2011 International Conference on Document Analysis and Recognition

OCR-Driven Writer Identification and Adaptation in an HMM Handwriting Recognition System

Huaigu Cao
Raytheon BBN Technologies
Cambridge, MA, USA
hcao@bbn.com

Rohit Prasad
Raytheon BBN Technologies
Cambridge, MA, USA
rprasad@bbn.com

Prem Natarajan
Raytheon BBN Technologies
Cambridge, MA, USA
pnataraj@bbn.com

***Abstract*—We present an OCR-driven writer identification algorithm in this paper. Our algorithm learns writer-specific characteristics more precisely from explicit character alignment using the Viterbi algorithm and shows significant reduction of close-set writer identification error rates, compared with the GMM-based method. With writers' identities retrieved, we improve the performance of handwriting recognition using the HMM trained adapted on the training data of that writer. In our system, writer identification and OCR are highly interactive. They improve the performance of each other and thus show close approximation of supervised text-dependent writer identification and writer-dependent HMM handwriting.**

Keywords- Handwriting recognition, writer identification, hidden markov model

I. INTRODUCTION

We present an OCR-driven approach to writer identification in handwritten document images. Although both writer identification and HMM handwriting recognition have their own applications (writer identification techniques can be applied to forensic signature verification, whereas handwriting recognition is still a challenge problem in OCR community with limited applications in constrained conditions, and has potential of being applied to automatically transcribing unconstrained handwritten documents in the future), they are in fact two highly interactive problems. On the one hand, writer dependent (WD) HMM systems are known to have significantly better performance than writer independent (WI) HMM systems. On the other hand, with labeled references provided, grouping and aligning instances of the same character become more reliable than unsupervised writer identification methods, and the error rate of writer identities can be reduced. Our idea is motivated by interactivity of writer identification and handwriting recognition. Multiple passes of OCR decoding are deployed in our HMM handwriting recognition system for this purpose (Fig. 1). The input (handwritten documents) is decoded with WI HMM to create preliminary transcriptions, with character boundaries indicated by the optimal HMM state sequence. For each character, writer labels are given by a component classifier of all writers in training. Character-level decisions are further fused at each line or page of text, if applicable. Finally, the input is decoded with the WD HMM of the identified writer.

The following is an overview of related works in writer identification and OCR-related applications. In [1], texture feature and *k* Nearest Neighbor classifiers are applied to writer identification in skew-corrected document images. In [1], image features extracted from macro and micro scales are investigated. The writer similarity score is computed using distance-based measures. With the success of HMM in handwriting recognition, speaker recognition techniques such as GMM [3][4] and GMM-SVM [5] can also be applied to writer identification. In [6], an error rate of 1.5% is obtained from close-set identification of over 650 writers from 1500 pages of the IAM data set [7]. Only a few efforts have been made in using writer identification to improve handwriting recognition or using handwriting recognition to improve writer identification. In [8], a GMM-based writer identification algorithm is applied to selecting WD models for keyword spotting. MAP adaptation is used to build WD models. In our prior work [9], we presented Arabic handwriting recognition using WD HMM created from MAP adaptation, and described a text-independent writer identification algorithm to select WD codebooks.

In this paper, we tested the GMM-based writer identification method using the handwriting recognition system based on WD HMM selection [9] and obtained a significant improvement in writer identification performance and handwriting recognition, compared with the text-dependent writer identification method described in [9]. We implemented the GMM –based writer identification method and showed advantage over the global features-based writer identification method [9]. But the OCR-driven text-dependent writer identification algorithm described in this paper shows substantially lower identification error rates than both of the text-independent methods. The impact of incorrectly identified writers on writer-dependent OCR performance is investigated and is proved to be negligible by our experiments.

[1] This paper is based upon work supported by the DARPA MADCAT Program.

