21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Date Field Extraction in Handwritten Documents

Ranju Mandal
Computer Vision and Pattern
Recognition Unit, Indian Statistical
Institute, Kolkata-108, India
ranjumandal@gmail.com

Partha Pratim Roy
Laboratoire d'Informatique
Université François Rabelais
Tours, France
partha.roy@univ-tours.fr

Umapada Pal
Computer Vision and Pattern
Recognition Unit, Indian Statistical
Institute, Kolkata-108, India
umapada@isical.ac.in

Abstract

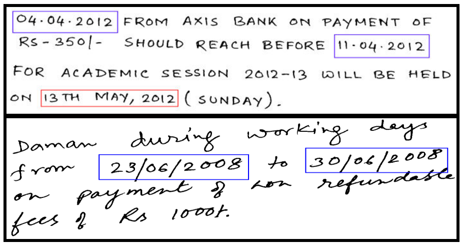
Automatic extraction of date patterns from handwritten document involves difficult challenges due to writing styles of different individuals, touching characters and confusion among identification of alphabets and digits. In this paper, we propose a framework for retrieval of date patterns from handwritten documents. The method first classifies word components of each text line into month and non-month class using word level feature. Next, non-month words are segmented into individual components and classified into one of alphabet, digit or punctuation. Using this information of word and character level components, the date patterns are searched first using voting approach and then we detect the candidate lines for numeric and semi-numeric date using regular expression. Gradient based features and Support Vector Machine (SVM) are used in our work for classification. The experiment is performed on handwritten dataset and we have obtained encouraging results from it.

I. Introduction

Date is useful and important information that could be used as key for searching and indexing of handwritten documents in administrative documents, historical archives, postal mails, etc. Some available OCR engines [1] do not work well in understanding handwritten documents. The output of such OCRs cannot be used for date extracting compilers because of poor recognition result. Hence, date extraction process from such documents will be very useful in searching and interpretation. To the best of our knowledge, there is no work that can search date pattern in printed/handwritten documents.

Date pattern detection and interpretation in handwritten documents is a challenging task due the unconstrained handwriting styles of different individuals. Alpha-numeric characters that represent date are sometimes touching and recognition confusion between numerals and alphabet makes the task more challenging. We have shown two examples of handwritten documents containing date information in Fig.1. It is to be noted that, the date patterns appear in different format in documents. Some of these formats of a single date are 12/03/2012 or 12th March, 2012 or March 12, 2012 or 12-03-2012 or 12.03.2012.or 12.03.12, etc. Automatic searching of such different date patterns from the documents is difficult.

Few research works have been published for automatic form field extraction from handwritten documents [2, 3, 4]. Recently, field based information retrieval got more popularity than recognition of full handwriting document. Koch et al. [3] proposed a method using HMM for numerical field extraction. To localize the desired numerical fields, syntactic analyzer has been applied over the handwritten text lines. Thomas et al. [2] proposed a HMM based classification model for alpha-numerical sequence recognition. Chatelain et al. [4] proposed an approach to locate numerical sequence using a segmentation-driven recognition. To extract the desired numerical sequence, a syntactical analysis has been performed on each line of text. Most of the papers mentioned before deals with alpha-numeric string extraction. This paper moves a step further in document interpretation and uses the recognition labels of alpha-numeric characters to locate the date fields in documents and this is the first work on date extraction.



04·04·2012 FROM AXIS BANK ON PAYMENT OF
RS-350/- SHOULD REACH BEFORE 11·04·2012
FOR ACADEMIC SESSION 2012-13 WILL BE HELD
ON 13TH MAY, 2012 (SUNDAY).

Daman during working days
from 23/06/2008 to 30/06/2008
on payment of non refundable
fees of Rs 1000/-.

Figure 1. Sample handwritten documents containing date fields. Numeric and semi-numeric date fields are marked with blue and red rectangle, respectively.

A block diagram of our proposed system is shown in Fig.2. A three-stage approach has been proposed here for date field extraction. In the first stage, month and non-month handwritten word blocks are separated. For this purpose, words blocks are extracted using morphological operation and the segmented word blocks are classified into month and non-month classes using word block level feature analysis. The second stage performs component analysis for each non-month handwritten word blocks. Isolated digits, punctuations and alphabets are identified using component level feature analysis. The components with low recognition confidence are analyzed further for touching segmentation [8]. We have used 400 dimensional gradient based features and Support Vector Machine (SVM) for classification in both word block and component level classification. Finally, in the third stage, numeric and semi-numeric (contains month field as text string) date patterns are searched from the sequence of the labeled components. To do so, candidate lines are selected first using voting approach and next a regular expression analysis is used to detect the date patterns.

