Dictionary Based Nepali Word Recognition using Neural Network

Ram Chandra Pandey, Babu Ram Dawadi, Suman Sharma, Abinash Basnet

Abstract—The Optical Character Recognition (OCR) systems developed for the Nepali language carry a very poor recognition rate due to error in character segmentation, ambiguity with similar character and has; unique character representation style. The purpose of this paper work is to take image of handwritten or printed Nepali characters and words as input, process the character, train the neural network algorithm, to recognize the pattern and convert to digital form of the input. In this paper, we propose an OCR system for Nepali text in Devanagari script, using multi-layer feed forward back propagation Artificial Neural Network (ANN), to improve the efficiency and accuracy. Adaptive learning rate with Gradient descent algorithm is proposed in Neural net with two hidden layers used with input and output and MMSE is the performance criteria.

Index Terms— Neural network, Nepali handwritten datasets, handwriting recognition, feedforward babk propagation ANN

1.0 INTRODUCTION

UNLIKE English character recognition, Nepali languages are complicated in terms of structure and computations. The Nepali languages is derived from Devnagari Script; written from left to right fashion having common features of containing straight line on top 'Shiro Rekha'. Character or word recognition is the mechanism for converting the text into a notational representation. It is a special problem in the domain of pattern recognition and machine intelligence. Automatic character recognition have many application areas like postal addresses reading, bank check verification, ancient document digitalization, handwritten form verification, forensic and medical analysis, etc. The field of character recognition can be divided into two different categories: on-line recognition and off-line recognition. The on-line case captured trajectory, pan up and pen down time, stroke orders, etc. of the written characters. Off-line mode deals with the recognition of the character or word present in the digital image of written text with holistic image. The early researches after the digital age were concentrated either upon machine-printed text or upon a small scale of well-separated handwritten symbols. Generally, template matching techniques were used for machine printed character recognition while statistical classification techniques were used for handwritten text recognition [1].

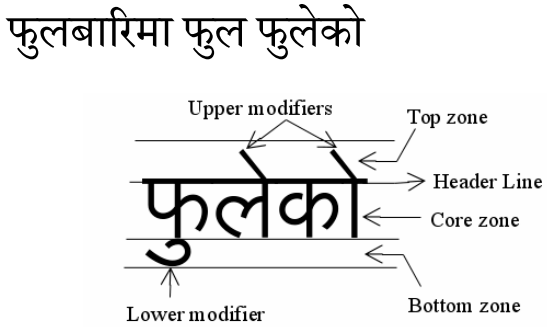

Fig. 1. Different zones of Nepali Word

As Nepali scripts has a horizontal line, which connects all the characters to make a word. This connector is called 'Shiro Rekha' or headline. Based on the horizontal lines, Nepali scripts can be split into three zones: top zone, core zone and bottom zone. The core and top are separated by the Shiro Rekha. Figure 1 shows the image of a word that contains three characters with both lower and upper modifiers.

1.1 Devanagari, the Script

Devanagari is an ancient script used to write languages such as Nepali, Sanskrit, Hindi, Marathi and several others, Nepali is being the official language in Nepal. Although a lot of work has been reported for online handwriting recognition in English and Asian languages such as Japanese and Chinese, there have been very few attempts at online Devanagari handwriting recognition. Thus, the need for more efficient online handwriting recognition algorithms and the under-represented status of Devanagari script set the premise for the work done under this paper.

Artificial Neural Network (ANN) is nonlinear parallel distributed highly connected computational network model having capability of adaptively, self-organization, fault tolerance, evidential response and closely resemble with physical nervous system. This paper not only concerns detecting printed character but also free handwritten characters. Either humans or other computer techniques can use neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, to detect trends and extract patterns that are too complex to be noticed. A trained neural network can be thought of as an expert, which can be used to project new situation.

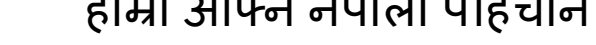

Fig. 2. Common Nepali sentence

Figure 2 shows that sentence is made of words, and each word has numbers of components, i.e, 'matras', above shiro Rekha, they are 'ekar or okar or ikar' and at lower part 'ukar' and etc. During segmentation and hence at recognition, these parts plays the vital role and due to various reason, the efficient system for precise recognition has not yet been developed.

