angle the line from point $\boldsymbol{p_i}$ to $\boldsymbol{p_{i+1}}$ makes with the horizontal.

$$\boldsymbol{\cos(\theta_i) = \frac{\Delta\ x(p_i, p_{i+1})}{l_i}} \quad (2)$$

$$\boldsymbol{\sin(\theta_i) = \frac{\Delta\ y(p_i, p_{i+1})}{l_i}} \quad (3)$$

- Curvature: The angle formed by the lines from $\boldsymbol{p_{i-1}}$ to $\boldsymbol{p_i}$ and $\boldsymbol{p_i}$ to $\boldsymbol{p_{i+1}}$.

$$\boldsymbol{\cos(\phi_i) = \cos(\theta_i)\cos(\theta_{i+1}) + \sin(\theta_i)\sin(\theta_{i+1})} \quad (4)$$

$$\boldsymbol{\sin(\phi_i) = \cos(\theta_i)\sin(\theta_{i+1}) - \sin(\theta_i)\cos(\theta_{i+1})} \quad (5)$$

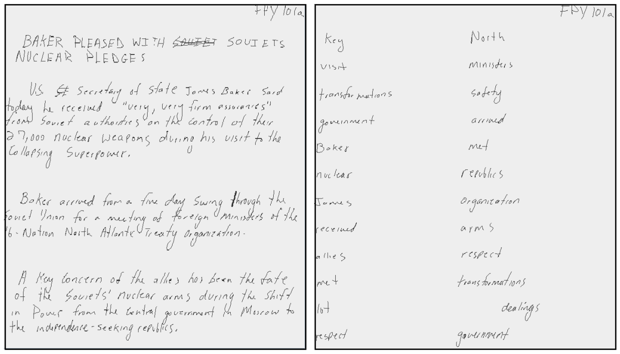
These DDC features are extracted from each point, and each of these are binned in separate histograms. The histogram bin widths are heuristically selected. The length of the point level feature vector is 900.

C. Datasets

1) IBM_UB_1: University at Buffalo (Center for Unified Biometrics and Sensors - *CUBS*) has released a dual (online + offline) handwriting dataset [8] 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 CrossPad, along with their corresponding offline pages.

The online data, presented in a standardized XML format - InkML [12], 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 that are organized in a twin-folio structure [8]. 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 4). 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 that is used to verify the correspondence between them. For this research, we have utilized only the summary text documents for building and testing our model. Out of the 4138 summary pages, 80% were used for training and 20% formed the test set.



FPY 101a

BAKER PLEASED WITH ~~SOVIET~~ SOVIETS
NUCLEAR PLEDGES

U.S. ~~St~~ Secretary of State James Baker Said today he received "very, very firm assurances" from Soviet authorities on the control of their 27,000 nuclear weapons during his visit to the Collapsing Superpower.

Baker arrived from a five day Swing through the Soviet Union for a meeting of foreign ministers of the 16-Nation North Atlantic Treaty Organization.

A Key concern of the allies has been the fate of the Soviets' nuclear arms during the Shift in Power from the central government in Moscow to the independence-seeking republics.

FPY 101a

Key
visit
transformations
government
Baker
nuclear
James
received
allies
met
lot
respect

North
ministers
safety
arrived
met
republics
organization
arms
respect
transformations
dealings
government

Fig. 4. Twin-Folio Structure of IBM_UB_1

2) IAM-OnDB: IAM-OnDB is another publicly available dataset that contains online handwritten text. This online data was written on a whiteboard and collected using an electronic interface [9]. It contains 1700 handwritten forms collected from 221 writers. In total it contains 86,272 word instances, 13,049 text lines and a lexicon of about 11,059 words [9].

In this dataset the online data is presented in an XML format and also contains metadata information about the writers such as native country, language, handedness and so on. All texts included in this database are taken from the LOB corpus [9].

For our experiments we use the standardized train/test split provided for this dataset.

IV. Experimental Results

Recapping briefly, our experimental design involved 40 comparative studies between the feature space model and writer-style space model, run on two datasets using two different sets of features each. The identification performance results from these studies are tabulated across Tables I and II. This data allows us to make specific comparisons and draw insights upon how the type of feature and the type of model interact to determine the effect of enrollment data upon identification performance (Figure 5).