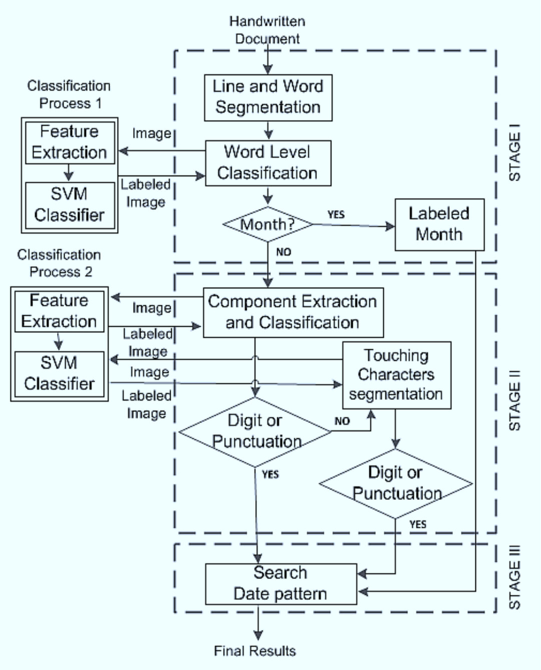


Figure 2. Block diagram of the proposed system

II. Feature Extraction and Classification

Gradient based feature of dimension 400 is used in our system for the recognition of word (month) and character (digits/punctuation/letters). The text image is normalized into 126x126 size and converted to gray-scale image by applying a set of mean-filtering. Next the resultant gray image is segmented into 9X9 blocks. Roberts filter is applied next to obtain gradient image. The direction of gradient is quantized into 16 directions and the gradient strengths are accumulated in each quantized direction. Histograms of 16 quantized directions are computed in each of 9x9 blocks. For better result, 9x9 blocks are down sampled into 5x5 by a Gaussian filter instead of directly dividing the image into 5x5. Thus, we get 5x5x16 = 400 dimensional feature. We feed this feature into a SVM classifier (Gaussian kernel with Radial Basis Function) for classification of text images. Details of SVM can be found in [7].

III. Date Field Extraction Approach

The document image is converted into binary image using global histogram-based Otsu binarization method. 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 [4].

These text lines are segmented into different component levels (words, digits, punctuations and letters) and the components are later used for date searching. Different steps used for this purpose are explained as follows.

A. Classification of Month and Non-month words

Horizontal Run Length Smoothing Algorithm (RLSA) [5] is applied on each text line to get individual words as a component. A connected component labeling is applied to find the bounding box of the word patches in the line. Next, using the patch mask, the original word is considered from the binary image and a classification (described in Section II) is performed for two-class problem: month and non-month word blocks identification. To train the classifier, a data set is used with different styles of month format that appear in date pattern. For example, “January”, “JANUARY”, “JAN”, “Jan”, etc. In Fig.3, the word blocks of text lines are classified as month and non-month blocks.

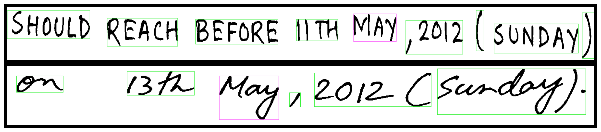

SHOULD REACH BEFORE 11TH MAY, 2012 (SUNDAY)
on 13th May, 2012 (Sunday).

Figure 3. Text lines showing detection of month blocks. Here, month, non-month blocks are marked with pink, green, respectively. (For better visibility see the PDF version)

B. Character Component classification

The words which are classified as non-month in earlier stage are considered here. A connected component analysis is employed to segment the non-month words into different components and component-wise classification is done to extract the character/digit components from these non-month words. For this purpose, connected components are fed to component level classification stage. These components are mainly classified into “punctuation”, “digit” or “alphabet” level (See Fig.4). There are some components which might be touching and these components could not be classified properly in this stage. Hence, the components with high recognition confidence are accepted and directly considered for date pattern matching. The rest of the components with low confidence are selected for touching component segmentation analysis. The confidence threshold is selected as 0.4 according to the experimental results. If some isolated characters are not identified properly, the touching character segmentation step (explained in next section) will identify it.

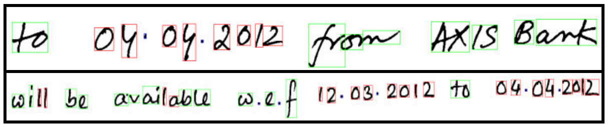

to 04·04·2012 from AXIS Bank
will be available w.e.f 12·03·2012 to 04·04·2012

Figure 4. Component classification results from handwritten lines into digit, punctuation and text. These are marked by red, blue and green, respectively.