1.2 Special Care for Devanagari

Apart from the precautions stated above some special care has to be taken for Devanagari script because of the complicated segmentation process. Segmentation increases the complication because unlike Latin script, descenders and ascenders of the characters won't be treated as the part of the character in Devanagari script. So no differentiating feature can be presented in the ascender or descender of the character.

Fig. 3. Removal of Core strip

When some important features of the character are at the level of Shiro Rekha or above it gets removed resulting in no recognition or recognizing a different character. For example भ has a curve at the level of Shiro Rekha which when removed results in looking like म. Similarly ध looks like घ when the Shiro Rekha is removed which can be seen in figure 3.

Also a few characters like इ झ ड have characteristic features extending to the bottom strip. When these features are removed the character might closely resemble other characters as shown in figure 4.

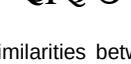

Fig. 4. Similarities between different characters once the bottom strip is removed

Also the graphical similarities in the letters in Devanagari are much more than that in Latin. Some of the letters have just a difference of a stroke like ष just has an additional diagonal stroke as compared to प. Also there are others which differ from each other only because of the presence of vertical line like न and म [4].

2.0 Script Recognition Process

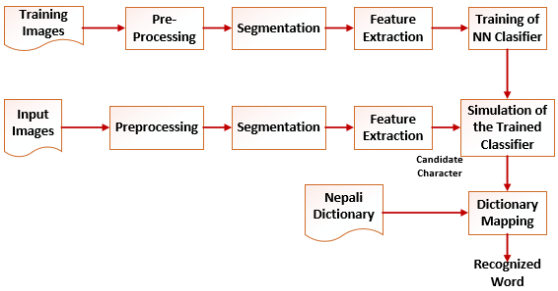

Fig. 5. Training and testing process for generic character recognition system

2.1 Pre-processing

The pre-processing is a series of operations performed on the scanned input image. Preprocessing converts the image into a form suitable for subsequent processing and feature extraction. The role of pre-processing is to segment the interesting pattern from the background. The various tasks performed on the image in pre-processing stage are:-

Edge Detection

Component of the gradient is found using:

$$\frac{\partial f(x,y)}{\partial x} = \Delta x = \frac{f(x+dx,y)-f(x,y)}{dx}$$

$$\frac{\partial f(x,y)}{\partial y} = \Delta y = \frac{f(x,y+dy)-f(x,y)}{dy} \quad (1)$$

Where, f(x,y) is the intensity at x, y coordinate , and the difference in intensity in corresponding coordinate determines edge which is found by different filter, i.e. spatial or frequency domain filter. Sobel operators for a 3×3 mask are given as:

$$\Delta x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

$$\Delta y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{bmatrix} \quad (2)$$

Edge detection helps to extract useful features for pattern recognition. The Sobel operator makes the operator less sensitive to noise.

2.1 Skeletonization

Skeletonization is a process of reducing object regions in a binary image to a skeletal remainder that largely preserves the extent and connectivity of the original object while throwing away most of the original object pixels. Mainly, skeletonize image of text that provides nature of text, which further works as feature that distinguished between varieties of characters.

2.3 Thinning the Image

Since the algorithm is based on the geometrical and structural properties of the Nepali characters, here thin the image to single pixel width so the contours are brought out more vividly. In this way, the attributes to be studied later will not be affected by the uneven thickness of edges or lines in the symbol. Thinning is a morphological operation that is used to remove selected foreground pixels from binary images. The thinning of an image I by a structuring element J is:

$$thin(I,J) = I - hit\_and\_miss(I,J) \quad (3)$$

Where the subtraction is a logical subtraction

2.4 Segmentation

The basic step in Character Recognition is to segment the input image into object from noisy background. This step separates out sentences from text and subsequently words and letters from sentences also. In this stage, an image of sequence of characters is decomposed into sub-images of individual character. After extracting the character need to be normalized to the proper size. In our proposed system, the pre-processed input image needs to isolate characters by assigning a number to each character using a labeling process, known as segmentation. This labeling provides information about number of characters in the image. Each individual character need to be uniformly resized for classification and recognition stage.

