ONLINE BANGLA HANDWRITTEN WORD RECOGNITION

Nilanjana Bhattacharya[1], Partha Pratim Roy[2], Umapada Pal[3] and Sanjit Kumar Setua[4]*

[1]Bose Institute, Kolkata, India
[2]Indian Institute of Technology Roorkee, India
[3]Indian Statistical Institute, Kolkata, India
[4]University of Calcutta, Kolkata, India

Email: nilibht@gmail.com, proy.fcs@iitr.ac.in, umapada@isical.ac.in, sksetua@gmail.com

DOI: https://doi.org/10.22452/mjcs.vol31no4.4

ABSTRACT

Bangla word recognition is extremely challenging and a limited number of works has been reported on online cursive Bangla word recognition. Bangla is a complicated script and it requires rigorous investigations to implement a better recognition system. While we have sophisticated classifiers like Hidden Markov Models or BLSTM Neural Networks for recognition of complicated scripts, there has been a limited number of comparative studies about the appropriate feature sets for such scripts. In this paper, our aim is to implement an appropriate recognition system for writer-independent unconstrained Bangla online words where a modified feature set is proposed. To construct the modified feature set, we have modified the existing feature sets and included new features to improve the recognition accuracy. We have tested the performances of various existing feature sets and the proposed feature set on a single dataset for fair comparison and reported the comparative results using various lexicons up to 20,000-word lexicon. An HMM-based classifier has been used to test each feature set. Finally, a recognition system is built over the combination of existing and modified feature sets.

Keywords: Online handwriting recognition, online character recognition, Bangla script, Indic script, cursive text recognition, N-Pen++ features, directional chain code feature, Hidden Markov Models.

1.0 INTRODUCTION

Bangla text recognition is more challenging than Latin text recognition in several ways. The Bangla writing system is more complex. There are more than 300 characters (including 50 vowel and consonant basic characters, more than 250 compound characters, and 11 vowel and consonant modifiers) in the Bangla alphabet. Detailed information about the basic characters, modifiers and compound characters can be found in [1]. Compound characters and modifiers can be attached beside, or to top or bottom of a basic character. Bangla is written from left to right. Generally, Bangla writing is cursive. The presence of many similar-shaped symbols and variability in writing make it more difficult to recognize.

There exists a limited number of works on online cursive Bangla word recognition. While sophisticated classifiers like Hidden Markov Models or Neural Networks have become standard classifiers for recognition of complicated scripts, there are a limited number of comparative studies about the appropriate feature sets for such scripts.

In this paper, our aim is to implement an appropriate recognition system for writer-independent unconstrained Bangla online words. Fig. 1 shows the sequence of experiments for the proposed method. Here, input data consists of a sequence of online points. At first, pre-processing is done. After interpolation of a continuous trajectory and removal of duplicate points, the skew is corrected. Next, we detect three zones/areas of writing to normalize the word size. Details of the pre-processing steps can be found in our work in [2]. After pre-processing, we generate three sets of features - Feature Set-1 contains the popular N-Pen++ features [3] which have been successfully applied for recognition of other scripts [4], Feature Set-2 contains the Mod-NPen features introduced in this paper, and Feature Set-3 consists of only one feature – the directional chain code feature [5] which has been used in offline recognition. We have built three HMM-based classifiers namely System-1, System-2 and System-3 based on the

three different feature sets discussed above. Experiments have been done on various lexicon sizes of up to 20,000 words. For improved recognition, the final classifier is built using a combination of features (Feature Set-2 and Feature Set-3) and the different comparative results are reported.

