Date Field Extraction from Handwritten Documents Using HMMs

Ranju Mandal*, Partha Pratim Roy†, Umapada Pal‡ and Michael Blumenstein*
*School of Information and Communication Technology, Griffith University, Queensland, Australia
ranju.mandal@griffithuni.edu.au, m.blumenstein@griffith.edu.au
†Dept. of Computer Science & Engineering, Indian Institute of Technology, Roorkee, India
proy.fcs@iitr.ac.in
‡Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
umapada@isical.ac.in

***Abstract*—Automatic document interpretation and retrieval is an important task to access handwritten digitized document repositories. In documents, the date is an important field and it has various applications such as date-wise document indexing/retrieval. In this paper, a framework has been proposed for automatic date field extraction from handwritten documents. In order to design the system, sliding window-wise Local Gradient Histogram (LGH)-based features and a character-level Hidden Markov Model (HMM)-based approach have been applied for segmentation and recognition. Individual date components such as month-word (month written in word form i.e. January, Jan, etc.), numeral, punctuation or contraction categories are segmented and labelled from a text line. Next, a Histogram of Gradient (HoG)-based features and a Support Vector Machine (SVM)-based classifier have been used to refine the results obtained from the HMM-based recognition system. Subsequently, both numeric and semi-numeric regular expressions of date patterns have been considered for undertaking date pattern extraction in labelled components. The experiments are performed on an English document dataset and the encouraging results obtained from the approach indicate the effectiveness of the proposed system.**

I. Introduction

Automatic handwritten document indexing has been an active research area in recent years. Date is a useful piece of information and it could be used as a key in various applications such as date-based document searching and indexing of document repositories such as administrative documents, historical archives, postal mails, etc. Automatic spotting/extraction of date information involves challenges due to different date patterns (Numeric and Semi-numeric dates consisting of different lengths), writing styles of different individuals, touching characters and classifier confusion between numerals, punctuation and text. Fig. 1 shows four handwritten lines containing date patterns. Significant work has been undertaken [1], [2] in the area of word spotting to make handwritten text available for searching and browsing. Hidden Markov Model (HMM)-based methods [1] are extensively used for modeling handwritten text, word spotting, etc. A Recurrent Neural Network-based approach has been proposed in [2] for word-based searching and indexing. Field-based information retrieval has also become more popular because of the poor performance and computationally expensive OCR engines[1] available in transcribing handwritten documents. Few research publications have also been available for automatic form field extraction from handwritten English documents [3], [4], [5]. Thomas et al. [3] proposed an HMM-based classification model for alpha-numerical sequence recognition. Koch et al. [4] also proposed a method for numerical field extraction based on HMMs. To localize the desired numerical fields, a syntactic analyser was applied over the handwritten text lines. Segmentation-driven recognition has been proposed by Chatelain et al. [5] to find numerical sequences. Most of the papers mentioned above deal with alpha-numeric string extraction. Handwritten date information processing from scanned documents still remains a very challenging task. Date pattern detection and interpretation in handwritten documents is challenging due to the unconstrained nature of the handwriting of different individuals, touching numerals, different patterns of a single date etc. It can be noted that, the date patterns appear in different format in documents. Some of these formats of English dates are used in 12/03/2012; 12th March, 2012; March 12, 2012 etc. Various formats of such date patterns make automatic searching more challenging. There is very limited published research available on automatic segmentation and recognition of dates from bank cheques. The existing methods work for specific formats of date and also work on specific documents such as bank cheques. Suen et al. [6] proposed an approach for automatic cheque processing i.e. the segmentation and recognition of dates written on bank cheques. First, the method segments a date image by using the separator information. Two separators are then used to detect the Year, Day and Month zones based on shape and spatial features. Next, numeric and non-numeric month fields are recognised by a connected digit recogniser and a cursive word recognizer. The recognition results are finally sent to a parser, which is used to interpret acceptable

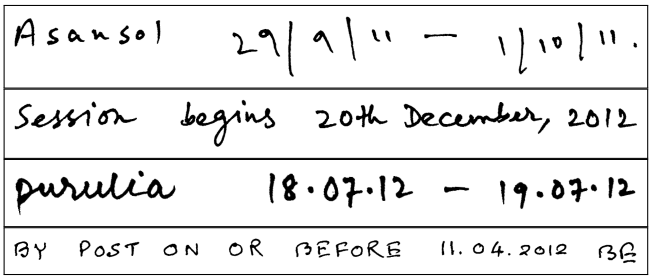


Fig. 1. Four samples of handwritten lines containing date information of different patterns.