2.5 Feature Extraction

Each character has some features, which play an important role in pattern recognition. Feature extraction describes the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure. Feature extraction stage in HGCR system analyses these Nepali character segment and selects a set of features that can be used to uniquely identify that character segment. Mainly, this stage is heart of HGCR system because output depends on these features.

2.6 Classification

During classification, a character is placed in the appropriate class to which it belongs. In training, the back propagation training algorithm subtracts the training output from the target to obtain the error signal then goes back to adjust the weights and biases in the input and hidden layers to reduce the error. Here more than 70 characters (36 Nepali alphabets, 12+ vowels, and 10 numeric characters, 10+ matras) are to be distinctly classified, so it needs many samples and hence takes time to train the neurons.

2.7 Post Processing

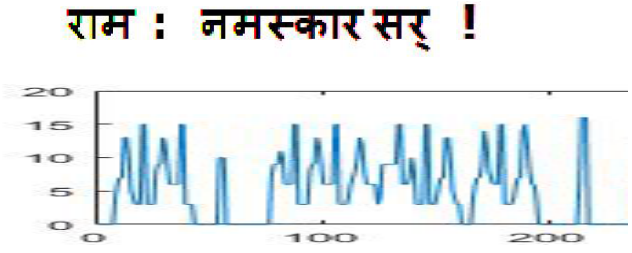

Fig. 6. Vertical Projection Profiles of a document for word separation

Evaluate MMSE which combines the train and test classifier and does for all trained classifiers. If MMSE is less then defined value then directly goes for next character recognition, which reduces execution time. Word segmentation is done in the same way as line segmentation but in place of horizontal profiling, vertical projection profiling is done as indicate in figure 6.

2.8 Training of Classification

Train the neural network such that with input data and target data clustered and returns the network after training it. The learning starts when all of the training data was showed to the network at least once. For every network learning algorithm, consists of the modification of the weights and use the gradient of the criteria field to determine the best weight/modification to minimize the mean square error. Training data sets are shown in figure 7 with other peculiarity of Devanagari character.

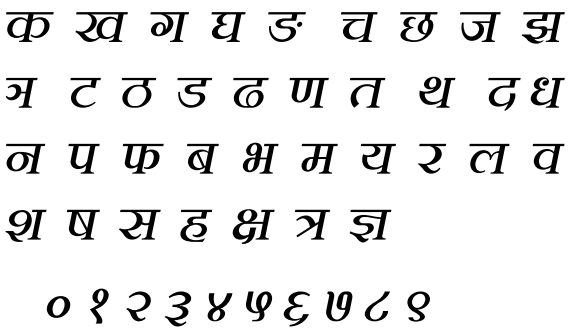

Fig. 7. Nepali alphabets and numeric character.

3.0 Result and Analysis

This paper has main purpose to improve accuracy in Nepali word recognition that uses preprocessed image as input and processes using Feed-Forward back-propagation neural network. In de-noising process: - thresholding is performed, first scanned colored RGB image is converted to gray scale and then gray scale image is converted to binary. Preprocessing has done to improve the accuracy of the recognition algorithm. Main steps in preprocessing are salt and pepper noise removal, binarization, and skew correction. Then boundary of each character is detected from histogram analysis 'Shiro Rekha' is detected as it has highest number of lower values of intensity pixels in text as shown on right portion of figure 8.

Intensity distribution gave the position of line, which had been used to separate different lines. Segmented separated single line is shown in figure 8, in addition with its horizontal and vertical intensity distribution plot. This clearly shows upper part of the horizontal histogram that "Shiro-Rekha" has maximum value as in all Nepali characters from which position of Shiro Rekha is further used in detection of upper and lower part and hence segmentation.

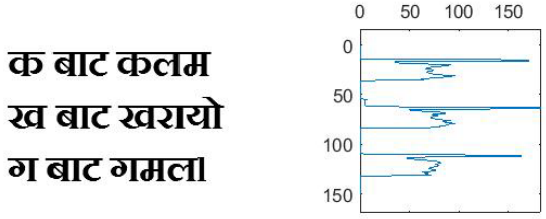

Fig. 8. Horizontal line Profiles of a document for line segmentation

This plot has got meaning in finding character position. It

shows the distribution varies character to character, so individual character can be identified with this shape of distribution only. But, it will arise several problems as two different characters can have similar shape. Therefore, this has two main advantages, firstly, to separate word to word as the Shiro line breaks while separating words, secondly, as a supporting feature to identify a character through its shape.