[2] The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.154

II. OCR-driven Writer Identification and Writer Dependent OCR

A. OCR-driven Writer Identification

Our OCR-driven writer identification method is text-dependent, *i.e.*, the reference of handwriting is required to perform identification. Since the transcribed reference is only available for our training data, we decode our test set and take the OCR hypothesis as the reference.

During training, given a line image from the training data and the corresponding reference of the line, first, we get the boundary for every character in the reference. This is done by finding the optimal state sequence for the WI HMM built from the reference using the Viterbi algorithm [12]. Then, we create a component writer classifier for each distinct character using directional element features [13] computed from 16 non-overlapping bins (4×4) of each character image with white space on top and bottom chopped and the Support Vector Machine (SVM) with the radial basis kernel. An SVM classifier only solves two-class problem. For n classes, we need to build $n(n\text{-}1)/2$ binary classifiers and take the label that is obtained for the most of time from these classifiers. Thus, the **component classifier** of a character is defined as the collection of all $n(n\text{-}1)/2$ binary classifiers as a voting system. We train the component classifier of each character on 90% of the training samples of the character, test and keep track of the accuracy of the component classifier using the remaining 10% of training samples.

When applying the SVM classifiers to the test set, first we decode the test set using the WI HMM and create character boundaries using the same method mentioned above. Then, we classify the writer of each character image using SVM component classifiers. Finally, we fuse the output of component classifiers as follows. The writer class of a region of k characters $c_1, c_2, \ldots, c_k$ is given by

$$w^* = \underset{w}{\operatorname{argmax}} \sum_{1 \le i \le k,\ C(c_i)=w} A(c_i) \qquad (1)$$

where $C(c_i)$ is the writer class generated by the component classifier of character c_i, and $A(c_i)$ is the accuracy of the component classifier of c_i estimated from the training set. A region can be a page, a paragraph, a line or even a word that can be assumed to be written by the same person in real-world applications. We only take the 1-best OCR hypothesis. Practically, this works very well. However, Eq. (1) can easily be modified to incorporate arc posterior probabilities from the word lattice.

As you can see from Eq. (1), if we assume that $A(c_i)$ is a constant, the decision becomes a simple voting by the component classification results of all characters in the region. The use of $A(c_i)$ in Eq. (1) takes advantage of the verified performance of component classifiers. Thus, component classifiers of more reliable writer identification performance are associated with higher weights when making the final decision.

Verification becomes a less important problem when our objective is to improve OCR performance since failure to detect a known writer is much more costly than failure to reject an unknown writer. On the one hand, the improvement of the WD HMM over the WI HMM is large once the WD HMM of the "right" writer is selected. On the other hand, we always create WD HMM by means of adapting WI HMM trained on large amount of data to small amount of writer-dependent data. This ensures that, even though the WD HMM is incorrectly selected, it does not necessarily have a lower performance than the WI HMM, which is justified in our experiments.

B. Writer Identification Using GMM

The GMM-based writer identification algorithm [3][8] is also evaluated in our OCR system. First, we train a GMM of 2048 Gaussian components from all writers' features using 3 iterations of the Expectation-Maximization (EM) algorithm as the universal background model (UBM). Then, we train a separate GMM for each writer using 2 iterations of EM with the UBM as the initial parameters of the EM algorithm. Features we use to train GMM are 17-dimensional LDA features projected from 3-frame concatenated various image features in the HMM handwriting recognition system [14] with two small modifications: features based on Gabor filters are added to the system and the dimensionality of LDA features is increased from 15 to 17 for better OCR performance.

The log-likelihood of GMM λ for a sequence of feature vectors $X = \{x_1, x_2, \ldots, x_N\}$ is computed as

$$\log p(X|\lambda) = \sum_{t=1}^{N} \log p(x_t|\lambda)/N \qquad (2)$$

where $p(x_t|\lambda)$ is the likelihood of model λ for feature vector x_t. The denominator N in Eq. (2) compensates for the incorrect assumption of independence of frames. This is useful for verification purpose. For identification, it is a constant over all classes and is not very useful. The writer who gives the highest log-likelihood is selected as the identified writer.