TABLE I. Model Comparisons By Feature Type: IBM_UB_1

IBM_UB_1				
Number of Words	Adjacent-Hinge		DDC	
	Feature Space	Style Space	Feature Space	Style Space
10	64.2	68.44	74.18	72.4
20	72.13	78.96	82.65	81.83
30	77.59	83.6	85.38	85.65
40	80.46	82.37	87.56	86.74
50	82.1	83.6	88.38	89.48
60	78.68	85.38	88.79	87.29
70	81.83	84.15	89.89	87.97
80	84.42	87.56	89.2	90.02
90	83.19	86.74	85.38	86.74
100	82.1	85.38	88.38	88.38

TABLE II. Model Comparisons By Feature Type: IAM-OnDB

IAM-OnDB				
Number of Words	Adjacent-Hinge		DDC	
	Feature Space	Style Space	Feature Space	Style Space
10	57.09	62.13	60.29	77.71
20	70.52	76.56	61.28	78.32
30	74.65	80.96	71.35	79.33
40	73.26	83.65	80.81	81.65
50	75.18	84.23	79.46	81.06
60	82.67	87.33	81.89	81.89
70	78.35	87.88	83.13	82.08
80	80.46	85.72	82.67	83.63
90	78.42	89.31	85.3	87.18
100	83.5	88.92	87.23	90.31

A. Data Sufficiency Study

Identification performance results suggest that, in general, both models' performance improves as the enrolled data increases beyond 10 words per page. As evident from the table, there is a substantial improvement in performance from 10 to 20 words per page across all model/feature combinations in both datasets. The improvement thereafter slows down. More importantly, the improvement trend diverges for the writer-style vs. feature-space models. We find that the writer-style model appears to asymptote for much lesser number of words per page (between 20/30 words) as compared to the feature-space model, which in certain instances requires as much as 60 words per page to plateau. Thus, across both datasets and features the writer-style model requires lesser data enrollment to achieve sufficient identification performance; as can be seen from Tables I and II, this model requires just around 20 words per page to achieve close to 80% accuracy.

B. Identification Performance

In this experiment we compare the identification performance of the two approaches averaged across both datasets and feature types. Doing so helps compare and evaluate the *overall* performance of each approach at different levels of data irrespective of feature type or dataset type. Our analysis shows that the writer-style model consistently outperforms the feature space model at each level of data enrollment (Figure 6). At the 20 word per page data enrollment level, the writer style model attains an identification performance of approximately 79% whereas the feature space model's performance is at approximately 72%. Similarly, their respective performances are at 88.25% and 85.3% at the 100 word per page level.

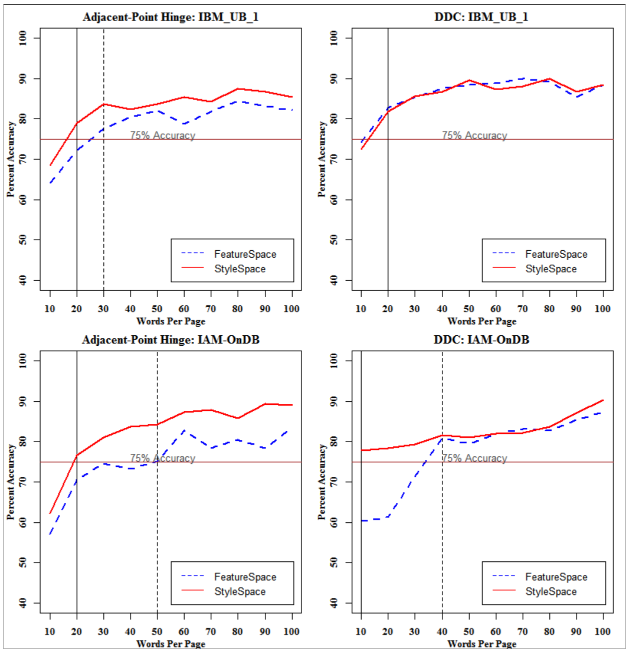


Fig. 5. Model Comparisons By Dataset and Feature Type

C. Sensitivity to Feature Type

In this experiment, we calculated the difference in performance for the two approaches when using the adjacent-point hinge feature vs. the DDC features. This was calculated at each data enrollment level. Comparing these differences enables us to judge how stable a given approach's identification performance is to the choice of features. Since the identification performance for both approaches reliably crosses the 60% threshold only after the 20 words-per-page mark, we conduct this experiment for 9 levels of data enrollment (20-100).