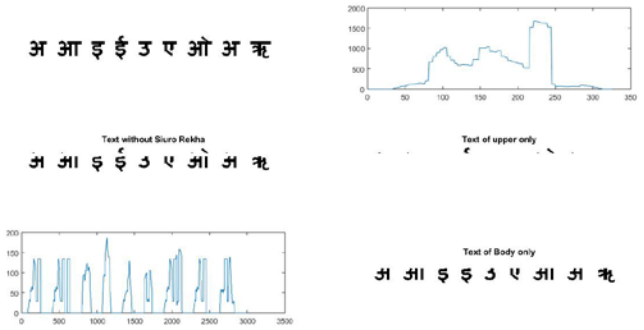

Fig. 9. Steps to make individual character

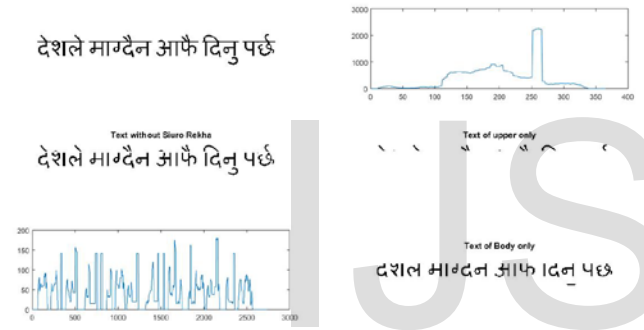

Fig. 10. Steps to make individual character for complex word

Stepwise illustration of finding different part of a single line is shown in this figure 9 and figure 10. Basically, it shows step from de-noised scanned image is fed to the system to separation of different part. From horizontal distribution, the position of Shiro line was detected, which separates different portion of text. While removing Shiro, it was very easy to distinguish various segments in segmentation. In segmentation, firstly area of three major different segments was estimated and then each of the segments were placed in different matrix and processed to find its features.

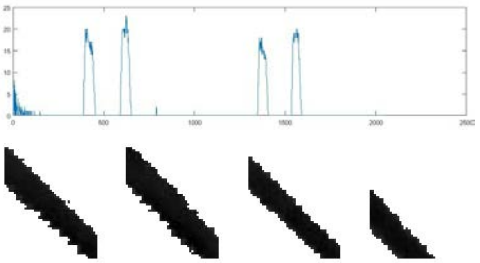

Fig. 11. Processing of upper part of line, and crop

After finding region of different part of text, segmentation was done to find individual component of upper part. For that, taking only upper part in consideration, presence of each component was found and stored the component character along with its coordinate. Histogram of upper part and obtained character is shown in figure 11.

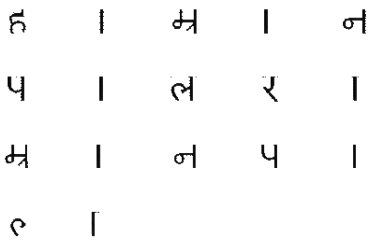

Fig. 12. Processing of Body part of line, and crop

As different part of text region has been stored separately, segmentation was done to find individual component character of body part. For that, taking only body part in consideration, presence of each component was found and stored the component character along with its coordinate. Each character has different shape, size, and length, for example, 'Ra' is shorter than any other character and so on. This creates problem in segmentation. Our approach was to find character by continuity check, which is based on hypothesis that from start to end of each character, there exist a continue intensity distribution. In addition to this hypothesis, Dictionary will reduce the inefficiency occurred due to segmentation. Obtained character is shown in figure 12.

To recognize Nepali word, one should be able to recognize character and all probable cases for the word, i.e. components present above or below to shiro and lower part of word respectively. For that case to work using computer, sufficient training of the component and character must be taken in place. Epochs, iteration required to perform learn, determined by training rate where if learning rate is inversely proportional to epochs. Nature of characters in Nepali language is complex in terms of size, shape, geometry, or other, which leads to ambiguity in similar looking characters and finally the consequence, will be inaccurate word recognition.

