DNN-HMM based Large Vocabulary Online Handwritten Assamese Word Recognition System

{S. Mandal, H. Choudhury, SRM Prasanna & S. Sundaram } IIT Guwahati, India

Objectives

1. Development of large vocabulary online handwriting Assamese word recognition system.
2. The use of DNN-HMM framework for online handwriting recognition.
3. Selection of 173 basic units that can characterizes 20K most frequent Assamese words.
4. Creation of Assamese word database.

Introduction

- The majority of the studies in Indic script considered to recognize isolated characters.
- Few works on word recognition task has reported on some of the Indic scripts such as Devanagari, Bangla and Tamil [1].
- The number of characters ranges from few hundreds to several thousand.
- The challenges include lack of common definition in basic unit (BU), availability database etc.

Assamese Handwritten Word Recognition System

- The preprocessing steps are smoothing, resampling, and size normalization.
- The following 16 features are considered: preprocessed x and y coordinates (2); first derivative of x and y coordinates (2); second derivative of x and y coordinates (2); writing direction (2); curvature (2); aspect ratio (1); linearity (1); slope (1); context map (1); ascender (1); and descender (1).
- In the proposed system, a deep neural network is trained to output HMM state probabilities.
- For training, GMM-HMM system is used to obtain alignment of the training data to HMM states.
- Next, we supply the label of each frame to the neural network for DNN training.

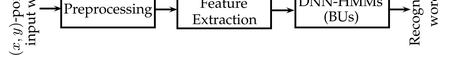

Figure 4: Block schematic of the developed online handwritten Assamese word recognition system.

- The word with highest score among all the lexicon words is declared as recognized word.

Assamese Script and Database Creation

অ (1)	আ (2)	ই (3)	ঈ (4)	উ (5)	ঊ (6)	ঋ (7)	এ (8)	ঐ (9)
ও (10)	ঔ (11)	ক (12)	খ (13)	গ (14)	ঘ (15)	ঙ (16)	চ (17)	ছ (18)
জ (19)	ঝ (20)	ট (21)	ঠ (22)	ড (23)	ঢ (24)	ণ (25)	ত (26)	থ (27)
দ (28)	ধ (29)	ন (30)	প (31)	ফ (32)	ব (33)	ভ (34)	ম (35)	য (36)
ৰ (37)	ল (38)	ৱ (39)	শ (40)	ষ (41)	স (42)	হ (43)	ক্ষ (44)	য় (45)
ড় (46)	ঢ় (47)	ৎ (48)	ং (49)	ঃ (50)	ঁ (51)	া (52)	ি (53)	ী (54)
ু (55)	ূ (56)	ৃ (57)	ে (58)	ৈ (59)	ো (60)	ৌ (61)	্ (62)	ক্ক (63)
ক্ট (64)	ক্ত (65)	ক্ম (66)	ক্ষ (67)	ক্স (68)	গু (69)	গ্ন (70)	গ্ম (71)	গ্ল (72)
গ্ল (73)	ক্ত (74)	ঙ্ক (75)	ঙ্গ (76)	ঙ্ঘ (77)	ঙ্ক্ষ (78)	চ্চ (79)	চ্ছ (80)	জ্জ (81)
জ্ঝ (82)	ঞ্ছ (83)	জ্ঞ (84)	ঞ্চ (85)	ট্ট (86)	ড্ড (87)	ঠ্ঠ (88)	ণ্ড (89)	ত্ত (90)
থ্থ (91)	ন্ত (92)	ন্থ (93)	ন্ম (94)	ন্দ (95)	ন্ধ (96)	ভ্ভ (97)	ম্ন (98)	ট্ট (99)
ত্ত (100)	ত্ত্ব (101)	ন্ত্র (102)	ন্ত্র (103)	ন্দ (104)	ন্দ্র (105)	ন্ন (106)	ন্থ (107)	ন্ম (108)
ঙ্গ (109)	ন্যে (110)	প্ট (111)	গু (112)	প্ত (113)	প্র (114)	ঙ্গ (115)	ব্ব (116)	ৰ্ক্ষ (117)
ৰ্ম (118)	ভ্র (119)	ম্ন (120)	স্প(121)	স্ফ (122)	স্ব (123)	স্ত (124)	ম্ভ (125)	ম্ম (126)
স্র (127)	ক্ক (128)	স্ট (129)	স্ত (130)	স্ত্র (131)	ব্ব (132)	স্ম (133)	স্ত্র (134)	শু (135)
শ্চ (136)	শ্ম (137)	শ্ব (138)	শ্ম (139)	শ্ম (140)	ষ্ক (141)	ষ্ক্র (142)	ষ্ট (143)	ষ্ঠ (144)
স্প (145)	স্ফ (146)	স্ম (147)	স্ক (148)	স্ত (149)	স্ত (150)	স্ত্র (151)	স্থ (152)	স্ন (153)
স্প(154)	স্ফ(155)	স্ব (156)	স্ম (157)	হ্ (158)	হ্ন (159)	হ্ন (160)	হ্ম (161)	হ্ম (162)
হ্ম (163)	ঽ (164)	৲ (165)	৷ (166)	৺ (167)	৻ (168)	৩ (169)	৾ (170)	ঘ্ন (171)
হ্ল (172)	স্থ (173)							

Figure 1: The basic unit (BU) set defined for recognizing 20K most frequent Assamese words.