[1] http://code.google.com/p/ocropus/

results and to reject invalid ones. Xu et al. [7] described a knowledge-based segmentation system for handwritten dates on bank cheques. The knowledge derived from writing style information are utilised, as well as syntactic and semantic constraints and different knowledge sources are adopted at different stages. Roy et al. [8] recently proposed a Deep belief Networks (DBNs)-based word hypotheses rescoring scheme for handwritten word recognition. In an early work by the present authors, a two-staged classification-based approach for date field extraction from handwritten documents [9] was proposed. A word-level and component-level classification was done to locate the date components. In an another work [10], date field extraction from multi-lingual (i.e. English (Roman), Devnagari and Bangla) handwritten documents was proposed.

In this paper, we present a two-step approach for date field extraction from handwritten documents using two classifiers namely HMMs and SVMs. This approach performs better than the previously proposed approach [9]. We observed that the present system works better due to the performance of the HMM on handwritten text. Unlike the previous method automatic segmentation of words into characters avoids the problem of pre-segmentation of date-fields. The approach we have adopted has been inspired by the method proposed by Roy et al. [11] for word recognition. The HMM is used for character level segmentation and recognition of date components such as numerals, alpha-numeric characters, punctuation, etc. to locate the date field in handwritten documents. HMM-based systems perform well for the automatic segmentation of characters in handwritten words, but due to generative properties of HMM, character recognition may fail occasionally. An SVM-based discriminative classifier trained with numerals and punctuation has been applied to improve the recognition results of labelled date fields obtained from the HMM-based system. A flow diagram of date field recognition system is shown in Fig. 2.

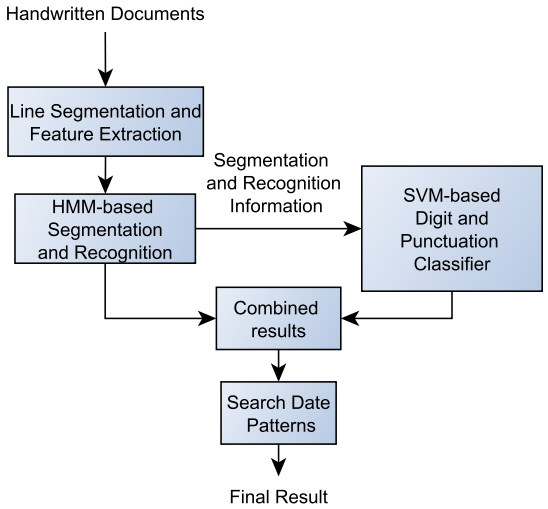


Fig. 2. Flow diagram of date field recognition system.

The organization of the rest of this paper is as follows: In Section II, the proposed methodology of date field extraction is detailed. Section III discusses the process of searching date patterns. The experimental results have been described and analysed in Section IV. Finally, conclusions are drawn in Section V.

II. Proposed Approach

As an initial step, Otsu's binarization method has been applied to convert the grey-scale images into binary images. Our date retrieval approach searches the date patterns in text line images. Hence, the binary document is segmented into individual text lines using a line segmentation algorithm [12] and the segmented lines are used for experimentation. Character-based HMMs [13] have been successfully used for recognition of arbitrary sets of words in English (Roman)/Latin scripts. An advantage of these systems is that they allow the recognition of unknown characters from the training data once the character models are trained. HMMs avoid the problem of pre-segmentation of words into characters so the errors of pre-segmentation can be eliminated. In our problem, as we consider all possible date patterns, a date information can consist of numerals, punctuation and month-words (month written in word form i.e. textual month). To handle all situations, our HMM-based recogniser has been trained with numerals strings, alpha-numeric strings and punctuations. Next, an SVM-based recognizer, especially trained with numerals and punctuation has been applied to refine the results obtained from the HMM-based recognition system. The alpha-numeric strings are not recognised by the SVM-based system and the results obtained from the HMM-based system are used in the final combination stage. The SVM-based numeral classifier initially takes the character alignment information produced by the HMM-based recogniser. The recognition errors generate by the HMM-based system due to segmentation problem have been handled by SVM-based isolated numeral and punctuation classifier. The detailed descriptions of the feature extraction and classification techniques using HMM and SVM-based approaches have been presented below in Section II-A and Section II-B respectively.