In contrast, training should be done in large number of dataset with different characteristics. We took more than 32K dataset for character recognition. Here, number of training data set chosen varies for different characters, for example, 'cha' has trained with less number of dataset as looks very unique and distinct from all other, whereas for 'nga' and 'da', dataset were very large, around 40 for each as shown in figure 12. The segmented character has been tested in trained NN. So, to set weight of each neuron to recognize 36 different characters, dataset of 360 characters have been trained, considering training a character, it needs 10 differently written characters. The resultant neural network architecture is shown in figure 14. In contrast, training should be done in large number of datasets with different characteristics. In figure 15, SSE is shown for anga, as there was less uniqueness in that character.

Creating template for Nepali fonts is comparatively diffi-

cult because for trained image, the size shall varies, i.e. KA, KHA are horizontal and anga is vertical in shape. So program need to run for each character but real condition is whole word is connected by horizontal line; is treated as noises. In addition, accuracy of network will further be increased using dictionary where it searches nearest Nepali word.

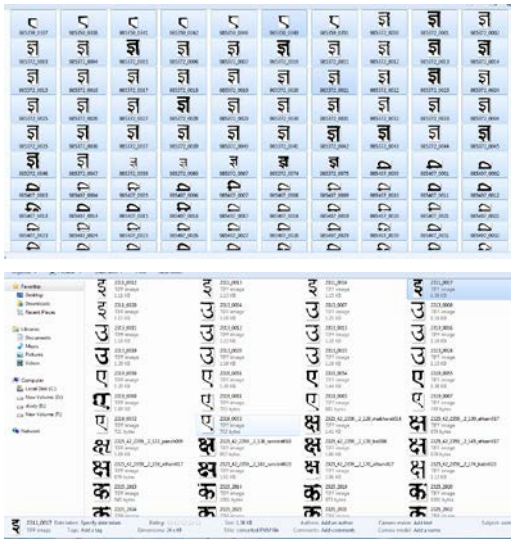

Fig. 13. Dataset of all Nepali characters and components to train NN

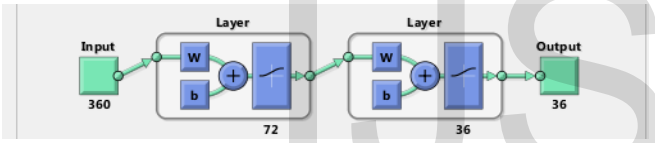

Fig. 14. Trained network architecture for 36 Nepali characters

Computer uses training of the component and character place in sufficient number, to learn character. But for individual dataset, the character should be trained until the recognition cross the definite accuracy level. Epochs means iteration required to perform learn, which determined by training rate shown in figure 16. Here Gradient descent algorithm to estimate learning rate is used, which has higher conversion rate, if learning rate is increases than number of epoch's decreases. Individual segmented character were preprocessed and resized before feeding to trained Neural Network. NN provided the corresponding similar character identity, and from the identity the component is recognized in character level. The trained neural network performed well on simple examples and successfully recognized using other methods previously. Those words which do not have upper and lower part is comparatively simple and has high accuracy in recognition as indicated on figure 17.

As neural network is trained with huge number of dataset for every probable character and component, the system is able to recognize handwritten words like shown in figure 18. It has found from experiment that recognition accuracy for such handwritten word is lower than the printed one. Here 'tha' is recognized in place of 'na' only because of handwritten script 'na' has similar characteristic with 'tha'. And the network is also able to recognize any Nepali fonts with different accuracy level. Result showed its best performance in Mangal font, Unicode compatible, which is freely available and popularly used in Nepalese offices.