We find that in the case of IBM_UB_1, the average change in identification performance (absolute value) for the feature space approach is 7.01% (standard deviation: 2.57%). In comparison, the writer-style space approach averages 2.93% change in performance (standard deviation: 1.66%). Further, a similar pattern emerges for the IAM-OnDB dataset as well. The feature space approach averages 4.75% change in performance (standard deviation: 2.69%) whereas the writer-style space approach averages a 2.82% change (standard deviation: 1.66%). This shows that the writer-style space approach is more stable and robust to variation in the choice of features used in the algorithm in comparison to a pure feature-space based approach.

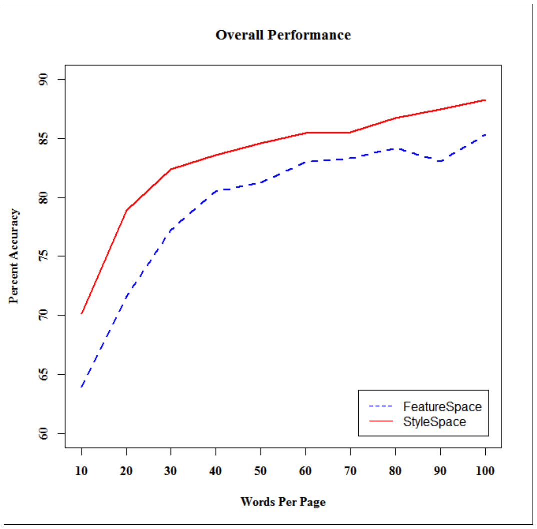


Fig. 6. Overall Model Performance Plots

V. Conclusion

In this research we undertook a formal data enrollment study for online writer identification. We compared two broad approaches to writer identification - feature space vs. writer-style space approaches. To this end, we studied the difference in performance levels between a writer-style space and a feature space model. This study was undertaken across two different datasets and used two different feature types. We found that the writer-style space model not only consistently outperformed the feature space model for both datasets and feature types (in terms of peak and overall identification performance), it also needed lesser data to meet or exceed comparable performance levels exhibited by the feature space model. In terms of data sufficiency, the writer-style space model required just around 20 words per page to achieve close to 80% accuracy.

Further, an examination of the sensitivity of these models to feature types revealed that the writer-style space model is relatively robust to the choice of features. Again, this was evidenced in both datasets. The change in performance levels on account of a change in features was lesser for this model type than that shown by the feature space model.

To conclude, at the outset we had asked the question of whether the writer-style framework's parsimony in modeling translates to a parsimony in data-sufficiency. Our findings clearly show that not only does this approach give higher identification performance for a given level of data, it achieves high performance levels with lesser data costs. Thus, it provides a promising avenue to develop identification related applications in future.

Acknowledgment

This work was partially funded by the NSF Center for Identification Technology Research (CITeR) (NSF IIP Award #:1266183)

References

[1] A. Brink, M. Bulacu, and L. Schomaker, "How much handwritten text is needed for text-independent writer verification and identification," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[2] H. Srinivasan, S. Kabra, C. Huang, and S. Srihari, "On computing strength of evidence for writer verification," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 844–848.

[3] A. Shivram, C. Ramaiah, U. Porwal, and V. Govindaraju, "Modeling writing styles for online writer identification: A hierarchical Bayesian approach," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, Sept. 2012, pp. 385 –390.

[4] L. Schomaker, "Advances in writer identification and verification," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, sept. 2007, pp. 1268 –1273.

[5] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

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

[8] A. Shivram, C. Ramaiah, S. Setlur, and V. Govindaraju, "IBM_UB_1: A dual mode unconstrained English handwriting dataset," in *Document Analysis and Recognition, 2013. Proceedings of the Twelfth International Conference on*, 2013, pp. 13–17.

[9] M. Liwicki and H. Bunke, "IAM-OnDB – an On-Line English Sentence Database Acquired from Handwritten Text on a Whiteboard," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005, pp. 956–961.

[10] A. Schlapbach, M. Liwicki, and H. Bunke, "A writer identification system for on-line whiteboard data," *Pattern Recognition*, vol. 41, no. 7, pp. 2381 – 2397, 2008.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[12] Ink Markup Language. http://www.w3.org/TR/InkML/.