Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 6 (2017) pp. 1595-1606

http://www.ripublication.com

Date Field Extraction from Gurmukhi Handwritten Documents

Gursimranjeet Kaur
M. Tech. Research Scholar (Computer Science & Engineering), Yadavindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India

Simpel Rani
Associate Professor (Computer Science & Engineering), Yadavindra College of Engineering, Punjabi University Guru Kashi Campus, Talwandi Sabo, Bathinda, Punjab, India

Abstract

Date field is an important part of any handwritten or printed document. Present days are generally fenced with computerized machine or devices. As a result, there is also a requirement of a system that may extract the date from handwritten scanned document or printed document. Date field extraction from handwritten Document is a challenging task due to different handwriting style of different persons. In this work, we have discussed the extraction of date field from Gurmukhi handwritten document. For Extraction, we have used Morphological operations and BRISK techniques as feature extraction techniques. Artificial Neural Network is used as a classifier to train the datasets by using Levenberg Marquardts algorithm. The method has been evaluated by using the dataset of 100 samples of handwritten numeric data from 0 to 9 in different handwriting styles used as a training data and 40 samples of Handwritten Gurmukhi script documents along with dates and non numeric data are used as testing data. We have achieved approximately 94.4% accuracy to extracting date from handwritten document by using morphological operation and BRISK technique. In this work, we have used the

numeric date format like, DD-MM-YYYY, DD.MM.YYY, DD/MM/YYYY,MM-DD-YYYY, MM.DD.YYY, MM/DD/YYYY e.g. 23-03-1992, 23.03.1992, 23/03/1992 or 03-23-2016, 03.23.2016, 03/23/2016.

Keywords:- Gurmukhi handwritten documents, Feature extraction, Morphological operations, BRISK technique, Artificial Neural Network.

1. INTRODUCTION

Recognition and extraction technologies plays a very important role in information processing. Date is very useful information that could help as key for finding and arranging of handwritten document in many filed. To extract the date from handwritten document is very challenging task due to unconstrained handwriting style of different person. Handwritten data is difficult to recognize due to different handwriting style of writer at same time or different time. Sometime same writer or person is not able to recognize its own handwriting at different time or same time. Among the prominent research work done in this field, Ranju et at. [1] have presented the extraction of date from handwritten document by using Gradient based feature and Support Vector Machine(SVM). They also tells about the framework for extracting the bangla date from handwritten bangla document. By using word level feature , they firstly classified the component in each and every text line into month /non-month and then they segmented the non-month words into individual components and classified them into digits, text or punctuation [4]. Arunkumar et at. [2] have proposed the method in Indian handwritten documents to localize the numerical date. They have used K-NN classifier. With the help of statistical analysis developed the algorithm to recognize the set of connected components in numerical date field. Ranju et al. [3] have described the different classifier to extract the date from automatic document in different languages or scripts like English, Bangla and Devanagari. The month and non- month classified by using Profile feature based approaches and Dynamic Time Warping (DTW). Gradient based feature and SVM classifiers are used to recognize the other numerical and punctuation data from a document. The date pattern are searched by using the information of word and character level components. Morita et al. [5] explained about an HMM-MLP hybrid system so that they recognize the handwritten dates on Brazilian Bank Cheques. The date sub field (day, month ,year) have recognized and segmented by using HMM. Multi layer perception Neural network is used to deal with string of digits. Ching et al. [6] have presented the recognition of handwritten data on cheques like legal amount, courtesy amount, signature, date and payee. Due to analyzing the problem they have focused on date written on bank cheques which is challenging task. Qizhi et al. [7] have explained about the knowledge -based method to segment the date on bank cheques with the different kind of knowledge at different stages of segmentation to improve the

