Handwritten and Typewritten Text Identification and Recognition using Hidden Markov Models

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

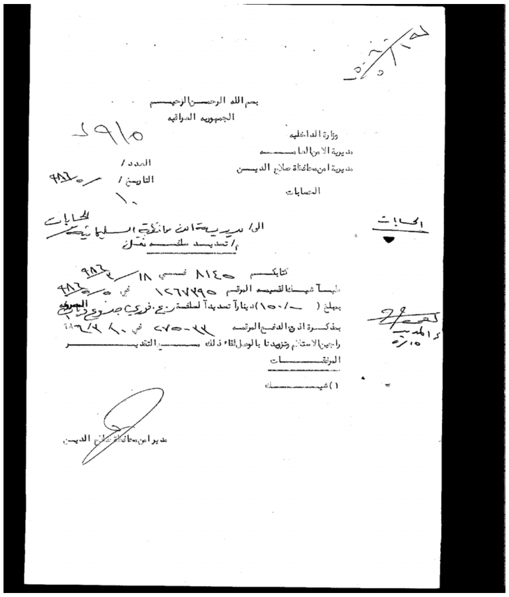
***Abstract*—We present a system for identification and recognition of handwritten and typewritten text from document images using hidden Markov models (HMMs) in this paper. Our text type identification uses OCR decoding to generate word boundaries followed by word-level handwritten/typewritten identification using HMMs. We show that the contextual constraints from the HMM significantly improves the identification performance over the conventional Gaussian mixture model (GMM)-based method. Type identification is then used to estimate the frame sample rates and frame width of feature sequences for HMM OCR system for each type independently. This type-dependent approach to computing the frame sample rate and frame width shows significant improvement in OCR accuracy over type-independent approaches.**

Keywords- optical character recognition, hidden Markov model, Gaussian mixture model

I. Introduction

In this paper, we describe an algorithm to identify handwritten and typewritten text from document images using hidden Markov models (HMMs). The existence of multiple types of text creates a big challenge in HMM-based OCR systems. As one can see from our data set of Arabic document images (Fig. 1), the document consists of both handwritten and typewritten text. The handwritten text exhibits diversity in sizes and poor alignment of each line of text, whereas the typewritten text, which is a special case of machine-printed text, is consistent in the character shapes and the alignment of text in each line. In order to achieve high-performance automatic transcription of mixed-type documents, much effort has to be made in modeling all possible variation in two distinct types of text. In an HMM-based OCR system, the challenge mostly presents in choosing proper sample rates and scales associated with the feature extraction algorithm. Practically, a single set of HMM parameters trained on both types of text included in the training set cannot handle text of mixed types well. Thus, identifying the type of text becomes an important problem for us to build specialized HMM for each type of text.

Identification of handwritten and machine-printed text has received significant attention [1][2][3][4]. In these related papers, features widely used for word spotting and OCR (projection profiles [1], Gabor features [2], moment features [3], and directional element features [4]) have been shown to be effective for text type identification. Type specific features such as the run-length histogram and texture features are also investigated in [2]. Most of the above approaches utilize classifiers such as *k*-NN and SVM. Contextual information with generative modeling methods such as the hidden Markov model [1] and the Markov random field [2] is also proved to be effective. The Gaussian Mixture Models (GMM)-based algorithm [5] has broad applications in speaker verification [5][6], writer identification [7] and word spotting [8]. We find that it is also suitable for text type identification since the definition



بسم الله الرحمن الرحيم
الجمهوريه العراقيه
وزارة الداخليه
مديرية الامن العامــــه
مديرية امن محافظة صلاح الديـــن
الحسابات

العدد /
التاريخ / ٥/ ٩٨٦

الى/ مديرية امن محافظة السليمانية
م/ تسديد سلفـــة نقل

كتابكــم ٨١٤٥ فـــي ١٨/ ٢/ ٩٨٦
طيــآ شيــكا القيمه المرقــم ١٢٦٧٧٩٥ في ٥/ ٥/ ٩٨٦
بمبلغ (١٥٠/-) دينارآ تسديدآ لسلفــة
بمذكــرة اذن الدفــع المرقمــه ٢٧٥-٧٧ في ٢٠/ ٢/ ٩٨٦
راجين الاستلام وتزويدنا بالوصل لقاء ذلك مع التقديـــر
المرفقــــات
١) شيــــك

مدير امن محافظة صلاح الديـن

Figure 1. Sample page of mixed Arabic handwritten and typewritten text.