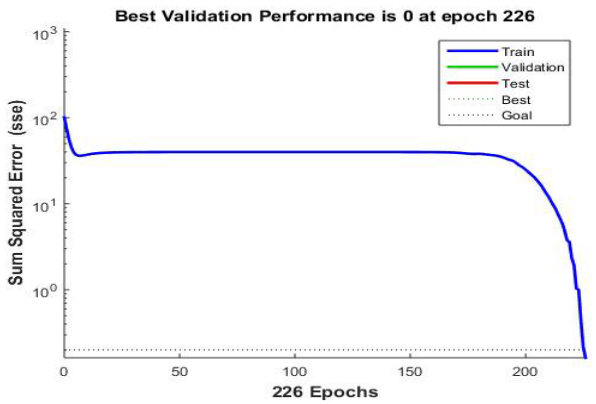

Fig. 15. Convergence of network for fuzzy data

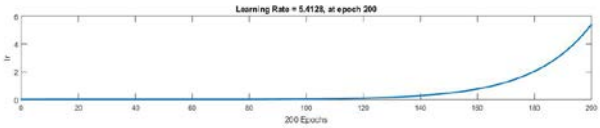

Fig. 16. Learning Rate versus epoch using gradient descent

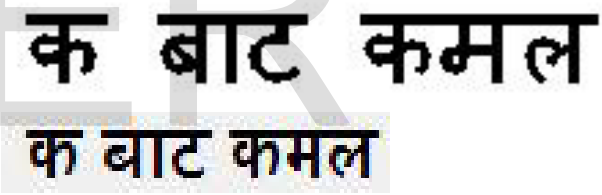

Fig. 17. Nepali word and sentence recognition

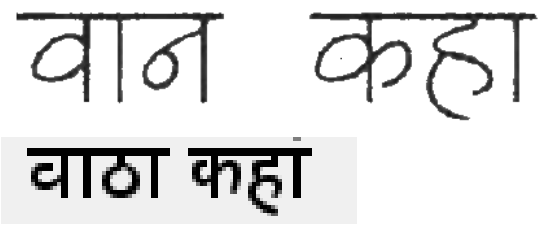

Fig. 18. Network result for handwritten words

The neural net has been trained with not only characters and component but also with half or semi part of that character shown in figure 19. This classifier made from those half characters lies near to the parent classifier so this may lead to misrecognition of the word. It is obvious that accuracy of any system decreases with increase in complexity. Here segmentation and joint classifier increases complexity. Even segmentation also brings inherent recognition problem as those half characters are hard to separate from connected words. Mostly those half words are connected with the next character and separation of that character and half character is hard as these are 8-connected. The length of combined characters is obvious-

ly larger than any longest character so presence of such character is estimated by that length hypothesis and precise segmentation has been achieved.

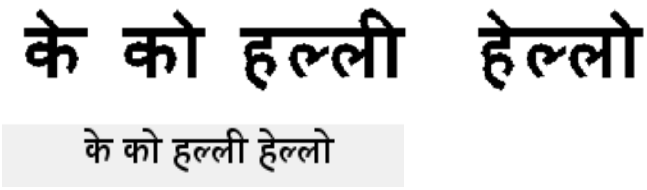

Fig. 19. Recognition including Half letter

4.0 CONCLUSION

In this paper, a solution for Nepali text recognition is demonstrated. Segmentation technique is explored thoroughly in this study and its pros and cons have been revealed. Training with huge number of dataset has done and this lead higher recognition accuracy, about 90% for simple word, 60% for complex word, and near about 50% for handwritten word are achieved. Optical Character Recognition is an open and challenging field for researchers to implement new ideas. The method described in this paper is useful in training Nepali Character dataset. For creating complete OCR system, this method is not very reliable due to similar looking characters and problematic on segmentation process.

The recognition accuracy of the prototype implementation is promising, but more work need to be done. In particular, no fine tuning of the system has been done so far. This character segmentation method also need to be improved so that it can handle a larger variety of touching characters, which occur fairly often in images obtained. Both of these challenges are font dependent. Hence, a detailed font study can help in finding good solutions for these challenges. Moreover, using this dataset exiting Nepali OCR can improve its accuracy.

REFERENCES

[1] Pant, Ashok Kumar, Sanjeeb Prasad Panday, and Shashidhar Ram Joshi. "Off-line Nepali handwritten character recognition using Multilayer Perceptron and Radial Basis Function neural networks." 2012 Third Asian Himalayas International Conference on Internet. IEEE, 2012.

[2] Banumathi, P., & Nasira, G. M. (2011, July). Handwritten Tamil character recognition using artificial neural networks. In Process Automation, Control and Computing (PACC), 2011 International Conference on (pp. 1-5). IEEE.

[3] Arnold, Rókus, and Póth Miklós. "Character recognition using neural networks." Computational Intelligence and Informatics (CINTI), 2010 11th International Symposium on. IEEE, 2010.