performances and also differentiate between numeric and alphabetic data by using various classifiers. In an another work [8] authors have discussed about the automatic recognition of handwritten date on Canadian Bank Cheques. Qizhi et al. [9] have proposed the recognition of handwritten month which is extracted from Canadian bank cheques. The combination of two classifiers HMM-MLP have used to extract the month from a cheque. Mohit et al. [10] have presented the processing system on cheque field like name, amount and also verify the signature and its authenticity. Koerich et al. [11] have discussed about the solution for data acquisition, MICR data identification, layout analysis, recognition and tracing recovering for extracting the data which is entered by user in bank cheques. Marcelo et al. [12] have developed the methods for recognition of the handwritten word written on Brazilian bank Cheques. The words are classified by the Global feature set and two architecture of artificial neural network. Mohammad et al. [13] have explained the recognition of handwritten courtesy amount and signature by using Neural network on bank cheques. Marisa et al. [14] have explained the first stage of recognition system which is applied on handwritten dates on cheques by using HMM. Vamsi et al. [15] have presented the extraction of signature which are based on a priori information of the documents or cheque by using sliding window. Pal et al. [16] have presented the method to recognition of handwritten numerals in Indian scripts such as Devnagari, Bangla, Oriya, Kannada, Tamil and Telugu by using Quadratic classifier which is based on scheme. The main aim is to use the system for Indian postal automation. Gajanam et al. [17] have described the recognition of handwritten Devnagari numerals by using JPEG algorithm. JPEG image compression algorithm is generated a unique vector to identify the each numerals.

2. OFFLINE HANDWRITTEN CHARACTER RECOGNITION

Optical character recognition is the action of recognizing the handwritten or printed document into machine encoded form. As compare to the recognition of printed document, the handwritten recognition is complicated task due to writing style of individual people or even different writing style of same people. Optical Character recognition is mainly divided into two types:-Printed/type-written character recognition and Handwritten Character Recognition. Handwritten Character is further divided into two types:- offline and online character recognition. Offline handwritten character recognition contain various stages like Image acquisition, pre-processing, segmentation, feature extraction and classification. In image acquisition, input image is converted into computerized form into bitmap image like .png, .bmp, .jpg, .pcx, .tiff etc. In pre-processing the image of document is converted into binary image and various other techniques are used to remove the noise, to make it ready and appropriate before feature extraction. Segmentation is used to partition the digital image into multiple segment. Feature extraction is the most important phase in

character recognition. Final phase is classification, which is used to extraction or recognization of character or words based on feature extracted in previous steps or phase. The diagram of steps involved in offline handwritten character recognition is shown in Figure 1.

