Problems in Extraction of Date Field from Gurmukhi Documents

Gursimranjeet Kaur[1], Simpel Rani[2]
[1]M.Tech. Scholar Yadwindra College of Engineering, Talwandi Sabo, Punjab, India
sidhus702@gmail.com
[2]Associate Professor, Yadwindra College of Engineering, Talwandi Sabo, Punjab, India
simpel_jindal@rediffmail.com

Abstracts

Date pattern is very important and useful information for any purpose and in any field. Date is the easy way to find the information. Date extraction is done in many handwritten or printed documents of any script like Roman, Bangla, Devanagari, Telugu etc. Date extraction is not a simple task due to presence of many problems like touching digits, handwriting styles, confusion between digits etc. In this paper, we have performed detail literature survey on automatic date extraction. Also we have discussed the problems and challenges which occurs in date extraction field in Gurmukhi handwritten documents.

1.Introduction

Automatic handwritten document retrieval has been evolved as an important research area in recent years. Date is an 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.. Alpha numeric characters that represent date are sometimes touching and confusion between numerals and alphabet makes the task more challenging. In date extraction process there exist numbers of challenges due to different date patterns, writing style of different people, touching characters and classifier confusion between numerals, punctuation and text. An Indian state generally uses three official languages. For example, the West Bengal State of India uses Bangla, Devnagari and English as official languages. Mandal et al. [1] proposed gradient based features and Support Vector Machine (SVM) for classification in both word block and component level classification. Arun et al. [2] described about the connected components in a numerical date field follows a particular structure and can be used for the localization task and target to find all classified date fields in each and every text line of the handwritten document. Umapada et al. [3] tells about a

framework for revival of Bangla 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 grouped into individual components and classified into one of text, digit or punctuation. Partha et al. [4] proposed different classifers for automatic date field extraction from multi-lingual like English, Devanagari and Bangla scripts handwritten documents. Dynamic Time Warping(DTW) and profile feature-based approaches are used for classification of month and non-month. Gradient-based feature and Support Vector Machine are two classifier which are used. Gajanan et al. [5] have proposed a speedy and effectual method for recognition of isolated handwritten Devanagari numeral which are based on JPEG image compression algorithm, it is less time consuming as compared to artificial neural network based recognition systems. Koch et al. [6] proposed method for automatic extraction of numerical field in handwritten incoming mail documents. Nang et al. [7] described about a approach extracting payee's name, legal amount and numerical format from bank cheque by hidden markov model. In this extraction of the information which the user filled in it like amount, name, date etc. Vansi et al. [8] has presented a novel method of automatically recognizing and segmentation of many information which are involves on cheques. They also presented four innovative features :- entropy, energy, aspect ratio and average fuzzy membership values. automatic processing of bank cheque involves recognition and extraction of handwritten information entered by user on their cheque such as courtesy amount, legal amount ,date and signature. Yousef et al. [9] have described the characteristics of Arabic languages and also described the steps that helps to complete the goal containing segmentation, binarization, tagging and validation. It also work on cheque processing which involves all the tasks to verifying the cheque filled by user is correct or not. Rajasehekararadhya et al. [10] have presented a zone-based feature extraction algorithm. The Feed forward Back propagation neural network (BPNN), Nearest Neighbor Classifier (NNC) and Support Vector Machine (SVM) classifiers are used for the classification and recognition of a numeral image. The most challenging task to recognition of handwritten Indian character which in similar shaped component.

There are standard pattern of date format:- DD-MM-YYYY; DD/MM/YYYY; DD.MM.YYYY. We can also write the dates in some other patterns like MAY 12, 2015; 12th MAY, 2015. Dates are used in many different fields like Handwritten documents in administrative, historical archives, postal mail, doctor slip's, defense document, bank cheque, Used for storing proper record for future.

2. Problems and challenges in date field extraction

Extraction of date fields from handwritten Gurmukhi documents is a challenging task a it may involve many problems. In this section, we have discussed the problem encountered during date field extraction.

2.1. Problem in handwriting styles of different people

There are two samples of dates which are written in different styles in figure 1. Different style mean there are different types of people who write the date in different-different ways. For example, writing style of the digits 2 & 4 are different in both dates in line 1 & 2 which are marked by red and green circles

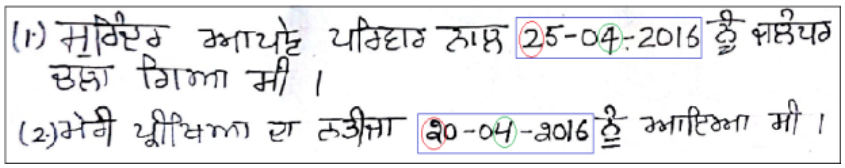

Figure 1: date fields containing different writing style of same digit.