C. Touching Character Segmentation

There may exist touching digits/characters in a component. Components with low confidence score at earlier stage of recognition are considered as touching and chosen for segmentation. Here, we use a dynamic programming based touching character segmentation scheme [8]. First, we find the cavity regions formed between touching characters. The cavity regions are obtained using Water Reservoir concept [5]. We use Top-Bottom reservoir analysis to find the cavity regions in a touching component. A set of candidate segmentation points is obtained from these regions using cavity region analysis. Next, the touching component is segmented into these candidate points to find different sub-images. Using dynamic programming, the recognition confidence of sub-images is analyzed and optimum segmentation path is found. Finally, based on the segmentation lines, the touching component is segmented. This approach segments touching digits in most of the cases. In Fig.5, the circle shows segmentation result of two touching digits 2 and 0.

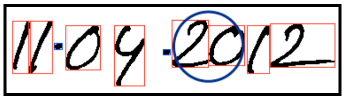


Figure 5. A segmentation result of touching digits.

D. Searching of Date Pattern

The text lines with its four different recognized components (month, digit, punctuation, text) are considered here for date pattern detection. This approach is divided into 2 parts: candidate line selection and pattern matching.

Candidate Line selection: The text lines that contain labelled months, digits and punctuation are selected here. For this purpose, we compute the total number of digits, punctuation marks, month string of a text line. Depending on the value of the individual counters we decide to search the date patterns in that text line in the next stage.

Pattern matching: The components in each candidate text lines are sorted in left to right direction using the CG (Centre of Gravity) and the positions of punctuation, digit and month text are noted. The date patterns are searched next using the sequence of labelled components. In our approach, we consider two different date pattern for searching, namely: numeric and semi-numeric patterns.

Numeric date matching: A date field consisting of only digits and punctuation is considered as numeric date field in our approach, e.g. (11-01-2011, 1/4/04, 11.02.99, etc.). For numeric date extraction we match sub-sequence of components with the following date regular expression:

(d|dd)(/.,\-)(d|dd)(/.,\-)(dd|dddd)

where, d represents digit and we are considering five types of punctuation in the date syntax. A complete numeric date field consists of single digit or double digits date information, single digit or double digits month information and double digits or four digits year information. In our searching algorithm we first find the position of the two punctuation marks. If we get one or two consecutive digit in the left of left punctuation, one or two digits in the middle of these two punctuation and two or four digits on the right of right punctuation, we consider this sequence as a valid numeric date field.

Semi-Numeric date matching: Other date fields that consist of textual month, digits and contraction (st,nd,rd,th) are considered as semi-numeric date. (e.g. May 31, 2010). For semi-numeric date field extraction we are searching the following regular expressions:

(md|mdd)(/.,\-)(dd|dddd) and
(d|dd)(contraction)(month)(/.,\-)(dd|dddd)

where, m represents a month field. There are two types of sequence for semi-numeric date fields. We find the entire pattern in the sequence of line components for matching with any pattern. In a semi-numeric date pattern textual month information may be in the front or in the middle of the sequence. We are accepting a labelled text as contraction followed by numeric date digits if we find a month field between date field and year fields.

IV. RESULT AND DISSCUSSION

To the best of our knowledge, there exists no standard database to evaluate date sequence extraction methods. For our experiment, we have collected 1200 (=10*2*60) lines (10 lines of uppercase letters and 10 lines of normal handwriting) containing date sequence of different valid patterns from 60 individuals of different profession.

A. Training Set

To train our classifier for detecting month blocks on handwritten document, we have used 2570 handwritten months of different forms (capital, small and short) collected from 80 individuals. 3450 handwritten words (non-month) are used for non-month classification. To recognize the digits we train the classifier with MNIST [9] dataset of handwritten digit.

B. Results

To evaluate the quantitative performance of the system, we have used precision (P) and recall (R) measure. Depending on the ground truth of the date, extracted sequence is considered to be valid date sequence or not.

Line selection test: We have tested a total 1200 handwritten lines for date field recognition. A filtering process is used to remove the lines without date pattern matching. The number of digit, punctuation and month components are counted in a line. It is noted, 63.7% (764 out of 1200) lines have been eliminated by keeping the counter to 6. Fig.7 shows the other result of line filtering process.