[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.

[3] Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.155

of the problem is very similar to writer identification.

In this paper, we apply GMMs [5] to text type identification. A careful survey of the literature shows that this is the first application of GMMs for text type classification. The type identification is performed at the word-level with word boundaries generated by the HMM OCR system using a small word bi-gram language model.We further generalize the GMM-based text type classifier to an HMM classifier for better modeling of contextual information and show significant improvement in classification accuracy. The recognized type information is applied to adjust the sample rate and frame width adaptively in feature extraction resulting in significant improvement in the performance of OCR on mixed-type. In our system, image feature computation and transformation for type identification can be performed using the same steps we use when we extract features for OCR. The diagram of our text type identification and mixed-type document recognition system is shown in Fig. 2.

II. Text Type Classification using Hidden markov Models

A. *Text Type Identification Algorithm*

In our system, we use the same set of features for both text type identification and OCR. The document is separated into line images. Each line image is further divided into sliding windows in the right-to-left order, *i.e.*, the reading order of Arabic. 20 image intensity percentile features, 12 angles and correlation features representing the orientation of the stroke, the frame energy (the number of black pixels in the window), 48 gradient features, 48 concavity features, and 48 Gabor filter features are computed for each sliding window of the image [12]. The features of every 3 adjacent frames are concatenated to create a large feature vector of 531 dimensions and projected into a vector space of 17 dimensions using the Linear Discriminant Analysis (LDA). The LDA transform is estimated using the HMM-based OCR training algorithm.

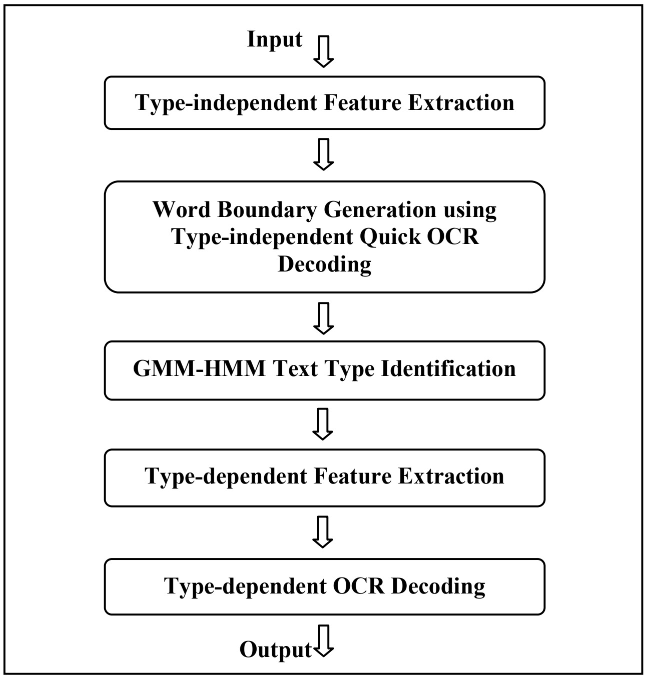

Input
Type-independent Feature Extraction
Word Boundary Generation using Type-independent Quick OCR Decoding
GMM-HMM Text Type Identification
Type-dependent Feature Extraction
Type-dependent OCR Decoding
Output

Figure 2. Text type classification and type-dependent OCR

We find that the Gaussian Mixture Models (GMM)-based speaker verification algorithm [5][6] is suitable for identification of handwritten and typewritten text due to its similarity to writer identification. In the GMM-based algorithm, features are represented in sequences of feature vectors. They are represented in the same way as features of an HMM-based OCR system. The feature vector of each type of text is represented by a GMM. The log-likelihood of a GMM with parameters λ for a sequence of feature vectors $X = \{x_1, x_2, \dots, x_N\}$ is computed as

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

where $p(x_t|\lambda)$ is the likelihood of model λ for feature vector x_t. The model giving the highest log-likelihood indicates the type of the text.