Fig. 1 presents the identified BU that are used to write Assamese. The BUs 1—11 are vowels, BUs 12—51 are consonants, BUs 52—60 are vowel modifiers, BUs 61 and 62 are consonant modifiers, BUs 63—163 and 171—173 are conjuncts, and BUs 164—170 are few strokes.

(a) ি + ক = কি (s53) (s12) (CV unit)

(b) গ + ্ + ম = গ্ম (s14) (s165) (CC unit)

(c) গ্ম + ু = গ্মু (s71) (s166) (New conjunct)

(d) ষ + ৩ = ষ্ঙ (s41) (s169) (New conjunct)

Figure 2: Generation of new basic unit (BU) by combining the BUs defined in Fig. 1

Set	Frequency (%)
Vowels	5.96
Vowels+Consonants	61.49
Vowels+Consonants+Conjuncts	64.16
Vowels+Consonants+Conjuncts+Modifiers (All BUs + their combination of Fig. 1)	99.76

Table 1: Number of characters (in %) covered by the selected BU set of Fig. 1 in Assamese OCR [2] corpus.

The set of 182 words is considered for data collection and it covers all the BUs in Fig. 1 and the characters formed by the BUs.

Set	# Writer	# Samples
Training	96	16208
Testing	48	8414
Validation	19	2941

Table 2: Samples and writers in Assam. word database.

Result and Discussion

Evaluation Methodology:

- Lexicon: 1K, 5K, 10K and 20K (close set).
- The most frequent of Assamese OCR is considered to create the lexicon.
- Accuracy $= 100 \times (t/n)$, t is the number of correct classification and n is the number of samples tested.
- The Kaldi toolkit is used for DNN-HMM.
- A left-to-right topology with 20 GMMs in each state is used.
- Highest accuracy is obtained for 17 states.

HMM State	GMM-HMM	DNN-HMM (# Hidden layer conf.)		
		L1:300	L2:300-300	L3:300-300-300
13	89.10	93.70	94.15	94.46
15	90.04	94.33	94.42	93.93
17	90.67	94.69	94.69	94.63
19	89.24	94.60	94.62	94.24

Table 3: Tuning of GMM-HMM and DNN-HMM models on validation set, considering 1K lexicon.

- DNN input consists of $(2L + 1)$ feature vectors and chosen L=4 as shown in Fig. 3.
- DNN input node = 144; output node = 2941 (173 HMMs with 17 states in each model).

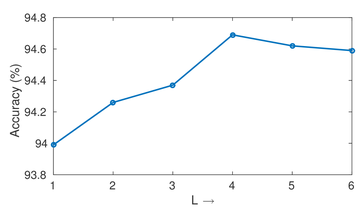

Figure 3: Tuning of the length of context window (L) for DNN training on validation set.

- Best accuracy is obtained for two hidden layers as given in Table 3.

Lexicon size (# word)	GMM-HMM	DNN-HMM
1K	81.15	89.75
5K	77.08	88.72
10K	73.87	88.03
20K	71.13	87.36

Table 4: Word recognition accuracy (in %) of the DNN-HMM system and GMM-HMM system with various lexicon word on test set.

- Proposed DNN-HMM system significantly outperform GMM-HMM system for all lexicon size.

Future Research

- The other preprocessing steps such as shirorekha detection can be incorporated. • Offline features can also be explored to make the model less sensitive to the temporal changes of the data.

References

[1] S. Bhattacharya, D. S. Maitra, U. Bhattacharya, and Swapan K Parui. An end-to-end system for Bangla online handwriting recognition. In *Proc. of Int. Conf. on Frontiers in Handwriting Recognitio*, pages 373–378. IEEE, 2016.

[2] Subhankar Ghosh, P. K. Bora, Sanjib Das, and B. B. Chaudhuri. Development of an Assamese OCR using Bangla OCR. In *Proc. of the Workshop on Document Analysis and Recognition*, pages 68–73. ACM, 2012.

Contact Information

Subhasis Mandal, Dept. of EEE, Indian Institute of Technology Guwahati, India - 781039
Email subhasis.mandal@iitg.ac.in
Phone +91-9085285110