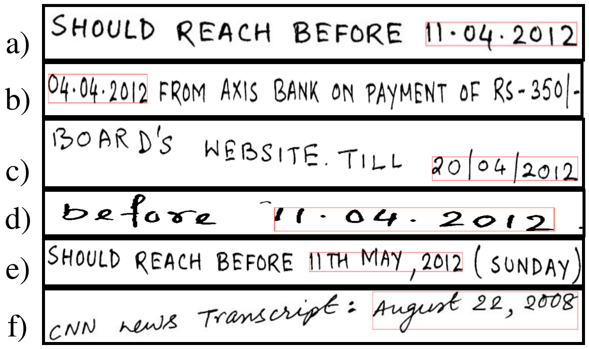

a)
SHOULD REACH BEFORE 11·04·2012
b)
04·04·2012 FROM AXIS BANK ON PAYMENT OF RS-350/-
c)
BOARD'S WEBSITE. TILL 20/04/2012
d)
before 11·04·2012
e)
SHOULD REACH BEFORE 11TH MAY, 2012 (SUNDAY)
f)
CNN news Transcript: August 22, 2008

Figure 6. Results of (a-d) numeric date field. (e,f) semi-numeric date field. Extracted date fields are marked with red box.

TABLE 1. ACCURACY ON DIFFERENT LEVEL

FR	Q1	Q2	Q3	Precision (Q1/Q3)	Recall (Q1/Q2)
Month	72	86	93	77.41	83.72
Date	310	414	348	89.08	74.87

FR:Field for recogntion,Q1: Fields retrieved and relevant, Q2: Relevant fields in dataset, Q3: Fields retrieved

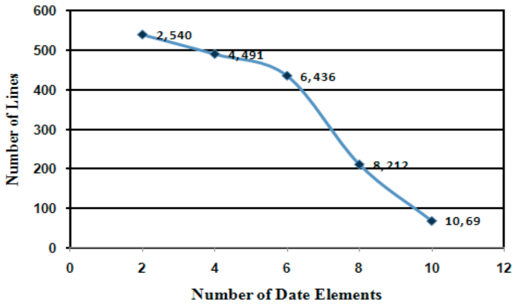


Figure 7. Line filtering results.

Date fields recognition results: Table 1 shows the results of our date pattern extraction experiment. Row 1 shows the performance of month text retrieval and Row 2 shows the overall date retrieval performance. Few examples of date field extraction are shown in Fig.6. The Precision-Recall for numeric and semi-numeric date field are computed separately and the results are presented in Fig.8. It is noticed that most of the errors are generated due to improper classification of textual months and texts. Some errors are found due to over segmentation and touching digit segmentation error in date fields.

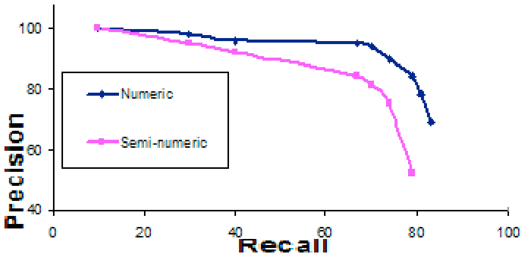


Fig.8. Precision Vs Recall curves for Numeric and Semi-Numeric date field extraction

V. CONCLUSION

In this paper, we have proposed an approach for extraction of date pattern from handwritten documents. To the best of our knowledge, this is the first work of its kind. Component labeling and SVM classification based method is applied to extract month, digit, punctuation and contraction (st, nd, rd, th). Finally, sub-sequence of labeled text lines are matched with date patterns by regular expression. We have obtained encouraging result from the experiment.

REFERENCES

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

[2] S. Thomas, C. Chatelain, L. Heutte and T. Paquet, "Alpha-numerical sequences extraction in handwritten documents", In Proc. ICFHR, pp.232-237, 2010.

[3] G. Koch, L. Heutte and T. Paquet, "Numerical Field Extraction in Handwritten Incoming Mail Documents", In Proc. PRIS, pp. 167-172, 2003.

[4] C. Chatelain, L. Heutte, and T. Paquet, "Segmentation-driven recognition applied to numerical field extraction from handwritten incoming mail documents", In Proc. DAS, pp. 564-575, 2006.

[5] P. P. Roy, U. Pal and J. Lladós, "Morphology Based Handwritten Line Segmentation Using Foreground and Background Information", In Proc. ICFHR, pp. 241-246, 2008.

[6] U. Pal, N. Sharma, T. Wakabayashi and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts". In Proc. 9th ICDAR , pp. 749-753, 2007.

[7] V.Vapnik, "The Nature of Statistical Learning Theory", Springer Verlang, 1995.

[8] P. P. Roy, U. Pal, J. Lladós and M. Delalandre. "Touching Text Character Segmentation in Graphical Documents using Dynamic Programming", Pattern Recognition, vol. 45 (5), pp. 1972-1983, 2012.

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