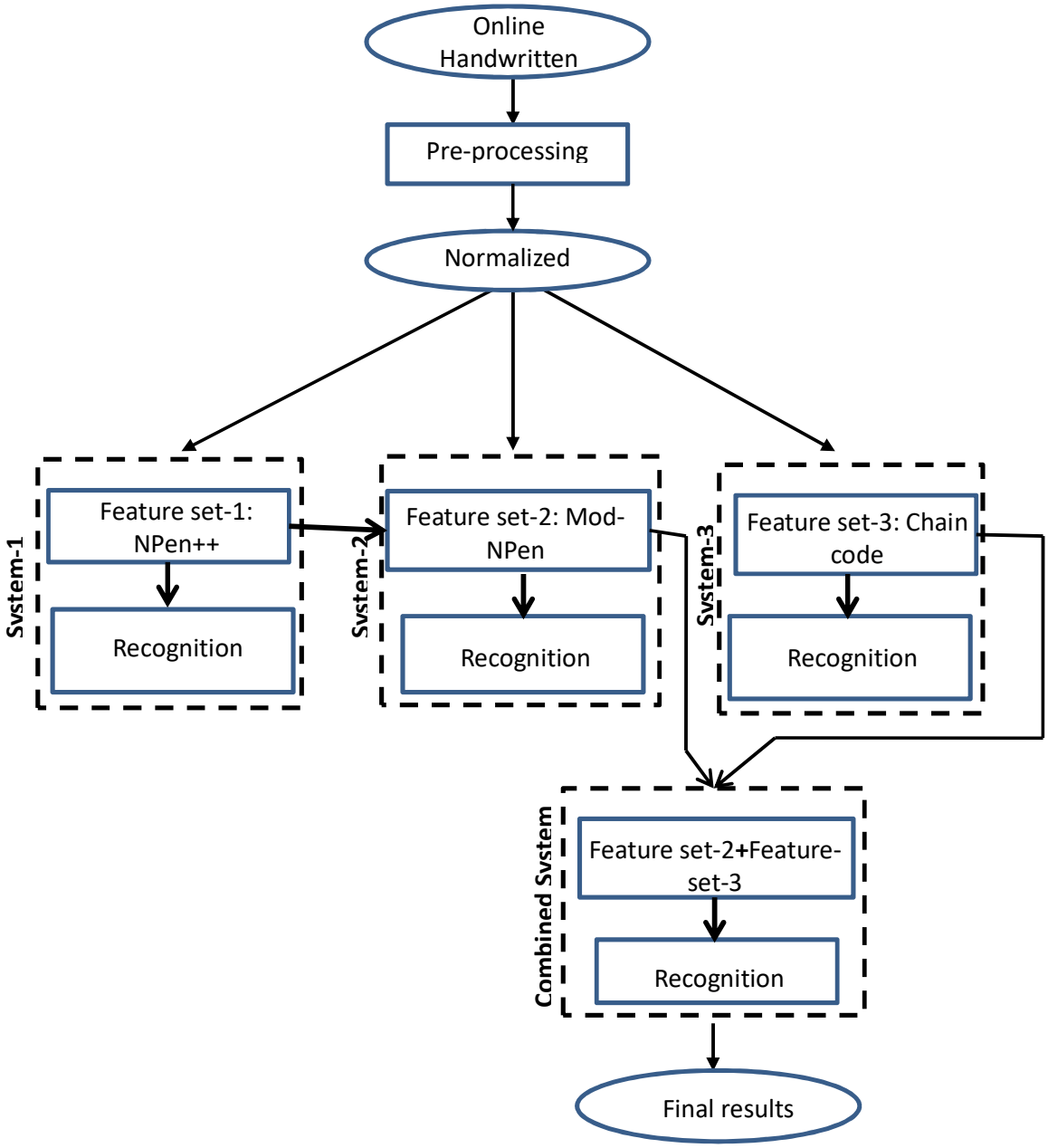

Fig. 1: Flowchart of our proposed system

The rest of the paper is organized as follows. In Section 1.1, a brief overview of related works on online recognition is given. Section 2 presents the feature sets. Classifiers and 'Combined System' are described in Section 3. Section 4 contains the details of data collection and lexicon formation. Experimental results and discussion are presented in Section 5. Finally, the conclusion is presented in Section 6.

1.1 Related Work

In [6], a survey on online recognition of different scripts could be found. Some recent works on experiments with different features for non-Indian scripts were reported in [7-11]. A few works were available on cursive text recognition of online Indic scripts. The work in [12] described a recognition system based on explicit segmentation of Bangla online handwritten words. From Bangla handwriting, 85 stroke classes were identified. Cursive words were explicitly segmented into primitives using an automatic rule-based segmentation algorithm. Next, the resulting primitives were classified using a SVM classifier. Grid-based 64 chain code histogram features were used. A similar method was described in [1] for words consisting of compound characters. In [13], each stroke was divided into sub-strokes, testing the angle encountered in the trajectory. Features used were length, angle, width, and normalized x and y coordinates. Then, HMM was used for recognition. The study [14] proposed an unsupervised feature generation approach based on dissimilarity space embedding (DSE) of local neighborhoods around the points along the trajectory. DSE has a high capability of discriminative representation and hence beneficial for classification. The fundamental idea of the feature extraction process was to first create a fixed set of reference points then use the list of dissimilarities of the current element to the reference points as features. To find the reference prototypes, two unsupervised approaches were investigated - clustering based and random prototype selection. Bidirectional long short-term memory (BLSTM) neural network was used for Bangla words recognition. In the same year, in [15], authors investigated different encoding schemes of Bangla compound characters and compared the recognition accuracies. In the traditional approach, compound characters are represented by unique symbols. As there are large numbers of compound characters in Bangla, authors classified only basic characters and used special nodes (in Neural Network) which react to semantic changes in-between the basic characters to identify compound characters. DSE-based features [14] were used here. The work in [4] addressed lexicon-free and lexicon-driven recognition problems for Type-1 (discretely written), Type-2 (cursively written) and Type-3 (written with delayed strokes) words of Devanagari and Tamil scripts. They used the features introduced in NPen++ recognizer [3] and implemented an HMM-based classifier for recognition.