C. HMM Adaptation Techniques

In most of the time, we do not have sufficient training data for each writer to train a WD HMM from scratch. Instead, we need to train the WI HMM from the entire training set of various writers first, and adapt the WI HMM to each target writer. We use the Maximum A Posteriori (MAP) adaptation algorithm to adapt the GMM mean vectors of the WI HMM. We do not update the GMM covariance and transition probabilities. In Map adaptation, instead of maximizing the auxiliary function $Q(\lambda, \hat{\lambda})$ as in the case of Maximum Likelihood (ML) HMM Training, a modified function $R(\lambda, \hat{\lambda})$ with one more component $\log G(\lambda)$ representing the prior distribution of HMM parameters is maximized [10]:

$$R(\lambda, \hat{\lambda}) = Q(\lambda, \hat{\lambda}) + \log G(\lambda). \qquad (3)$$

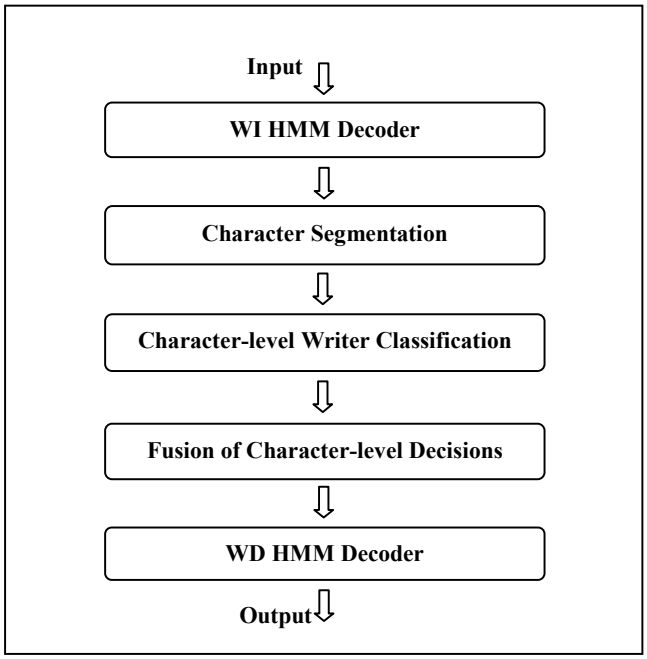

Figure 1. OCR-driven writer identification and WD HMM selection.

Figure 2. A sample page from our training corpus.

The updating equation of the mean of the k-th Gaussian component is

$$\hat{\mu}_k^{MAP} = \frac{\tau_k \mu_k^{WI} + c^k \hat{\mu}_k^{ML}}{\tau_k + c_k}, \quad (4)$$

where μ_k^{WI} is the k-th WI GMM mean, τ_k is the prior counting weight of the WI mean, $\hat{\mu}_k^{ML}$ is the updated mean vector using maximum likelihood estimation, and c^k is the occupancy of the k-th Gaussian component.

Maximum Likelihood Linear Regression (MLLR) [11] is another commonly used adaptation technique. In MLLR, an adapted mean of GMM is assumed to be an affine transformation of the original mean:

$$\hat{\mu}^{MLLR} = A\mu^{WI} + b, \quad (5)$$

and estimated using the EM algorithm.

MAP adaptation has more parameters to estimate than MLLR adaptation. MAP adaptation is more suitable for adaptation with large amount of training data than MLLR adaptation, *e.g.*, when thousands of word images of the same writer are available. MLLR is more suitable for adapt HMM to a single page of manuscript composed of a hundred word images or so.