A. Hidden Markov Model-based recognition system

Feature Extraction: Sliding window-based Local gradient histogram (LGH) [14] features have been computed in our method. Here, a sequence of overlapping sub-images have been produced by a fixed width sliding window which traverses the handwritten text image from left to right in order. The sliding window images are subdivided into a 4×4 (4 rows and 4 columns) regular grid and from each cell a histogram of gradient orientations is calculated by using the pixels available in it. The direction of the gradient is quantized into L directions and the gradient strengths are accumulated with each of the quantized directions. Each bin specifies a particular octant in the angular radian space. Here we consider 8 bins $360°/45°$ of angular information. The histogram is formed by adding up $m(x, y)$ to the bin indicated by quantized $\Omega(x, y)$. A 128-dimensional feature vector has been computed from each sliding window position by concatenation of the 16 histograms of 8 bins.

Hidden Markov Model (HMM): The feature vector sequence is processed using left-to-right continuous density HMMs [15]. One of the important features of HMM is the capability to model sequential dependencies. An HMM can be defined by initial state probabilities π, state transition matrix $A = [a_{ij}], i, j = 1, 2, \cdots, N$,where a_{ij} denotes the transition

probability from state i to state j and output probability $b_j(O_k)$ modelled with a continuous output probability density function. The density function is written as $b_j(x)$, where x represents a k-dimensional feature vector. A separate Gaussian Mixture Model (GMM) is defined for each state of the model. Formally, the output probability density of state j is defined as

$$b_j(x) = \sum_{k=1}^{M_j} c_{jk} \mathcal{N}(x, \mu_{jk}, \sum jk) \quad (1)$$

where, M_j is the number of Gaussians assigned to j, and $\mathcal{N}(x, \mu_{jk}, \sum)$ denotes a Gaussian with mean μ and covariance matrix $\sum$ where c_{jk} is the weight coefficient of the Gaussian component k of state j. For a model λ, if $\mathcal{O}$ is an observation sequence $\mathcal{O}$ = ($\mathcal{O}_1$, $\mathcal{O}_2$, $\cdots$, $\mathcal{O}_T$) which is assumed to have been generated by a state sequence $\mathcal{Q}$ = ($\mathcal{Q}_1$, $\mathcal{Q}_2$,...,$\mathcal{Q}_T$), of length T, we calculate the observations probability or likelihood as follows:

$$P(\mathcal{O}, \mathcal{Q}|\lambda) = \sum_{Q} \pi_{q1} b_{q1}(\mathcal{O}_1) \prod_{T} a_{qT-1qr} b_{qT}(\mathcal{O}_T) \quad (2)$$

where μ_{q1} is the initial probability of state 1. The sliding window-based feature vectors along with the line-wise transcriptions of the handwritten line images have been used in order to train the character-level HMMs. The HTK toolkit [16] implementation model of the HMM developed for speech signal modelling has been used in our experiments where recognition is based on the Viterbi algorithm.

Computation of character boundaries: A Viterbi forced alignment (FA) algorithm has been applied to calculate the character boundaries in handwritten lines. The algorithm finds the optimal alignment of a set of Hidden Markov Models. An iterative alignment and retraining process called embedded training has been used to refine character segmentation boundaries. The character segments ($S_1, S_2, \cdots, S_n$) of a given hypothesis have been obtained using the alignment algorithm. An N-best Viterbi list composed of N hypotheses are generated. Among all, the best hypothesis has been chosen based on the addition of a maximum likelihood (ML) segmentation at the character level. Character level segmentation performance of this algorithm on sample handwritten lines has been shown in Fig. 3.

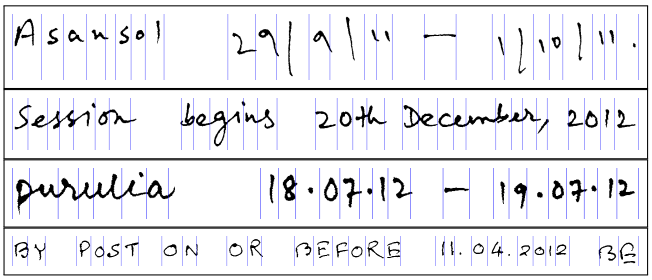


Fig. 3. Samples of handwritten lines following character level segmentation.