Main contributions of this work are as follows. Here we address the recognition problem of cursive or mixed cursive Bangla words where a modified feature set (Mod-NPen feature set) is proposed. Our recognition technique uses an analytical approach. We experiment with state-of-the-art features as well as modified features. Finally, a combination of feature sets is implemented in an HMM-based classifier for improved recognition rates. Moreover, a dataset is also created for Bangla online handwriting from individuals of various backgrounds for the experiment. Also, a 20,000-word lexicon is created from the most commonly-used Bangla words to test the feature sets. The database and the lexicon will be available to other researchers.

2.0 FEATURES

Here, we describe three feature sets which are used in our experiments. Feature Set-1 consists of the popular NPen++ features which were originally introduced for online cursive Latin text recognition in [3] and later were used for online cursive word recognition of two Indian scripts (Devanagari and Tamil) in [4]. Through experiments, we have modified the feature set for higher recognition accuracy on the Bangla dataset and the resulting feature set is called as 'Mod-NPen' feature. Mod-NPen features form Feature Set-2. Feature Set-3 is the traditional directional chain code feature [5]. This feature has been extensively used for offline recognition. It has also been used in a few works for isolated online character recognition. It is assumed that our research is the first in testing the performance of this feature for online cursive text recognition.

Feature Set-1: N-Pen++ Features

This feature set contains N-Pen++ features, introduced in N-pen++ recognizer [3] for the Latin text dataset. We choose to implement N-pen++ features because these features were successfully used for cursive word recognition of two Indian scripts (Devanagari and Tamil) in [4]. The features are - Vertical position, Writing direction - sin α and cos α, Curvature - sin β and cos β, Pen-up/pen-down, "Hat"-feature, Aspect, Curliness, Linearity, Slope - cosine of angle, Ascenders/Descenders, and Context maps. Detailed definitions of the features can be found in [3].

Feature Set-2: Mod-NPen Features

To improve the recognition accuracy of the NPen++ feature set, we use the 'Mod-NPen' features customized for the Bangla dataset. To form Feature Set-2, some of the features from Feature Set-1 are deleted during a subset selection experiment and some features are included/replaced for increased recognition accuracy. The features taken are: Writing direction - sin α and cos α, Curvature - sin β and cos β, Aspect, Curliness, Slope, and Context maps. For Slope, sine of angle is added. The linearity feature is replaced with a new one. The average squared distance between every point in the vicinity of (x(t), y(t)) and the straight line joining the first and last point in this vicinity is called the "Linearity". In Feature Set-2, we use "Modified Linearity" which is the average squared distance between every point in the vicinity of (x(t), y(t)) and the center-of-gravity of the points in the vicinity."Modified Linearity" is a novel feature introduced in this work. If the position of first and/or last point in the vicinity changes, "Linearity" value changes a lot, but "Modified Linearity" is a more noise resistant feature, as the center-of-gravity depends on all the points of the vicinity. The 18-dimensional Feature Set-2 provides a notable improvement of recognition accuracy over Feature Set-1.

Feature Set-3: Directional Chain Code Feature

The third feature set contains only one feature - directional chain code feature [5], which, at any online point, takes into account only the influence of the previous point. This feature has been extensively used for finding the direction along the contour of an offline image for offline handwriting recognition. It has also been used in a few works for isolated online character recognition. Perhaps, no one has tested the performance of this feature for online cursive text recognition to date. To compute this feature, the change in direction while moving from one point to the next in a writing trajectory is calculated and is quantized into one of 8 possible values, from 0 to 7 according to Freeman's direction code as shown in Fig. 2.