The Gaussian mixture models for handwritten and machine-printed texts are trained as follows. First we train a GMM of 2048 Gaussians as the Universal Background Model (UBM) using the 17-dimensional LDA features of word images of both types from our training data. We exclude the features of between-word gaps since it does not have any useful information. The initial parameters of the UBM are obtained using the k-means clustering algorithm. We update initial parameters using 3 iterations of the Expectation-Maximization (EM) algorithm to get the final parameters of the UBM. With the parameters of the UBM as initial parameters, we train a GMM for each type of text using 2 iterations of EM.

In our application, we do not assume that a line or even a document consists of only one type of text. Thus, we need to detect changes of types within a line. First, we use our HMM OCR system to decode the test image and get hypothesized word boundaries. A small language model (usually a word 2-gram) is used to speed up this step. Practically, the word boundaries are very accurate and more reliable than the OCR results. From each word image, we can compute the log-likelihood associated with each type using Eq. (1).

Owing to the fact that features from a word image are not always sufficient for us to make a decision, we generalize the GMM to an HMM so we can use information from more contextual information. In the first-order HMM of Fig. 3, the type of each word image is represented by a hidden variable. Each pair of adjacent word images are connected with an edge showing their conditional dependency. The observation of a hidden variable is evaluated by the log-likelihood with each type using Eq. (1). The transition probabilities between states (handwritten and typewritten) are estimated using the

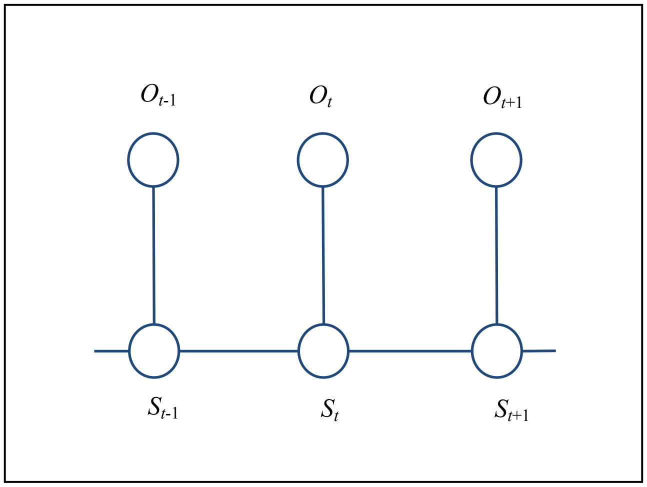


Figure 3. First-order HMM for text type identification with each hidden variable represent the state of a word image.

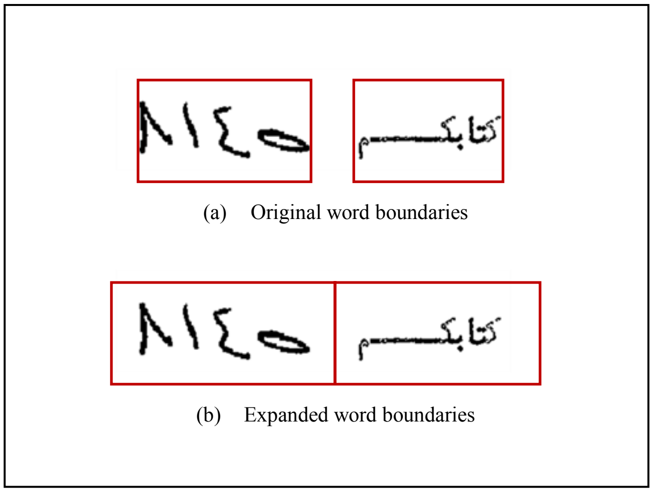


Figure 4. Word boundary expansion for frame-level evaluation.

counts of transitions in our development data set. For each hypothesized word sequence from our test set, the optimal label sequence is obtained using the Viterbi algorithm [11].