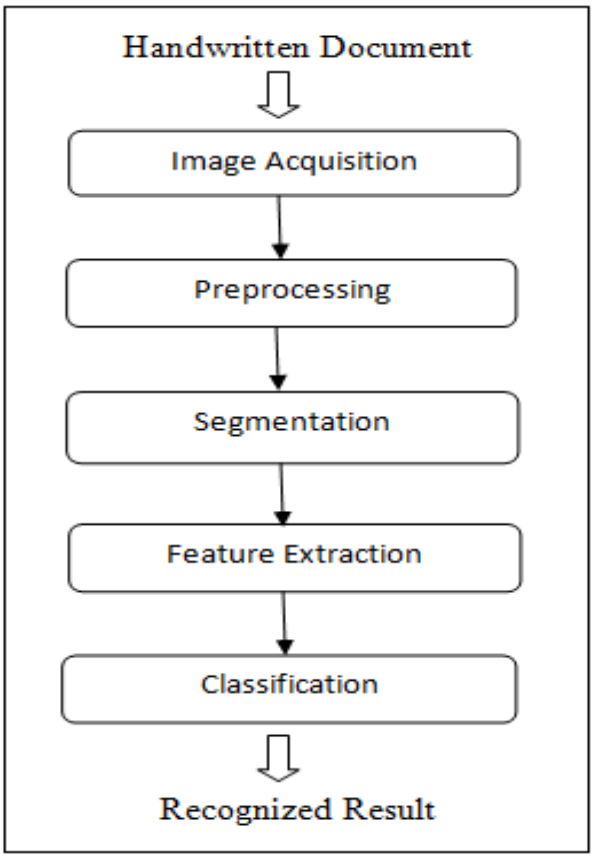


Figure 1. Steps involved in Offline handwritten character recognition

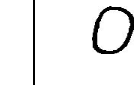
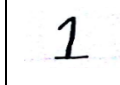
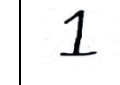
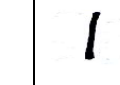
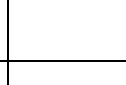
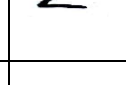
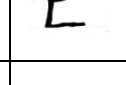
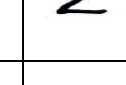
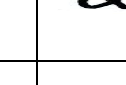
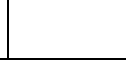
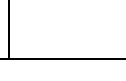
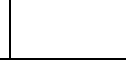
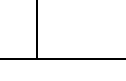
3. DATA COLLECTION

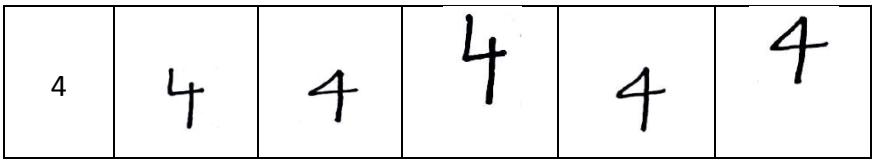
In this work, we have extracted the date field from Gurmukhi Handwritten Documents which is very challenging task due to different handwriting styles of writers or person. We have collected 100 samples of numeric character from 0 to 9 in different handwriting style and also collected 40 handwritten documents with numeric and non-numeric data with different line levels. In Figure 2, we have shown some of the samples of handwritten documents collected.

Figure 2. Sample documents

In this work, we have collected 100 samples of numeric data which is written on A4 size white paper. Each character is written 10 times in different handwriting styles. Table 1 shows some sample of numeric data.

Table 1. Samples of training dataset

Digits	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
0					
1					
2					
3					



4. FEATURE EXTRACTION

Feature extraction is very useful and important step in extraction and recognition of texts, character or digits. Every character or numeral has its own features or shapes which are very important in recognition or extraction of character in system.

4.1 MORPHOLOGICAL OPERATION :- Morphological operation is used to take binary image in pre-processing step. It is also known as Thinning. which is used to remove foreground pixels which are selected from binary image like erosion and opening. Morphology is a huge set in processing operation that process the images based on the shapes. It is applied on structuring elements to an input image and getting an output image of same size. Morphological operation is assembled by selecting the shape and size of the binary image.

4.2 BRISK :- BRISK stands for Binary Robust Invariant Scale Keypoints. It is the combination of both Descriptor and scale space keypoint detections. It is also well-known as Descriptor composition, Feature detection and Keypoint matching. It compares the value of Gaussian window which leading between 0 and 1. BRISK techniques is 512 bit binary descriptor which calculate the weighted Gaussian average.

The Methodology step are below:-

Step 1: Design and develop a proper GUI of propose date field extraction from the Handwritten Gurmukhi documents.

Step 2: Develop a code to upload test image.

Step 3: Apply pre-processing on uploaded test image.

Step 4: After the 3rd step apply text segmentation technique and segment each and every text region from the image.

Step 5: Develop a code for the feature extraction from pre-processed test image using the morphological operation and BRISK technique.

Step 6: Initialized artificial neural network with extracted feature sets of uploaded images as an input of artificial neural network. Train artificial neural network according to the feature sets.

Step 7: After that in testing phase extract the date field only from the uploaded image with structure of artificial neural network and calculate the Accuracy of propose work.

5. CLASSIFICATION

Classification is the process of inspecting the feature of character to predict its class. Various classifiers are used for classification. The classification is done after the step of feature extraction in the system and it works according to the feature extracted in the system.

5.1 NEURAL NETWORK :- Neural network is a system of data structure and programs like a human brain. It is being used mainly in various application fields due to the capability of learning. It is also known as Artificial Neural Network. Neural network are generally arranged in the layers which are made up of mainly interconnected nodes. There are three types of layers in neural network are Input layer, Hidden layer and Output layer helps in system to communicate. Input layer is used to given input and where the actual processing is done hidden layer which is connected to output layer where the actual result is obtained.