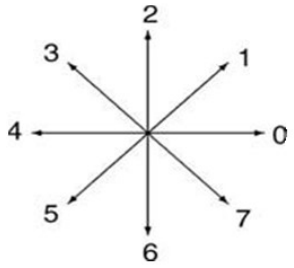

Fig.2: Chain code directions for feature computation

3.0 CLASSIFIERS AND COMBINED SYSTEM

3.1 Classifiers

Here, a Hidden Markov Models (HMM) based classifier [16-18] is used for explicit-segmentation-free recognition. HMM has been proved to be successful to classify temporal patterns [19]. For 65 character classes, 65 HMMs are built. Each HMM is capable of modelling variations in a single model via different paths. Each HMM is composed of a constant number of hidden states ($S_1, S_2, ..., S_N$) using a left-to-right linear Bakis topology. The first state is the input state and the last state is the output state. Other states emit observable feature vectors O whose output probability distributions are modelled by a Gaussian Mixture Model (GMM). For a model λ, if an observation sequence $O = (O_1, O_2, ..., O_T)$ is generated by a state sequence $Q = (Q_1, Q_2, .., Q_T)$, the observation's likelihood is given by:

$$P(O, Q|\lambda) = \sum_Q \pi_{q1} b_{q1}(O_1) \prod_T a_{qT-1\,qT} b_{qT}(O_T) \qquad ----(1)$$

Where π_1 is the initial probability of state 1, a_{ij} is the transition probability from state i to state j and $b_i(O)$ is the output probability of state i. The number of states and Gaussian mixture densities are found through experiments.

For training, sequences of features extracted from the word along with the transcription of the word are used. The Viterbi algorithm is used for recognition. The output is the transcription of the most likely word from the lexicon. The HTK toolkit [20] is used here for implementation.

We have implemented three separate HMM-based classifiers for Feature Set-1 (N-Pen++ features), Feature Set-2 (Mod-NPen features) and Feature Set-3 (directional chain code feature). We name these classifiers as System-1, System-2 and System-3, respectively. Dictionary (Lexicon) information is used in all three systems.

3.2 Combined System

The goal of the experiments on different feature sets is to design a better recognition system. It has been observed from our experimental results that some systems outperform other systems; the sets of words misclassified by the different systems are not the same. It shows that different feature sets offer complementary information for recognizing different shapes. Thus, we are motivated to make the final classification decision by combining individual systems. Here, System-2 is a modified version of System-1 and from experimental results we see that System-2 provides better accuracy than System-1, i.e. System-2 is an improved version of System-1. Hence, we combine System-2 and System-3 which are based on Mod-NPen features and the directional chain code feature, respectively. The combination strategy is simple. It is just feature-level combination, i.e. 'Combined System' is an HMM-based classifier built upon a feature set which includes Mod-NPen features as well as the directional chain code feature.

4.0 DATA COLLECTION

A dataset of online handwritten Bangla words is not freely available to date. Our dataset consists of 11,800 Bangla words from 205 writers. Each writer contributed at least 50 words. We use an iball A414 Take Note tablet. Writers are from different origins and educational standards. Handwritten words obtained are cursive or mixed cursive. To form the dataset, 250 words (word classes) are chosen so that all characters and modifiers are present in the data. To get an idea of the dataset, Table 1 shows the distribution of the lengths of words (along with their frequencies) belonging to this list. Training and testing sets contain distinct sets of inputs. For training and testing, we use 7,800 and 4,000 words, respectively.

Table 1: Distribution of word lengths in our dataset

Length of the word class (i.e. number of characters/modifiers in a word)	**2**	**3**	**4**	**5**	**6**	**7**
Number of word classes	22	67	67	46	33	15
Number of samples	1038	3162	3162	2171	1558	709

Lexicon Generation

To test the influence of vocabulary on the recognition performances of different feature sets, we have prepared a lexicon with 20,000 words. The first 250 entries of the lexicon are taken from the list of 250 word classes. For the remaining 19,750 words, we used 200 WebPages from the online edition of a popular Bangla newspaper named "Anandabazar Patrika".

5.0 RESULTS AND DISCUSSION

5.1 Results of Different Individual Systems

As we mentioned earlier, three systems are developed and they are tested here for a comparative study. System-1 (built upon Npen++ features) provides 92.40% and System-2 (built upon our Mod-NPen features) produces 93.45% word recognition accuracy on a test set of 250 different word classes. System-3 (built upon the directional chain code feature) produces 92.33% word recognition accuracy on the same test set.