[4] Mani, Nallasamy, and Bala Srinivasan. "Application of artificial neural network model for optical character recognition." Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on. Vol. 3. IEEE, 1997.

[5] Alpaydin, E. "Optical character recognition using artificial neural networks." Artificial Neural Networks, 1989., First IEE International Conference on (Conf. Publ. No. 313). IET, 1989.

[6] M.Gunasekaram, S.Ganeshmoorthy,"OCR Recognition System Using Feed Forward And Back Propagation Neural network", Department of MCA ,Park College of Engg & Tech, Coimbatore.

[7] Puri, S. (2010). U.S. Patent No. 7,747,070. Washington, DC: U.S. Patent and Trademark Office.

[8] C. Y. Suen, M. Berthod, and S. Mori, "Automatic recognition of handprinted characters: The state of the art," Proceedings of IEEE, vol. 68, no. 4, pp. 469–487, Apr. 1980.

[9] V. K. Govindan and A. P. Shivaprasad, "Character recognition: A review," Pattern Recognition, vol. 23, no. 7, pp. 671–683, 1990.

[10] R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11, no. 1, pp. 68–83, Jan. 1989.

[11] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR research and development," Proceedings of IEEE, vol. 80, no. 7, pp. 1029–1058, Jul. 1992.

[12] C. C. Tappert, C. Y. Suen, and T. Wakahara, "The state of the art in on-line handwriting recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-12, no. 8, pp. 787–808, Aug. 1990.

[13] J. R Ward and T. Kuklinski, "A Mode l for Variability Effects in Handprinting with Implications for the Design of Handwriting Character Recognition System", IEEE Transactions on Systems, Man and Cybernetics, Vol. 18, No.3, May/June 1988.

[14] H. Arakawa, K. Odaka and 1. Masuda, "On-line recognition of handwritten characters Alphanumeric, Hiragana, Katakana, Kanji", Proc. 4th Int. Joint Con$ Pattern Recognition, pp. 810-812, Nov 1978.

[15] C. Kan and M. D. Srinath, "Invariant character recognition with Zernike and orthogonal fourier-mellin moments," Pattern Recognition, vol. 35, no. 1, pp. 143–154, Jan. 2002.

[16] N. Sharma, U. Pal, F. Kimura, and S. Pal, "Recognition of off-line handwritten devnagari characters using quadratic classifier," in Indian Conference on Computer Vision, Graphics and Image Processing, 2006,

[17] K. C. Santosh and C. Nattee, "Template-based nepali handwritten alphanumeric character recognition," Thammasat International Journal of Science and Technology (TIJSAT), Thammasat University, Thailand.

Author Profiles

Ram Chandra Pandey received his BE in Computer Engineering from Tribhuvan University, Nepal in 2009, and Master's degree in Computer Systerm & Knowledge Engineering from Tribhuvan University, Nepal in 2014. Currently he is working as a Senior Engineer at Nepal Telecom.

(rcpandey.chhetri@gmail.com)

Babu Ram Dawadi Received his B.Sc. in Computer Engineering, M.Sc. in Information and Communication Engineering and Masters in Public Administration degree from Tribhuvan University, Nepal. He worked as an Assistant Director for 3 years at Nepal Telecommunications Authority. He is the full time faculty and research scholar at department of electronics and computer engineering IOE Pulchowk Campus, Nepal. His area of interest is Networking, Distributed Computing and Data Mining. (*baburd@ioe.edu.np*)

Suman Sharma received the Bachelor's degree in Electronics and Communication Engineering from Tribhuvan University, Nepal in 2009, and Master's degree in Information and Communication Engineering from Tribhuvan University, Nepal in 2014. Currently he is working as a Senior lecturer at Kathmandu Engineering College, Department of Electronics and Communication Engineering. He is currently the coordinator for Electronics and Communication Engineering. His research interest includes the wireless communication, image processing, adaptive signal processing, and neural network application on wireless communication.

Abinash Basnet received the Bachelor's degree in Electronics and Communication Engineering from Tribhuvan University, Nepal in 2012, and Master's degree in Information and Communication Engineering from Tribhuvan University, Nepal in 2016. Currently he is working at Nepal Telecom in wireless department. His research interest includes the wireless communication, Cognitive radio, image processing, and neural network application on wireless communication.