III. Experimental Results

A. Overview of the HMM Handwriting OCR System and Data Corpus

We used the HMM handwriting system [14] to test our method. In the system, a 14-state HMM is defined for each character with the left-to-right configuration and at most one skipped state is allowed in state transition. A preliminary Writer-independent character-tied HMM of 512-GMM was trained and applied to creating state labels of training data that we needed to estimate the LDA feature transform. With the estimated LDA matrix, a state-tied HMM of 1500 Gaussians per state (GMM) was trained and applied as our WI HMM to decoding test data. For each writer in training data, a state-tied WD HMM of 1500 Gaussians per state was created by 2 iterations of MAP adaptation of the WI HMM. Selection of WD HMM is done using either the proposed OCR-driven writer identification algorithm or any text-independent writer identification algorithm. Unsupervised MLLR adaptation and duration adaptation [9] are performed per page using the OCR hypotheses obtained from WD decoding. This seems to be a very complex scheme of multiple adaptation methods. But, we can evaluate if the proposed method will still make the improvement with other similar techniques existing in a practical system.

We tested our method on Arabic handwritten documents. We collected 37,608 8.5x11-inch pages of Arabic handwriting written by 259 people as our training set. There were about 100 words in 15-20 lines in each page. Each writer contributed to 50 to 250 pages of these documents. All pages were transcribed and locations of words were marked. These pages were scanned monochromatically in 600 dpi TIFF lossless format. We also collected over 1,300 pages of similar style as our development and test sets. We used a development set of 868 pages for optimizing the balance between glyph HMM and language modeling scores. Line boundaries created by manual annotation are used in our experiments since we do not focus on the line finding algorithm in this paper. A typical sample page from our training corpus is shown in Fig. 2.

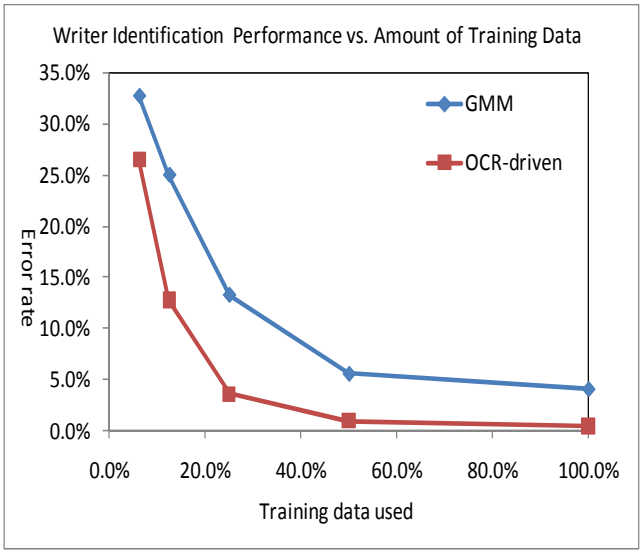

Figure 3. Impact of percentage of training data used on writer identification accuray.

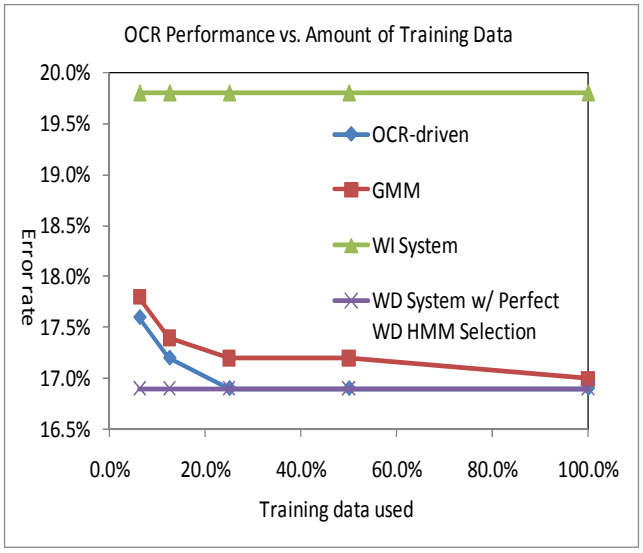

Figure 4. Impact of percentage of training data used on WD OCR performance.