We have also investigated the change in accuracy depending on the size of the vocabulary. Fig. 3, Fig. 4 and Fig. 5 show the influence of vocabulary on the performance of System-1, System-2 and System-3, respectively, while keeping the test set constant. Lexicon sizes are increased additively, i.e. each larger lexicon contains some new entries along with all the entries of the previous lexicon. We have tested using 500, 1,000, 2,000, 5,000, 10,000 and 20,000 word lexicons considering the top 1 choice, top 2 choices and top 5 choices of the classifier. In Fig. 3, Fig. 4, and Fig. 5, accuracies obtained from the top choices are shown in blue; while increases in accuracies obtained from the top 2 choices are shown in red; further increases in accuracies obtained from the top 5 choices are shown in green. From the graphs it can be seen that, as usual, the accuracy decreases when the lexicon size increases. For larger lexicons, accuracy increases a lot when we consider more choices; while for smaller lexicons, increase in accuracy is small if more choices are taken into consideration. The accuracies of System-1, System-2 and System-3 on the lexicon of 20K words (considering only the top choice) are 72.58%, 74.30% and 71.68%, respectively.

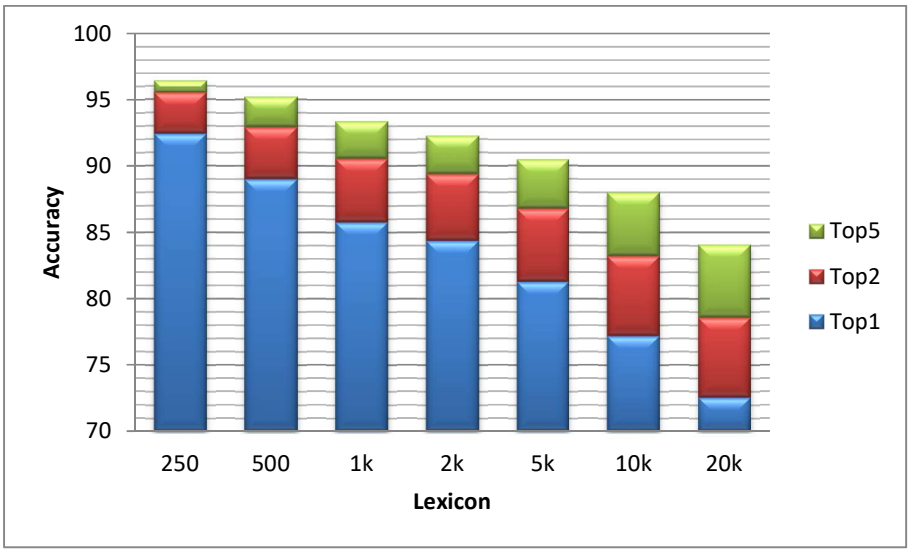

Fig. 3: Influence of lexicon size on performance of System-1

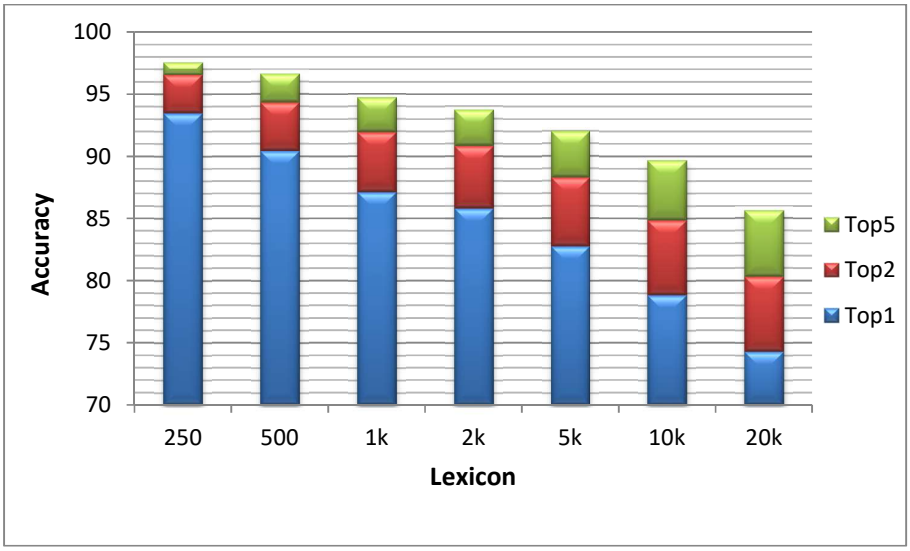