5.2 LEVENBERG-MARQUARDT ALGORITHM :- The levenberg - marquardt algorithm has developed by Kenneth levenberg and Donald marquardt which is used to solve the non-linear problem by providing numerical solution. The problems occurred in least square curve fitting . In MATLAB ,algorithm levenberg marquardt is used as a trainlm(levenberg marquardt) in neural network to train the data. It is also called Damped least square method which has been assembled to work specially with the loss function which take the form of sum of squared error. It works with the gradient vector and Jacabian matrix without calculating hessain matrix.

The hessain matrix is can be

$$\mathbf{H} = \mathbf{J}^T\mathbf{J}$$

And, the gradient can be calculated as

$$\mathbf{g} = \mathbf{J}^T\mathbf{e}$$

where ' J' is Jacabian matrix that contain the first derivatives of network error with the weight and bias. 'e' is vector of network error. The hessain matrix is much complicated than Jacabian matrix which is computed by a standard backpropagation techniques. The levenberg-marquardt algorithm uses with hessain matrix like:-

$$\mathbf{x}_{k+1}=\mathbf{x}_k-[\mathbf{J}^T\mathbf{J}+\mu\mathbf{I}]^{-1}\mathbf{J}^T\mathbf{e}$$

When the scalar μ is zero, this is just Newton's method, using the approximate Hessian matrix. When μ is large, this becomes gradient descent with a small step size.

6. EXPERIMENTAL RESULTS

Step 1:- When the main menu run then the GUI of main menu is displayed on screen.

Step 2:- The GUI appears on screen click on start button to go to next step and if you want to close it then press the exit button.

Step 3:- The GUI is displayed on the screen after the step 2 where actually the processing is done to extract the date from document.

Step 4 :- While clicking on training panel the data is started to train by the Levenberg marquardt algorithm in neural network .

Step 5 :- When training is done, upload the image by clicking on upload button as the test image is appeared on GUI .

Step 6:-When the test image is uploaded click on pre-processing button, then it divides the image into further image. Binary and reconstructed image .

Step 7:- When pre-processing is done, we segment the line or text from pre-processed image.

Step 8:- When all processing is completed then click on result button and we get extracted date from handwritten document .

Table 2 shows the accuracy results of 10 document out of 40 documents which are taken as a sample of document with different handwriting and also containing dates and non- numeric data along with them.

Table 2. Sample of 10 documents with accuracy.

S. No.	Gurmukhi Images	Real Date in Images	Recognized Date from Images	Percentage Accuracy
1	ਗੁਰੂ ਨਾਨਕ ਦੇਵ ਜੀ ਦਾ ਜਨਮ 14/04/1469 ਨੂੰ ਹੋਇਆ । ਮਾਪ ਜੀ ਦਾ ਵਿਆਹ 24-09-1487 ਨੂੰ ਹੋਇਆ ।	14/04/1469 24-09-1487	14/04/1469 54-09-1487	95
2	26·12·1666 ਨੂੰ ਗੁਰੂ ਗੋਬਿੰਦ ਸਿੰਘ ਜੀ ਦਾ ਜਨਮ ਹੋਇਆ । ਗੁਰੂ ਅਮਰਜ਼ ਜੀ ਦਾ ਜਨਮ 15·04-1563 ਨੂੰ ਹੋਇਆ । ਮਾਪ ਜੀ 30·05·1606 ਨੂੰ ਜੋਤੀ-ਜੋਤ ਸਮਾਂ ਗਏ ।	26.12.1666 15-04-1563 30.05.1606	26.12.1666 15-04-1563 30.05.1600	96
3	ਗੁਰੂ ਨਾਨਕ ਦੇਵ ਜੀ ਸਿੱਖਾਂ ਦੇ ਪਹਿਲੇ ਗੁਰੂ ਸਨ । ਇਹਨਾਂ ਦਾ ਜਨਮ 15-04-1469 ਨੂੰ ਹੋਇਆ ਸੀ । ਇਹਨਾਂ ਦਾ ਵਿਆਹ 24·09·1487 ਨੂੰ ਮਾਤਾ ਸੁੱਲਖਣੀ ਜੀ ਨਾਲ ਹੋਇਆ ਸੀ । 25 ਸਾਲਾਂ ਵਿੱਚ ਚਾਂ ਤੋਂ ਦੂਜੀ ਚਾਂ ਗੁਰਬਾਣੀ ਦਾ ਪ੍ਰਚਾਰ ਕੀਤਾ। ਆਪ ਜੀ 09/22/1539 ਨੂੰ ਜੋਤੀ-ਜੋਤ ਸਮਾਂ ਗਏ ।	15-04-1469 24.09.1487 09/22/1539	15-04-1469 24.09.1487 09/22/1539	100