The language model used in our HMM handwriting recognition system was a word tri-gram trained on Arabic text collection of over 217 million words, with a dictionary of over 300,000 distinct words.

B. *Writer Identification Performance*

We trained our writer identification system using all 259 writers in the training set. We tested the close-set performance of the OCR-driven writer identification method presented in this paper using a test set of 196 pages. Each of these pages was written by one of the writers contributing to the training set. Using the same test set, we compared the page-level writer identification performance of the method using global features and SVM [9], the GMM-based method, and the OCR-driven method. Eq. (2) can easily be modified to represent the log-likelihood of an entire page by summing over the log-likelihood of feature vectors in all frames of the page. Thus, N in Eq. (2) should be the number of frames in the page. The close-set writer identification error rates are shown in Table I. The OCR-driven approach led to tremendous reduction of identification. The GMM-based approach outperformed [9] by 50% relative in the error rate. However, it is still not comparable to the OCR-driven approach.

The impact of amount of training data used on identification performance of the GMM-based method and the OCR-driven method was evaluated by training on randomly selected subsets of pages from the training set. The sampled training data shown in Fig. 3 are 6.25%, 12.5%, 25%, 50%, and 100% of the entire training set, respectively, with the exception that 100% training data used in the OCR-driven method actually means that 90% are for training the SVM and 10% are used for evaluating the SVM. Fig. 3 shows the reduction of error rates by adding more pages to the training set. From the curves in Fig. 3, the OCR-driven method consistently provides higher accuracy than the GMM-based method. Table II shows the average and standard deviation of numbers of pages involved in training for all writer classes in the test set. Fig. 3 and Table II show that the OCR-driven writer identification algorithm is able to produce extremely high accuracy (<5%) when around 20 pages (2000 words) per writer are available for training, and can still maintain reasonably high performance when less training data (200-2000 words per writer) are available.

TABLE I. CLOSE-SET ERROR RATES OF THREE WRITER IDENTIFICATION ALGORITHMS

Methods	Global features & SVM	GMM	OCR-driven
Page-level error rates %	58.2	4.1	0.5

TABLE II. AVERAGE AND STANDARD DEVIATION OF NUMBER OF PAGES PER WRITER USED IN TRAINING

% Sampled	6.25	12.5	25	50	100
Avg. #. pages	5.7	11.4	22.8	45.7	91.3
Std. dev. #. pages	2.2	4.5	9.0	17.9	35.8

TABLE III. OCR PERFORMANCE OF MISMATCHED WD HMM

Model	WI HMM	Mismatched WD HMM
WER %	17.7	17.5

C. *Improvements in OCR Performance*

We evaluated the OCR performance on the same test set we used to test writer identification. Our baseline system performed WI decoding followed by decoding using page-wise MLLR and duration adapted HMMs. The second system performed WD decoding using models selected by the GMM-based writer identification algorithm followed by

the same page-wise adapted decoding. To evaluate the OCR-drive writer identification algorithms, we created the third system that has the following steps:

1) Decode the test image with WI HMM;

2) Perform Writer identification using OCR hypotheses created in the WI HMM decoding stage as the reference;

3) Decode the input image again using the WD HMM selected by writer identification;

4) Perform page-wise MLLR and duration adaptation on the selected WD HMM, and re-decode the input image with the page-wisely adapted HMM.

We used the Word Error Rate (WER) (the number of substituted, inserted and deleted words over the number of words in the reference) to measure the performance of our OCR systems. The OCR word error rates of writer identification models trained with different amount of data are shown in Fig. 4. The green line with triangle markers that is above all other curves shows the performance of the WI system. The purple line with cross markers that is below all other curves shows the performance of the system with WD HMM selected using the ground truth. The blue curve with diamond markers represents the WD HMM system using the OCR-driven writer identification method. The Red curve with square markers represents the WD HMM system using the GMM-based writer identification method. We can see the OCR system using the OCR-driven writer identification consistently outperforms the OCR system using the GMM-based writer identification. When 25% or more training data were used in writer identification, the WER of the OCR system using the OCR-driven writer identification method almost reaches the theoretical bound indicated by the purple line.