2.2. Problem due to confusion between numeric figures with date fields

As shown in figure 2, there are two numeric values written in lines along with dates. sometime in extraction of date filed there is confusion between numeric values because dates are also in numeric. This makes the date extraction more difficult.

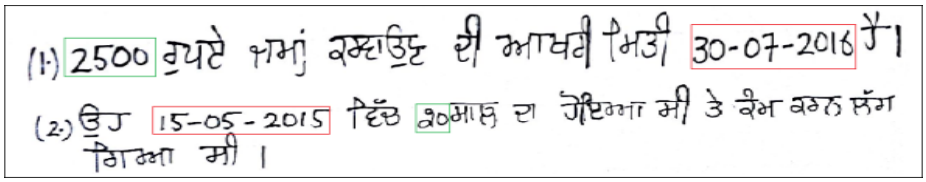

Figure 2: Numeric values (in green rectangle) along with date fields (in red rectangle).

2.3. Problem when month format is different such as numeric style, alpha-numeric

There are two types of date sample in figure 3. One is numeric and other is alpha numeric. Numeric date means the date written in only in numeric values marked by blue rectangle in figure 3. Alpha numeric date is that in which numeric and alphabets are used as shown with pink rectangle. It creates confusion between numeric and alpha-numeric dates. Therefore, it is always a challenging task to extract correct date field.

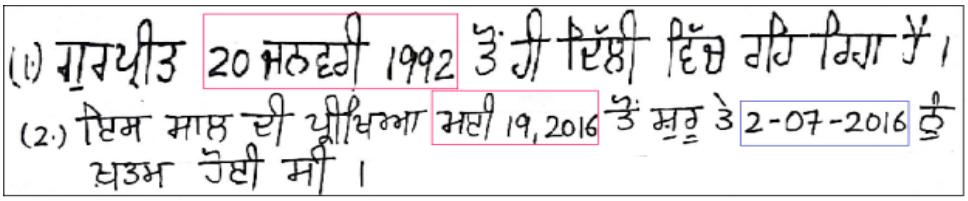

Figure 3: Sample of handwritten dates in numeric and alpha-numeric date field.

2.4. Problem in length of the date patterns are different

In figure 4, there are three samples of dates which are in different pattern like dd-mm-yy, mm-dd-yyyy, dd/mm/yyyy. These all three patterns have different lengths and because of this, date field extraction may be difficult.

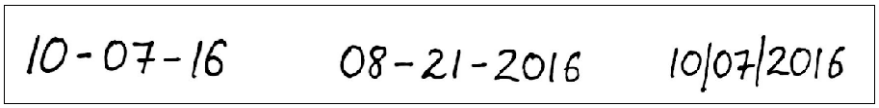

Figure 4: Sample of dates in different lengths

2.5. Problem of touching characters

This problem arises due to different style of writing. While writing the date, one digit touches the other digit as shown in figure 5 marked by black circle. In figure 5 the touching digits are 2, 5 and 2, 0. These touching digits may create problem in extraction of date field and further in recognition of digits.

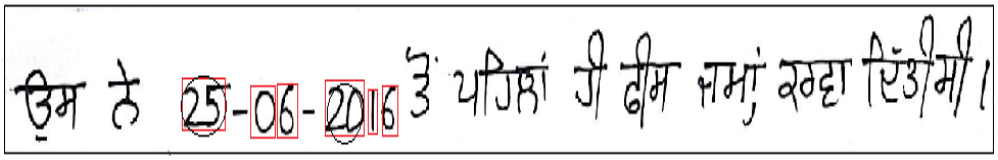

Figure 5: Date filed containing touching digits

2.6. Problem due to poor handwriting

In figure 6, few components of date have been written very poorly and it is very difficult for classifier to correctly recognize the digits. The digit shown in red rectangular box in figure 6 is very difficult for any classifier to classify it as either 5 or 8.