4	ਗੁਰੂ ਅੰਗਦ ਜੀ ਦਾ ਜਨਮ 31/03/1504 ਨੂੰ ਹੋਇਆ ਸੀ। ਆਪ ਜੀ ਦਾ ਵਿਆਹ 17-01-1520 ਵਿੱਚ ਹੋਇਆ ਸੀ। ਆਪ ਜੀ 29-03-1552 ਨੂੰ 47 ਦੀ ਉਮਰ ਵਿੱਚ ਜੋਤੀ-ਜੋਤ ਸਮਾਂ ਗਏ	31/03/1504 17-01-1520 29.03.1552	32/03 17-01-1520 29.03.1552	83
5	ਹਰਪੀਤ ਨੇ ਬੀ.ਟੈਕ ਦੀ ਪੜ੍ਹਾਈ 14-07-2012 ਵਿੱਚ ਸ਼ੁਰੂ ਕੀਤੀ ਸੀ। ਉਸ ਦੀ ਬੀ.ਟੈਕ ਦੀ ਪੜ੍ਹਾਈ 27-07-2014 ਵਿੱਚ ਖਤਮ ਕੀਤੀ।	14.07.2012 27-07-2014	14.07.2012 57-07-2014	95
6	01-01-2016 ਤੋਂ 15-01-2016 ਤੱਕ ਅਸੀਂ ਜਸ਼ਨ ਮਨਾਏ। 17-08-2016 ਨੂੰ ਅਸੀਂ ਨਵੀਂ ਕਾਰ ਖਿੱਤੀ ਸੀ।	01-01-2016 15-01-2016 17-08-2016	01-01-2016 15-01-2016 17-08-2016	100
7	ਮਹਾਤਮਾ ਗਾਂਧੀ ਦਾ ਜਨਮ 02-02-1869 ਈਸਵੀ ਨੂੰ ਹੋਇਆ ਸੀ। 30/01/1948 ਨੂੰ ਜਦੋਂ ਇਹ ਬਿਰਧ ਮੰਦਰ ਵਿਚ ਪ੍ਰਾਰਥਨ ਲਈ ਗਏ ਤਾਂ ਇਕ ਨੱਥੂ ਰਾਮ ਗਾਡਸੇ ਨਾਮੀ ਵਿਅਕਤੀ ਨੇ ਇਹਨਾਂ ਨੂੰ ਗੋਲੀ ਮਾਰ ਕੇ ਇਹਨਾਂ ਦੀ ਹੱਤਿਆ ਕਰ ਦਿੱਤੀ।	02.02.1869 30/01/1948	02-02-1869 30/01/1948	90
8	ਸਾਇੰਸ ਸਿਟੀ ਦਾ ਨੀਂਹ-ਪੱਥਰ ਸਾਬਕਾ ਪ੍ਰਧਾਨ ਮੰਤਰੀ ਸ੍ਰੀ ਇੰਦਰ ਕੁਮਾਰ ਗੁਜਰਾਲ ਵੱਲੋਂ 17-10-1997 ਨੂੰ ਰੱਖਿਆ ਗਿਆ। 19-03-2005 ਨੂੰ ਇਸ ਨੂੰ ਲੋਕਾਂ ਲਈ ਖੋਲ੍ਹ ਦਿੱਤਾ ਗਿਆ। ਜੰਗੀ ਹਥਿਆਰਾਂ ਲਾਲ ਸੰਬੰਧਿਤ ਗੈਲਰੀ ਵਿੱਚ ਰੱਖਿਆ ਗਿਆ 'ਮਿੱਗ 23/08/1980 ਵਿੱਚ ਬਣਾਇਆ ਗਿਆ ਅਤੇ 1981-1982 ਵਿੱਚ ਭਾਰਤੀ ਹਵਾਈ ਫੌਜ ਨੂੰ ਸੌਂਪਿਆ ਗਿਆ।	17-10-1997 19-03-2005 23/08/1980	17-10-1997 19-03-2005 23/08/198	96
9	ਮੇਰੇ ਪਿਤਾ ਜੀ ਦਾ ਜਨਮ 22/03/1954 ਵਿੱਚ ਹੋਇਆ ਸੀ। ਮੇਰੇ ਪਿਤਾ ਜੀ ਨੇ 01-03-1990 ਤੋਂ 31-03-2015 ਤੱਕ ਸਰਕਾਰੀ ਨੌਕਰੀ ਕੀਤੀ ਅਤੇ 58 ਸਾਲ ਦੀ ਉਮਰ ਵਿੱਚ ਸੇਵਾ ਮੁੱਕਤ ਹੋ ਗਏ।	22/03/1954 01-03-1990 31-03-2015	22/03/1954 01-03-1990 31-03.2015	96
10	ਪਵਨ ਦੀ ਪ੍ਰੀਖਿਆ 01-02-2017 ਨੂੰ ਸ਼ੁਰੂ ਅਤੇ 28-02-2017 ਨੂੰ ਖਤਮ ਹੋਵੇਗੀ। ਉਸ ਦਾ ਨਤੀਜਾ 13-03-2017 ਨੂੰ ਆਵੇਗਾ।	01.02.2017 28.02.2017 13.03.2017	01-02-2017 28.02.2017 13.03.2017	93