We also tested the OCR performance of our approach on 197 pages collected from Arabic handwriting written by people who did not contribute to the training set. We decoded this test set using the WI OCR system and the WD OCR system with the OCR-driven writer identification method, respectively. These two systems were also what we ran to create results in Table II. The performance is shown in Table III. Surprisingly, there was even a small improvement of WER (from 17.7% to 17.5%) from the mismatched WD system. This shows that a verification stage is not necessary since the mismatched WD HMM did not lower the performance of OCR. The reliable performance in the event of incorrectly picked writer labels is mostly owing to the use of adaptation technique that retains the writer-independent prior distribution.

IV. CONCLUSION

We presented an OCR-driven writer identification algorithm in this paper. The text-dependent writer identification, enabled by initial WI model decoding, had distinct advantage over the conventional writer-independent identification approaches. Our approach also showed significant improvement of Arabic handwriting OCR performance when applied to selecting writers for WD decoding. Tested in a system with multiple adaptation techniques integrated, the WD HMM selection and decoding showed its unique contribution to the performance improvements and cannot be substituted with per-page adaptations. The test on unseen writers showed that MAP adapted HMM was able to maintain the same performance as the WI HMM when writers were selected incorrectly.

REFERENCES

[1] Said Peake Tan, H. E. S. Said, G. S. Peake, T. N. Tan, and K. D. Baker, "Writer Identification from Non-uniformly Skewed Handwriting Images," Proc. 9th British Machine Vision Conference, 1998, pp. 478-487.

[2] S. N. Srihari, S-H. Cha, H. Arora and S. Lee, "Individuality of Handwriting," Journal of Forensic Sciences, vol. 47(4), 2002, pp. 856-872.

[3] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication, vol. 17 (1-2), August 1995, pp. 91-108.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, 2000, vol. 10, pp. 19-41,.

[5] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, 2006, vol. 13, pp. 308-311.

[6] A. Schlapbach, H.Bunke, "Off-lineWriter Identification Using Gaussian Mixture Models," Proc. 18th International Conference on Pattern Recognition, 2006, vol. 3, pp. 992-995.

[7] U.-V. Marti and H. Bunke, "The IAM–database: An English sentence database for off-line handwriting recognition," Int. Journal of Document Analysis and Recognition, 2002, vol. 5, pp. 39-46.

[8] J.A. Rodriguez, F. Perronnin, G. Sanchez, and J. Llados, "Unsupervised writer style adaptation for handwritten word spotting," Proc. 19th International Conference on Pattern Recognition, 2008, pp.1-4.

[9] Huaigu Cao, Rohit Prasad, and Prem Natarajan, "Improvements in HMM Adaptation for Handwriting Recognition Using Writer Identification and Duration Adaptation," Proc. International Conference on Frontier of Handwriting Recognition, 2010, pp. 154-159.

[10] J.-L. Gauvain, Chin-Hui Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Transactions on Speech and Audio Processing, Apr 1994, vol. 2 (2), pp. 291-298.

[11] C. J. Leggetter, and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, April 1995, Vol. 9 (2), pp. 171-185.

[12] G. A. Fink, "Markov Models for Pattern Recognition: from Theory to Applications," Springer Press, 2007, pp. 75-76.

[13] N. Kato, M. Suzuki, S. Omachi, H. Aso, and Y.Nemoto, "A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, Mar 1999, vol. 21 (3), pp. 258-262.

[14] Shirin Saleem, Huaigu Cao, Krishna Subramanian, Matin Kamali, Rohit Prasad, Premkumar Natarajan, "Improvements in BBN's HMM-Based Offline Arabic Handwriting Recognition System," Proc. International Conference on Document Analysis and Recognition, 2009, pp. 773-777.