Fig. 4: Influence of lexicon size on performance of System-2

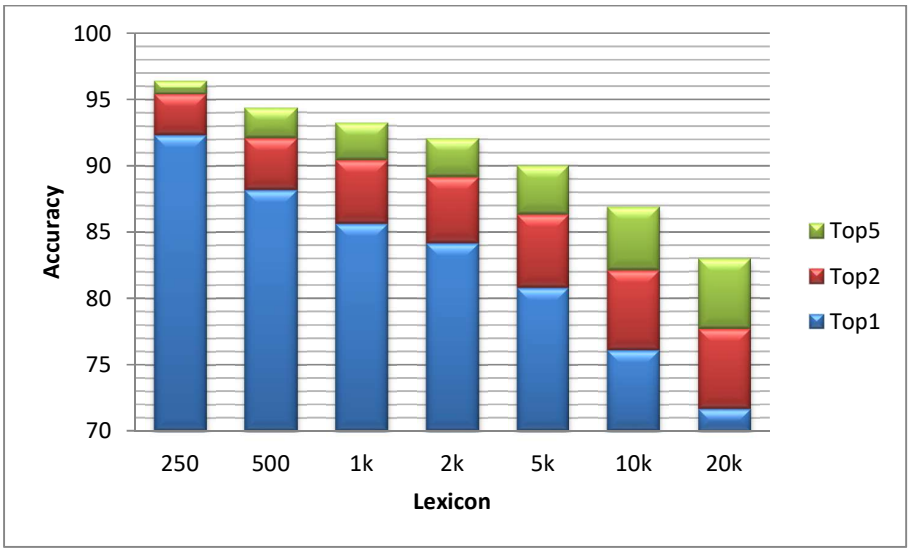

Fig. 5: Influence of lexicon size on performance of System-3

5.2 Results of Combined System

Influence of the vocabulary on performance of Combined System is shown in Fig. 6. As we can see, Combined System provides recognition rates of 94.10% and 76.30% on the test set using 250 and 20,000 word lexicons, respectively. As expected, the performance of Combined System is better than the performances of all three of the individual systems.

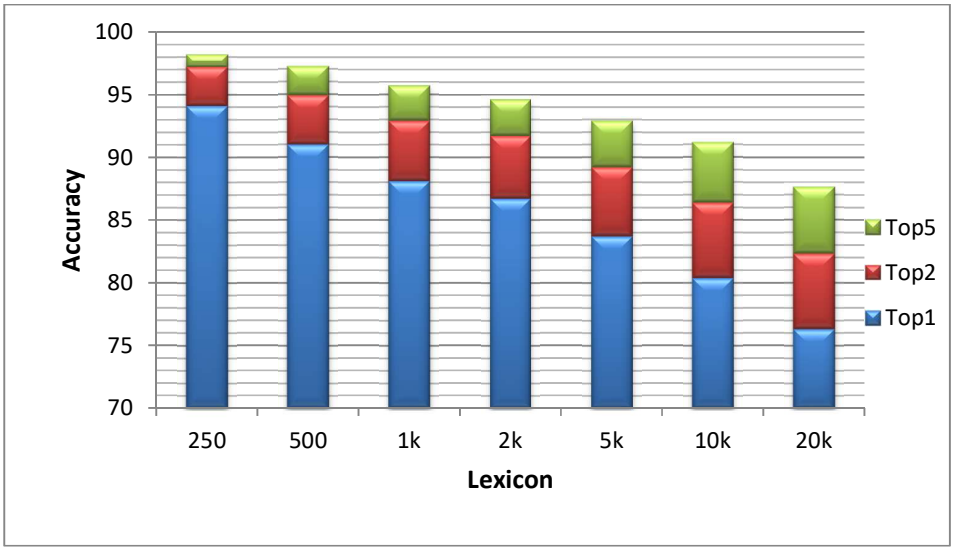

Fig. 6: Influence of lexicon size on performance of Combined System

A graphical representation of the comparative results of different systems on lexicon sizes of 250, 500, 1,000, 2,000, 5,000, 10,000 and 20,000 words (considering the top choice) is also shown in Fig. 7, from which we notice that Combined System is the best (76.30% on 20K lexicon) among all the systems. A little less accuracy (74.30% on 20K lexicon) is obtained from System-2, which is in turn better than System-1 (72.58% on 20K lexicon). Though the minimum accuracy (71.68% on 20K lexicon) is obtained from System-3, it is worth noting that it has only one-dimensional feature. From this work, we have come to know that chain code feature alone can compete with existing feature sets for online cursive handwriting recognition. As usual, for all systems, accuracy decreases with larger lexicons.