Dates are extracted from documents by using Morphological operation and BRISK feature extraction technique .We have obtained approximately 94.4% accuracy. Figure 3 shows the accuracy result of 10 image of the extracted date from handwritten Gurmukhi document

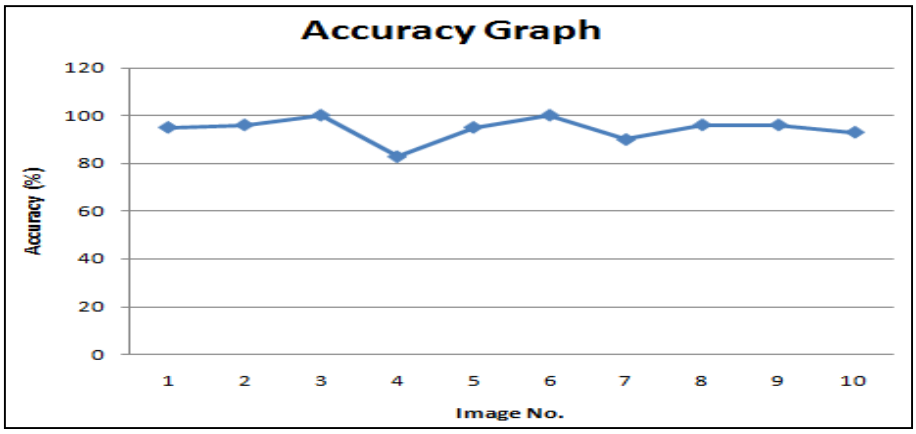


Figure 3. Percentage accuracy of 10 documents.

7. CONCLUSIONS AND FUTURE SCOPE

This paper described the method to extract the date from Gurmukhi handwritten document. The main aim was to extract the numeric dates by using Morphological operation and BRISK technique. Levenberg marquardt algorithm is used in artificial neural network to train the numeric data. We have obtained approximate 94.4% accuracy. In future scope, we can try to work on Gurmukhi dates by another techniques and also increase our stroke of database.

REFERENCES

[1] Ranju Mandal, Partha Pratim Roy and Umapada Pal, "Date field extraction in handwritten documents", *In Proc. International Conference on Pattern Recognition (ICPR),* Vol. 45, 2012, 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 application,* Vol.3, No.9, 2012, pp.111-114.