B. SVM-based classifier

SVM-based classification system has been used here as an isolated character classifier. The alignment information produced by the Viterbi-forced alignment of characters has been used for this system also. The feature extraction technique and classifier details are described below.

Histogram of Oriented Gradients (HoG): HoG [17] is a robust feature descriptor which is very popular in computer vision and image processing for object detection. Dalal and Triggs [17] first described the HoG descriptors and primarily focused on pedestrian detection in static images. The basic idea behind the HoG descriptor is that the shape and appearance of the object within an image can be described by the intensity gradient distribution or the edge directions. The HoG descriptors are typically computed by dividing an image into small spatial regions called 'cells'. A histogram of the gradient direction of the pixels within the cells is computed. The histogram bins/channels are evenly spaced over 0° to 180° or 0° to 360° based on the usage of signed or unsigned gradient values. Combining the histogram of all the cells produces the features. HoG features suit the problem well because it operates on the localized cells and it is capable of describing the shape and appearance of the handwritten numerals in the present context. Here, 8 bin/orientations have been considered over 7×7 blocks for feature extraction, which resulted in a 392-dimensional feature vector.

SVM: SVM is a popular classification technique which can successfully be applied to a wide range of applications [18]. So, in our experiments, we have used an SVM-based numeral and punctuation classifier. In our experiments, the Gaussian kernel SVM outperformed other non-linear SVM kernels, hence we are reporting our recognition results based on the Gaussian kernel only. The hyper parameters of the SVM were set using a validation process as follows; kernel type = RBF, $\gamma = 0.04$ and $C = 3$. The best results have been achieved by setting the above values for these parameters.

C. Combination of recognition scores

In this procedure, the inputs are the N-best list score from the individual classifiers and the score of both classifiers are usually normalized before fusion. The resultant score has been calculated by combination of scores obtained from the classifiers as follows. Let $D_1, D_2, \cdots, D_L$ be the set of L classifiers [19]. The output of the i^{th} classifier is denoted as $D_i(x) = [d_{i,1}(x), \cdots, d_{i,c}(x)]^T$, where $d_{i,j}(x)$ is the degree of support given by classifier D_i to the hypothesis that test where class x comes from and c refers to the crisp class label. We construct $\widehat{D}$, the fused output of the L level classifier as

$$\widehat{D} = F(D_1(x), \cdots, D_L(x))$$

where F is the aggregation rule of the maximum average and product operator. Top three likelihoods along with the confidence scores have been estimated for particular segmented components from the HMM-based system. The SVM-based system estimates the confidence measure of the top three labels obtained from HMM's output. Finally, the label of a segmented part which receives a combined maximum score has been considered as the final class label. Fig. 4 shows an example of how the combination has been done using top three scores obtained from the classifiers.

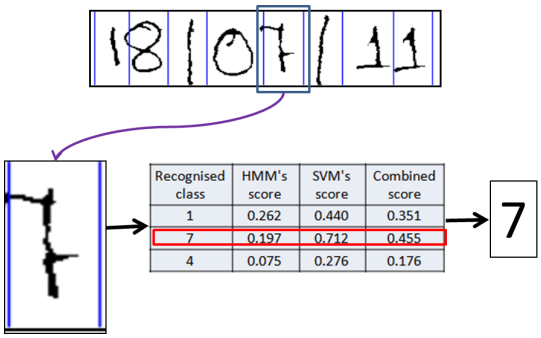


Fig. 4. An example of scores combination obtained from the HMM and the SVM classifiers. An average value has been considered as combined score.

III. Searching of Date Patterns

Text lines with their four different types of recognized components (month-word, numeral, punctuation and text) are considered here for date pattern detection.

A. Date pattern matching

The labelled components in each candidate text lines such as punctuation, numeral and months-word (ex. Jan, Feb) are noted. Next, the date regular expressions are searched using the sequence of labelled components. In our approach, we consider two different date patterns for searching, namely: numeric and semi-numeric patterns.

Numeric and semi-numeric date matching: A date field consisting of only numerals and punctuations is considered as a numeric date field in our approach, e.g. (15/08/2012, 12-01-12 etc.). Other date fields that consist of month-word, numerals and contraction (st, nd, rd, th) are considered as semi-numeric dates e.g. 31st March, 2011. For numeric date extraction we search a sub-sequence of components with the following date regular expression:

(d—dd)(/.-)(d—dd)(/.-)(dd—dddd)

A complete numeric date field consists of at least one or at most two numerals for day information, at most two numerals of month information and a maximum of four numerals for year information. The following date regular expressions have been used to search for semi-numeric dates.