From the above discussions (in Sub-section 5.1 and Sub-section 5.2), it can be noted that although for smaller lexicons, the performance of System-2 (built upon our Mod-NPen features) is only marginally better than System-1 (built upon Npen++ features), the accuracies achieved by System-2 are much higher than that of System-1 for larger lexicons. Moreover, Combined System shows a large improvement over the others.

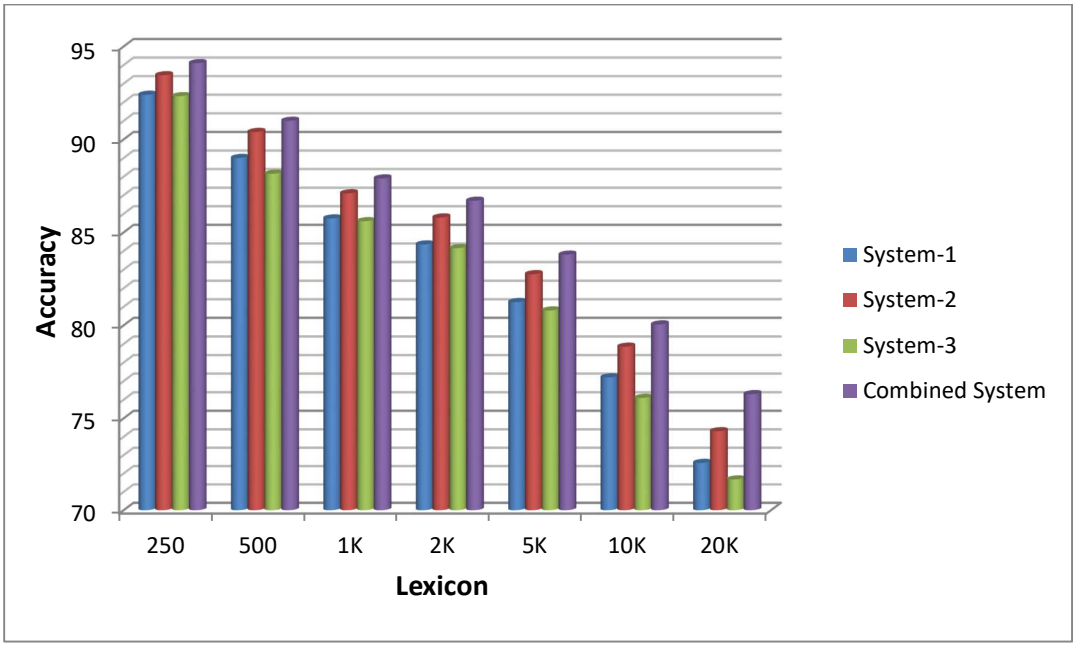

Fig. 7: Visualization of comparative results of different systems

5.3 Comparison of Results with Other Studies

The work on Bangla online word recognition in [13] reported a maximum accuracy of 93.10% for a 50-class problem. The work in [4] obtained 57.55% accuracy on an online Devanagari dataset and 89.82% accuracy on an online Tamil dataset using a lexicon-driven approach. In each case, a 20K-word lexicon was used. Our work shows performance comparisons of different feature sets on the same dataset (Sub-section 5.2). It is noted that Combined System achieves 94.10% accuracy for a 250-class problem (lexicon size is also 250 here) and 76.30% accuracy using a 20K-word lexicon on online Bangla cursive words.

6.0 CONCLUSION

We have addressed a writer-independent unconstrained Bangla online word recognition problem which is yet to get full attention from researchers. In this paper, we select some of the state-of-the-art features which have been proven to be effective for other applications and test their performances on online cursive Bangla words. We also proposed Mod-NPen features (modified over Npen++ features) for better accuracy. Moreover, we propose a recognition system composed of a combined set of modified features for online Bangla word recognition. A dataset is also created with the help of Bengali people with various backgrounds and a 20,000-word lexicon is created from the most commonly used Bangla words to test feature sets. We have chosen an HMM-based classifier for explicit-segmentation-free recognition. Comparative results are reported and it is observed that Combined System provides encouraging word recognition accuracies on large lexicons. As mentioned earlier, the database and the lexicon will be available to other researchers free of cost upon request.