B. Evaluation Metric

In our system, word boundaries are located automatically by OCR decoding. It is not reasonable to evaluate the word-level identification error rate since errors may occur during word segmentation. Thus, we evaluate the frame-level type identification performance. First, we create the ground-truth label for each sliding window (frame) of line images using our word-level reference. A frame is labeled as handwritten, typewritten or gap. Next, we expand the boundaries of two adjacent hypothesized word images towards each other for the same length so that they touch each other (Fig. 4) and assign the label of each word to all frames in the expanded boundary of the word. Finally, we compute the identification error rate using

$$Err = \frac{\#\{f|T(f) \neq G(f), G(f) \neq \text{gap}\}}{\#\{f|G(f) \neq \text{gap}\}} \qquad (2)$$

where f is a frame, $T(f)$is the identified type of f and $G(f)$ is the ground-truth reference of f. The exclusion of gaps from the ground truth and expansion of word boundaries are introduced to avoid unnecessary penalty on small amount of white space included in word images.

We also measure the identification error rate of imbalanced decisions. We introduce an offset Δ to the log-likelihood of handwritten:

$$\log p(X|\lambda_{\text{HW}}) = N\Delta + \sum_{t=1}^{N} \log p(x_t|\lambda_{\text{HW}}) \qquad (3)$$

The error rate of each type is evaluated from the decisions made using multiple values of Δ. This is equivalent to plotting the Detection Error Trade-off (DET) curve of either type. In our OCR system described in the next section, we only use the balanced decisions (Δ=0) made by the HMM type classifier.

III. Mixed-type Document Recognition

A. Sample Rates and Widths of Sliding Window Frame in Feature Extraction of Mixed-type Documents

Identification of handwritten/machine-printed text can potentially be applied to improving OCR performance on documents of mixed types. In our data set, handwritten text and typewritten text are very different in sizes and alignment. Therefore, to perform OCR on these documents, the biggest challenge is how to select proper sample rate and width of each frame. If these problems can be solved, the HMM trained on text of mixed types is able to handle both types of text well. In our OCR system, in principle, we ignore the variation of character durations due to font or writer difference, and assume the size of text is proportional to the estimated average height of text *h*. Thus, in our HMM OCR system, the interval between two adjacent frames and the width of frames are proportional to *h*. For example, a frame is defined every *h*/60 pixels in our system. All features from the same frame are concentric and differ in widths of sliding window frames. The window width for percentile and energy features is *h*/20, the window width for angle and correlation features is *h*/6, and the window width for gradient, concavity, and Gabor features is *h*/12. Thus, the estimation of *h* is crucial for feature extraction. We describe three ways to estimate *h* as follows.

B. Locally Adaptive Height h_L

To estimate the locally adaptive height h_L, we divide each line image horizontally into 4 pieces of equal length and compute the vertical difference between the highest and lowest black pixels of piece. h_L is defined as the maximum of the 4 differences. We apply the same procedure to lines in both training and testing data.

C. Globally Adaptive Height h_G

The locally adaptive height increases the variation of training data unnecessarily. The globally adaptive height h_G we present here has better performance in handwriting recognition where the size of text changes all the time but the dynamic range is not very wide. Here, we assume the text in each page has the same size. We estimate the locally adaptive height for each line of text in the page and compute the median of all locally adaptive heights as h_G.

D. Clusttered Heights h_C

With the type identification algorithm described in Section II, we can cluster line images of two types in a page and estimate the clustered height h_C for each type separately. We process our test data as follows: a line image is split into multiple continuous pieces of smaller line images, each having a unique type of text; for typewritten text, we use the average height of the text in our data set as the estimated height h_C; for handwritten text, h_C is estimated as the median of locally adaptive heights using lines of the page that are classified as handwritten.