(md—mdd)(.,-)(dd—dddd) and
(d—dd)(contraction)(m)(.,-)(dd—dddd)

where, d represents numerals, m represents a month-word and we consider three types of punctuation in the date syntax. Two types of regular expression for semi-numeric date fields have been considered. In a semi-numeric date pattern month-word may be in the front or in the middle of the sequence. Contractions can be found before alpha-numeric month information. The 'grep' command-line utility available in Unix-like systems for pattern searching has been used to search all the above-mentioned date patterns from the transcribed lines.

IV. Results and Discussion

Two different sets of data have been used to train the HMM-based and SVM-based recognisers. The HMM-based system used handwritten text lines and handwritten numeral strings as training data. The IAM English sentence dataset [20] and 1500 samples of 6 digit Indian PIN code string were used to train the HMM-based system, whereas to train the SVM-based numerals and punctuation recogniser, English digits from the MNIST[2] dataset and 904 handwritten punctuations were used. The test dataset used for the experiments included 1240 handwritten text lines collected from English documents which are written by different individuals from various professions. The dataset contains a date sequence of different valid patterns. A 5-fold cross validation technique has been employed to compute the recognition accuracy. A validation dataset was used to vary the HMM parameters such as the number of states, the number of Gaussian distributions and width of the sliding window. 6 states for each character model and 8 GMMs for each state were selected on the basis analysing the performance. In LGH feature extraction, a sliding window width of 10 pixels with a 50% overlapping ratio provides the best result in our experiments. Table I shows the performance of month-word recognition accuracy by HMMs only. Table II shows the results produced in these experiments solely by the HMM-based system. Table III shows the improved results obtained by the HMM-SVM hybrid rescoring system. A few sample images in Table IV show the qualitative results obtained from the experiments.

TABLE I. Performance of HMM-based month-word field recognition

Total samples	Correctly recognised samples	Precision(%)	Recall(%)
184	160	100	86.95

TABLE II. HMM-based recognition accuracy of date fields

Data	Total samples	Correctly recognised samples	Accuracy(%)
Numeric Date	256	208	81.25
Semi-Numeric Date	163	129	79.14

TABLE III. Recognition accuracy based on combined results of HMM and SVM

Data	Total samples	Correctly recognised samples	HMM+SVM Accuracy(%)
Numeric Date	256	225	87.89
Semi-Numeric Date	163	134	82.20

Comparative analysis: The proposed approach performs better than our previous approach [9]. Table V and VI shows the comparison on month-word and date field recognition in precision-recall measure.

Error analysis: We have found that some errors occurred due to segmentation problems generated by the HMM on some low quality images. SVMs also failed to recognise those characters if there was improper character alignment. Table VII shows some samples of erroneous results produced in the experiments.

V. Conclusions

An automated system for date field extraction from handwritten documents has been presented here. Sliding window-

[2]http://yann.lecun.com/exdb/mnist/

TABLE IV. QUALITATIVE RESULTS FROM THE HMM-BASED AND COMBINED APPROACH ON SAMPLE IMAGES

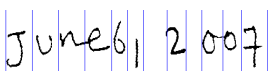
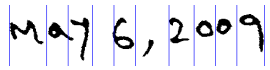
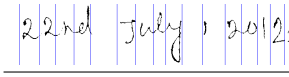
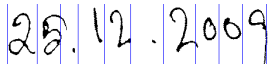
Date sample	HMM results	HMM+SVM results
	17/07/1r	17/07/12
	June6,,2 007	June6, 2007
	May 6,20o9	May 6,2009
	22nd July, a0i2	22nd July, 2012
	t5.12. 2009	25.12. 2009

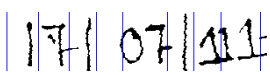
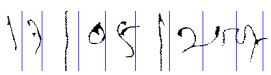
TABLE V. COMPARISON WITH THE PREVIOUS METHOD ON MONTH-WORD RECOGITION.