[3] Ranju Mandal, Partha Pratim Roy, Umapada Pal, and Michael Blumenstein , "Multi-lingual date field extraction for automatic document retrieval by machine," *Information Sciences*, Vol. 314, 2015, pp. 277-292.

[4] Ranju Mandal, Partha Pratim Roy and Umapada Pal, "Bangla Date Field Extraction in offline Handwritten Document", Proceedings of the workshop on Document Analysis and Recognition, 2012, pp.37-41.

[5] M. Morita, R. Sabourine, F. Bortolozzi and C.Y. Suen , "Segmentation and Recognition of Handwritten Date: An HMM-MLP hybrid approach",

International Journal document Analysis Recognition, Vol. 6, 2004, pp. 248-262.

[6] Ching Y. Suen , Qizhi Xu , and Louisa Lam , "Automatic Recognition of Handwritten Data on Cheques-Fact or Fiction", *Pattern Recognition Letters*, Vol. 20, 1999, pp.1287-1295.

[7] Qizhi Xu , Louisa Lam , and Ching Y. Suen , "A Knowledge Based Segmentation System for Handwritten dates on bank cheque", *Proceeding of the 6th International conference on Document Analysis and Recognition*, 2001, pp. 384-388.

[8] Qizhi Xu , Louisa Lam , and Ching Y. Suen "Automatic Segmentation and Recognition System for Handwritten Dates on Canadian bank cheques", *Proceeding of the Seventh International Conference on document Analysis Recognition*, Vol. 2, 2003, pp.704-708.

[9] Qizhi Xu , Jin Ho Kim, Louisa Lam , and Ching Y. Suen "Recognition of Handwritten month words on bank cheques", *Proceeding of the 8th International workshop on Frontiers in Handwriting Recognition*, pp. 111-116, 2002.

[10] Mohit Mehta, Rupesh Sanchati, and Ajay Marchya ,"Automatic cheques processing system", *International Journal of Computer and Electrical Engineering*, Vol. 2, No. 4, 2010, pp.1793-8163.

[11] A.L. Koerich, and L.L. Lee, "A novel approach for automatic extraction of the user entered data from bank checks", *3rd IAPR International Workshop on Document Analysis system (DAS),*Vol. 11, 1998, pp.141-144.

[12] Marcelo N. Kapp, Cinthia O. De. A Freitas and Robert Sabourin , "Methodology for the design of NN-based month-word recognizer written on Brazilian bank checks",*Image and Vision computing,* Vol. 25, 2007, pp. 40-49.

[13] Mohammad Badrul Aiam Miah, Mohammad Abu Yousuf, Md. Sohag Mia and Md Parag Miya, "Handwritten courtesy amount and signature recognition on bank cheque using neural network", *International Journal of Computer Application*, Vol. 118, No. 5, 2015, pp. 21-26.

[14] Marisa Emika Morita, Edouard Lethelier, Abdenaim El Yacubi, Flavio Bortolozzi, and Robert Sabourin, "Recognition of handwritten dates on bank checks using HMM approach", *In: Proceedings of XIII Brazilian symposium on computer graphics and image processing*, 2000, pp.113-120.

[15] Vamsi Krishma Madasu, Mohd. Hafizuddin , Mohd. Yusuf ,M. Hanmandlu, and Kurt Kubik, "Automatic extraction of signature from bank cheques and

the other document", *Proc.7th digital image computing: Techniques and Application*, pp.591-600, 2003.

[16] U. Pal, T. Wakabayashi, N. Sharma, and F. Kimura, "Handwritten numeral Recognition of six popular Indian Scripts", *Proceeding of the 9th international conference on document analysis and recognition (ICDAR),* Vol.2, 2007, pp.749-753.

[17] Gajanam Birajdar, and Mansi Subhedar, "Use of JPEG algorithm in handwritten Devnagari Numerals Recognition", *International Journal of Distributed and Parallel system (IJDPS)*, Vol.2,No.4, 2011, pp.152-160