REFERENCES

[1] N. Bhattacharya, U. Pal, and F. Kimura, “A System for Bangla Online HandwrittenText”, *12th Int'l Conf. on Document Analysis and Recognition*, 2013, pp. 1367-1371.

[2] N. Bhattacharya, V. Frinken, U. Pal, and P. P. Roy, “Overwriting Repetition and Crossing-out Detection in Online Handwritten Text”, *3rd Asian Conference on Pattern Recognition*, 2015, pp. 680-684.

[3] S. Jaeger, S. Manke, J. Reichert, and A. Waibel, "Online Handwriting Recognition: The NPen++ Recognizer", *International Journal on Document Analysis and Recognition*, Vol. 3, 2001, pp. 169-180.

[4] A. Bharath and S. Madhvanath, "HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34 No. 4, 2012, pp. 670-682.

[5] S. Madhvanath, G. Kim, and V. Govindaraju, "Chaincode Contour Processing for Handwritten Word Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21 No. 9, 1999, pp. 928-932.

[6] R. Plamondon and S. N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, 2000, pp. 63–84.

[7] J. Schenk and G. Rigoll, "Novel Hybrid NN/HMM Modelling Techniques for On-line Handwriting Recognition", *10th Int'l Workshop on Frontiers in Handwriting Recognition*, 2006, pp. 619–623.

[8] G. Mujtaba, L. Shuib, R.G. Raj, R. Rajandram, K. Shaikh, M.A. Al-Garadi, "Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection", PLoS ONE, Vol 12(2): e0170242, 2017. https://doi.org/10.1371/journal.pone.0170242.

[9] Y.-F. Lv, L.-L. Huang, D.-H. Wang, and C.-L. Liu, "Learning-Based Candidate Segmentation Scoring for Real-Time Recognition of Online Overlaid Chinese Handwriting", *12th Int'l Conf. on Document Analysis and Recognition*, 2013, pp. 74–78.

[10] H. El Abed, M. Kherallah, V. Märgner, and A. M. Alimi, "On-line Arabic handwriting recognition competition, ADAB database and participating systems", *International Journal on Document Analysis and Recognition*, Vol. 14, 2011, pp. 15-23.

[11] K. Islam, and R.G. Raj, "Real-Time (Vision-Based) Road Sign Recognition Using an Artificial Neural Network", Sensors, Vol. 17(4), 2017, pp. 853. MDPI AG. http://dx.doi.org/10.3390/s17040853.

[12] N. Bhattacharya and U. Pal, "Stroke Segmentation and Recognition from Bangla Online Handwritten Text", *13th Int. Conf. on Frontiers in Handwriting Recognition*, 2012, pp.736-741.

[13] G. A. Fink, S.Vajda, U. Bhattacharya, S. K.Parui, and B. B. Chaudhuri, "Online Bangla Word Recognition Using Sub-Stroke Level Features and Hidden MarkovModels", *12th Int'l Conf. on Frontiers in Handwriting Recognition*, 2010, pp. 393-398.

[14] V. Frinken, N. Bhattacharya, and U. Pal, "Design of Unsupervised Feature Extraction System for On-Line Bangla Handwriting Recognition", *11th IAPR Int'l Workshop on Document Analysis Systems*, 2014, pp. 355–359.

[15] V. Frinken, N. Bhattacharya, S. Uchida, and U. Pal, "Improved BLSTM neural networks for recognition of on-line Bangla complex words", *IAPR Joint Int'l Workshops on Statistical Techniques in Pattern Recognition + Structural and Syntactic Pattern Recognition, Lecture Notes in Computer Science,* Springer, 2014, pp. 404-413.

[16] C. Bishop, *Pattern Recognition and Machine Learning*, Springer Verlag, 2006.

[17] F. Jelinek, *Satistical methods for speech recognition*, MIT Press, 1994.

[18] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, Vol. 77 No. 2, 1989, pp. 257-286.

[19] W.Cho, S.W.Lee, J.H.Kim, "Modeling and recognition of cursive words with hidden Markov models", *Pattern Recognition*, Vol. 28 No. 12, 1995, pp. 1941–1953.

[20] S. J. Young et al., *The HTK Hidden Markov Model Toolkit Book*. Entropic Cambridge Research Laboratory, 1995.