We need to train two sets of HMM parameters: one set for handwritten text and the other set for typewritten text. We use all training data available to train two sets of parameters and the only difference is the way we estimate text sizes for feature extraction. When we train the handwritten HMM, we extract features using the page-wise globally adaptive heights rather than clustered heights. In our training data, the amount of handwritten text is about 10 times the amount of typewritten text. Thus, the estimation of text size is dominated by handwritten text. For a training set in which handwriting does not dominate, one can re-estimate HMM parameters on a subset with adjusted ratio between two types using adaptation techniques [9][10]. The reason why we extract features for training using globally adaptive heights instead of clustered heights is to make sure the trained HMM also works well when typewritten is incorrectly classified as handwritten. Similarly, when we train the HMM for typewritten text, we extract features using the average height of typewritten lines to improve the performance of the HMM when handwritten is incorrectly classified as typewritten.

IV. Experimental Results

A. Data Collection

Our data set has over 8000 images of Arabic documents of mixed handwritten and typewritten text scanned over 30 years ago monochromatically with 200 dpi. These documents were selected from correspondences, memos, and cursive drafts under a triage according to the legibility the documents. The selected document images were still very hard to read given the nature of real-world documents and poor image capturing condition. We used 8211 images as the training set, 320 images as the development set, and 313 images as the test set. The locations and transcriptions of words and lines in these images were annotated manually. The type identification and OCR experiments discussed in this section were performed with annotated line boundaries since several automatic line-finding algorithms are available and they are not the main interest of this paper.

TABLE I. Text Type Identification Performance of the GMM and HMM Methods

Method	Frame Error Rate %		
	Handwritten	Machine-printed	All
GMM	7.62	4.95	6.33
HMM	5.34	4.21	4.75

TABLE II. Improvements in OCR Performance from Clustered Text Heights

Text Height	WER %		
	S_{MP}	S_{HW}	S_{MX}
h_L	17.3	49.9	44.7
h_G	16.5	46.8	44.7
h_C	16.6	46.6	40.5

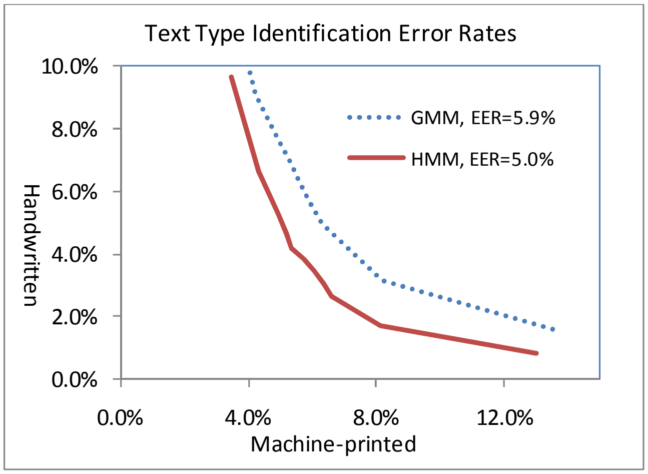


Figure 5. Type-specific identification error rates.

B. Text Type Identification Results

We trained Gaussian mixture models for handwritten and typewritten text, respectively using 2400 handwritten word images and 2400 typewritten word images from our training set.

We created word boundaries of the development and test sets using an HMM OCR system trained on features with globally adaptive text heights h_G and a word 2-gram trained on all training data. Our HMM OCR system consisted of 14-state character HMM with the right-to-left state transition configuration. Each state had a state-tied mixture model of 256 Gaussians. GMM means and covariances were shared for each state of tri-phones of the same center-phoneme. GMM weights were not shared.

For the HMM text type classifier, the transition probabilities were estimated on the development set and applied to the test set. To avoid under-estimation of interstate transition probabilities, we excluded lines that have a unique type of text in this step.