Method	Precision(%)	Recall(%)
Previous Approach [9]	77.41	83.72
Proposed Approach	100	86.95

TABLE VI. COMPARISON WITH THE PREVIOUS METHOD ON COMPLETE DATE FIELD RECOGITION

Method	Precision(%)	Recall(%)
Previous Approach [9]	89.08	74.87
Proposed Approach	100	85.68

TABLE VII. ERRONEOUS RESULTS

Date sample	HMM results	HMM+SVM results
	/07/111	17/4/18
	17/05/207	17/05/209

wise LGH-based features with HMMs have been used for segmentation and recognition of date components such as numerals, punctuation, month-word, etc. Next an SVM-based recognition technique with HoG-based features was applied to refine the results obtained from the HMM-based recognition approach. Finally, different date patterns were searched from the sub-sequence of labelled components. Overall, the results of the proposed approach are encouraging. However, although the proposed date field extraction method works well on English handwritten documents, in future, we would focus alternative segmentation-free approaches to enhance the results. Moreover, the system can be extended for date field extraction from multi-script documents.

REFERENCES

[1] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "HMM-based word spotting in handwritten documents using subword models," in *Proc. International Conference on Pattern Recognition*, 2010, pp. 3416–3419.

[2] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on Recurrent Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3, no. 3, pp. 211–224, 2012.

[3] S. Thomas, C. Chatelain, L. Heutte, and T. Paquet, "Alpha-numerical sequences extraction in handwritten documents," in *Proc. International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2010, pp. 232–237.

[4] G. Koch, L. Heutte, and T. Paquet, "Numerical field extraction in handwritten incoming mail documents," in *Proc. International Workshop on Pattern Recognition in Information Systems (PRIS)*, 2003, pp. 167–172.

[5] C. Chatelain, L. Heutte, and T. Paquet, "Segmentation-driven recognition applied to numerical field extraction from handwritten incoming mail documents," in *Proc. International Workshop on Document Analysis System (DAS)*, 2006, pp. 564–575.

[6] C. Y. Suen, Q. Xu, and L. Lam, "Automatic recognition of handwritten data on cheques - Fact or fiction?" *Pattern Recognition Letters*, vol. 20, pp. 1287–1295, 1999.

[7] Q. Xu, L. Lam, and C. Y. Suen, "A knowledge-based segmentation system for handwritten dates on bank cheques," in *Proc. International Conference on Document Analysis and Recognition*, 2001, pp. 384–388.

[8] P. P. Roy, Y. Chherawala, and M. Cheriet, "Deep-belief-network based rescoring for handwritten word recognition," in *Proc. International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 506–511.

[9] R. Mandal, P. P. Roy, and U. Pal, "Date field extraction in handwritten documents," in *Proc. International Conference on Pattern Recognition (ICPR)*, 2012, pp. 533–536.

[10] R. Mandal, P. P. Roy, U. Pal, and M. Blumenstein, "Multi-lingual date field extraction for automatic document retrieval by machine," *Information Sciences*, vol. 314, pp. 277–292, 2015.

[11] S. Roy, P. P. Roy, P.Shivakumara, and U. Pal, "Word recognition in natural scene and video images using Hidden Markov Model," in *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, 2013, pp. 1–4.

[12] P. P. Roy, U. Pal, and J. Lladós, "Morphology based handwritten line segmentation using foreground and background information," in *Proc. International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2008, pp. 241–246.

[13] A. B. Bernard, F. Menasri, R. H. Mohamad, C. Mokbel, C. Kermorvant, and L. Sulem, "Dynamic and contextual information in HMM modeling for handwritten word recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2066–2080, 2011.

[14] J. R. Serrano and F. Perronnin, "Handwritten word-spotting using Hidden Markov Models and universal vocabularies," in *Proc. International Conference on Pattern Recognition*, 2009, pp. 2106–2116.

[15] A. Hasnat, S. M. Habib, and M. Khan, "A high performance domain specific OCR for Bangla script," in *Novel Algorithms and Techniques In Telecommunications, Automation and Industrial Electronics,LNCS,Springer*, 2008, pp. 174–178.

[16] S. Young, *The HTK Book*. Version 3.4. Cambridge University English Department, 2006.

[17] N. Dalal and B. Triggs, "Histogram of Oriented Gradients for human detection," in *Proc. CVPR*, vol. 1, 2005, pp. 886–893.

[18] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[19] L. Kuncheva, J. C. Bezdek, and R. P. W. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern recognition*, vol. 34, no. 2, pp. 299–314, 2001.

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