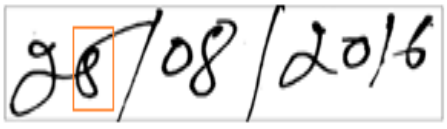

Figure 6: Date written in poor handwriting

2.7. Numeric looks similar to alphabets

This problem arises due to the different writing style of different people. People writes some digits similar to some Roman alphabets which makes it very difficult to extract the dates from database. The date shown in figure 7 marked by green rectangle box contains the digits which are written in date having shape very similar to Roman alphabet. For example:- numeric 0 is looking like Roman alphabet O, numeric 5 is looking like Roman S, numeric 2 is looking like Roman Z.

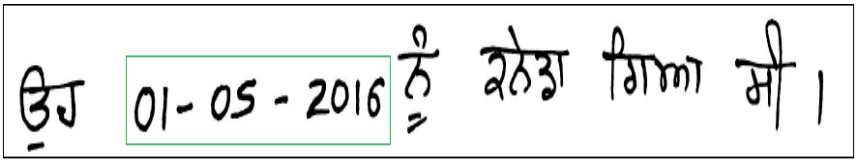

Figure 7: Shows the numeric digits having shapes like Roman alphabets

2.8. Problem of broken digits

This problem appear due to improper writing of digits. Writing style of people produces this kind of problem. The date shown in below figure 8 is broken due to improper writing which are marked by red circles. The broken digits makes the problem of date extraction even more difficult.

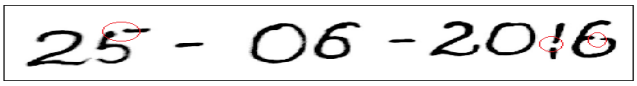

Figure 8: Date containing broken characters

2.9. Merged character

In this problem, one digit is very much merged with another digit, due to different writing style of different people. This also happens due to writing habit or style of an individual and two digits are merged with each other. It creates difficulty in understanding the digits. In figure 9 marked by red circle are merged digits which cannot be understood properly.

Figure 9: Merged digits

3. Conclusions

Date field extraction is one of the important step of OCR. Date extraction is very challenging task due to different writing style of different people. In this paper, we have discussed various

problems which arises in the extraction of date. In future, we are trying to implement the good and suitable method for solving the problems which arises in date field extraction.

4. References

[1] Ranju Mandal, Partha Pratim Roy and Umapada Pal, “Date Field Extraction in Handwritten Documents”, 21st International Conference on Pattern Recognition (ICPR 2012), November 11-15, 2012, Tsukuba, Japan, pp. 533-536.

[2] S. Arunkumar, Pallab Kumar Sahu, Sudeep Gorai and Kalyan Ghosh, “Localisation of Numerical Date Field in an Indian Handwritten Document”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 3, No. 9, 2012, pp 111-114.

[3] Umapada Pal, Ranju Mandal and Partha Pratim Roy, “Bangla Date Field Extraction in Offline Handwritten Documents”, Vol. 2, December 16 2012, Mumbai, India, pp. 37-41.

[4] Partha Pratim Roy, Ranju Manda, Umapada Pal and Michael Blumenstein, “Multi-lingual date field extraction for automatic document retrieval by machine", Information Sciences, Vol. 314, September 2015, pp.533-536.

[5] Gajanan Birajdar and Mansi Subhedar, "Use of JPEG algorithm in handwritten Devanagari numeral recognition", International Journal of Distributed and Parallel system (IJDPS),Vol.2,No.4, July 2011,pp 152-160.

[6] 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.

[7] Nang Aye Aye Htwe, San San Mon and Myint Myint Sein, “Recognition on User-Entered Data from Myanmar Bank Cheque”, pp. 114-118

[8] Vansi Krishna Madasu and Brian Charless Lovell, “Automatic Segmentation and Recognition of Bank Cheque Fields”, In Proceedings of Digital Image Computing: Techniques and Applications (DICTA'05), 2005 Dec 6, pp. 33-33.

[9] Yousef al-ohali, Mohamed Cheriet and Ching Suen, “Databases for recognition of handwritten Arabic cheque”, Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, Amsterdam, ISBN 90-76942-01-3, Nijmegen: International Unipen Foundation, September 11-13 2000, pp 601-605.

[10] S.V.Rajasehekararadhya and P. Vanaja Ranjan, "Handwritten of Numeral/Mixed numerals recognition of south Indian scripts:The zone based feature extraction method", Journal Theoretical and Applied Information Technology, Vol. 7, No.1, pp 63-79.