We evaluated the error rates of the GMM and HMM text type classifiers using our test set. There were totally 2,845,383 non-gap frames in the test set including 1,367,754 typewritten frames and 1,477,629 handwritten frames. The error rates are shown in Table I. Remember that the GMM and HMM methods used the same GMM parameters. The only difference between them were the use of transition probabilities in the HMM. Table I shows that the GMM-based method was improved significantly from the contextual information of the HMM. The consistent improvement from the HMM-based method can be justified by DET curves (Fig. 5) of both GMM and HMM computed using Eq. (3). The equal error rates (EER) of GMM and HMM are 5.9% and 5.0%, respectively.

C. OCR Results using Clustered Text Heights

With text types classified, we redid feature extraction on the test set using clustered text heights, and decoded the test set using two sets of type-dependent HMM parameters trained by following those steps described in Section III (D). The language model we used were word 3-gram trained on transcription of all training data. For comparison, we also ran two other HMM OCR systems, one with locally adaptive text heights and the other with globally adaptive text heights, on the test set. To better present our results, we analyzed the composition of each page and divided the test into three subsets.

- S_{MP} consisted of 115 pages. Each of them has 90% or more machine-printed words,
- S_{HW} consisted of 151 pages. Each of them has 90% or more handwritten words, and
- S_{MX} consisted of 47 pages. Each of them has 10-90% machine-printed words.

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 performance of three OCR systems is shown in Table II. First of all, the first two rows of results show big improvements from using globally adaptive heights (h_G) as opposed to locally adaptive heights (h_L) on both subsets with one predominant type (S_{MP} and S_{HW}). However, h_G did not make any improvement on the subset of mixed types (S_{MX}). The last two rows of results show that clustered text heights (h_C) improved the OCR performance on S_{MX} significantly by9.4% relative (from 44.7% to 40.5%). On two other subsets, h_C did not bring much degradation in OCR performance (from 16.5% to 16.6% on S_{MP} and from 46.8% to 46.6% on S_{HW}, nearly negligible).

V. Conclusion

We described a GMM-HMM based approach to identifying machine-printed and handwritten text in this paper. The use of contextual constraints in the HHM was proved to be highly effective for reducing type identification errors. We also showed that the word error rate on document images of mixed types of text can be reduced significantly using our type identification method.

References

[1] J. K. Guo, M. Y. Ma, “Separating handwritten material from machine printed text using hidden Markov models,” Proc. Sixth International Conference on Document Analysis and Recognition (ICDAR '01), Washington, DC, USA, 2001, pp. 439-443.

[2] Y. Zheng, H. Li, D. Doermann, “Machine printed text and handwriting identification in noisy document Images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26(3), 2004, pp. 337-353.

[3] R. Kandan, Nirup Kumar Reddy, K. R. Arvind, and A. G. Ramakrishnan, “A robust two level classification algorithm for text localization in documents,” Proc. 3rd International Conference on Advances in Visual Computing (ISVC'07), Vol. 2. Springer-Verlag, Berlin, Heidelberg, 2007, pp. 96-105.

[4] S. Chanda, K.Franke, and U. Pal, “Structural handwritten and machine print classification for sparse content and arbitrary oriented document fragments,” Proc. ACM Symposium on Applied Computing (SAC '10), 2010, pp. 18-22.

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

[6] 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,.

[7] A. Schlapbach, H.Bunke, “Off-line writer identification using Gaussian mixture models,” Proc. 18th International Conference on Pattern Recognition, 2006, vol. 3, pp. 992-995.

[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] 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.

[10] 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.

[11] G. A. Fink, “Markov models for pattern recognition: from theory to applications,” Springer Press, 2007, pp. 75-76.

[12] S. Saleem, H. Cao, K. Subramanian, M. Kamali, R. Prasad, P. Natarajan, “Improvements in BBN's HMM-based offline Arabic handwriting recognition system,” Proc. International Conference on Document Analysis and Recognition, 2009, pp. 773-777.