2012 10th IAPR International Workshop on Document Analysis Systems

Lexicon Reduction Technique for Bangla Handwritten Word Recognition

Tapan Kumar Bhowmik *, Utpal Roy † and Swapan K. Parui ‡
* *Faculty of Mathematics and Natural Sciences*
University of Groningen
Netherlands
Email: tkbhowmik@ai.rug.nl
† *Department of Computer and System Sciences*
Visva Bharati, Santiniketan
India
Email: roy.utpal@gmail.com
‡ *Computer Vision and Pattern Recognition Unit*
Indian Statistical Institute, Kolkata
India
Email: swapan@isical.ac.in

***Abstract*—In this paper we introduce a stroke based lexicon reduction technique in order to reduce the search space for recognition of handwritten words. The principle of this technique involves mainly two aspects of a word image to constitute a feature vector: one is word-length and the other is shape of the word. The length of the word image is represented by the number of specific vertical strokes present in the word image and, on the other hand, the shape of a word image is realized with the combination of both horizontal and vertical strokes. The experiment has been carried out with a database of 35,700 off-line handwritten Bangla word images. Though our proposed lexicon reduction technique is developed for recognition of Bangla handwritten words, its generalization property can easily be exploited for recognition of handwriting in other scripts also.**

***Keywords*-Lexicon reduction; Stroke extraction and representation; Automatic clustering; Off-line handwritten Bangla word recognition;**

I. Introduction

Off-line handwritten word recognition is the transcription of handwritten data into a symbolic (ASCII) electronic format. It has several applications such as reading addresses on postal system [1], reading amounts on bank checks [2], [3], extracting census data on forms, reading address blocks on tax forms etc. There are two major approaches to recognition of a handwritten word image: analytical approach [4], [5] and holistic approach [6]. The idea of analytical approach is to recognize the input word image as a series of segmented sub-images, called characters/glyphs. The holistic approach, on the other hand, considers the word image as a single, indivisible entity, and attempts to recognize the word from its overall shape. Both the approaches have some advantages and disadvantages. The error found in the segmentation based analytical approach originates from both the segmentation technique and the recognition procedure. The errors of these two procedures propagate to the final recognition. But the advantage of the segmentation based recognition is that it can cope with a large lexicon size of handwritten words [7]. On the other hand, for a small lexicon size, the holistic approach produces a high recognition accuracy. Usually in the holistic approach, a hidden markov model (HMM) is used as the recognition engine. In this approach one HMM is constructed for each distinct word in the lexicon. For this it is inefficiently applicable to large lexicon systems due to the growing number of models in respect to the size of lexicon. For example, for a practical problem having a lexicon size 10,000 it is not feasible to compare with 10,000 numbers of HMM during run time against unknown inputs. That makes the problem more complicated as well as time consuming. If we look into a real-life problem, say reading amounts on bank checks or others [8], we see that the lexicon size is not so large. Even if the lexicon size appears larger for some practical problems, we can cope with it by reducing the lexicon size or reorganizing the search space, or using heuristics to limit the search efforts.

In this study we have developed a lexicon reduction technique to reduce the search space during recognition of Bangla handwritten words. The technique is applied on a database of 35,700 off-line handwritten Bangla word images with a lexicon set of 119 words.

II. Lexicon reduction

There are some basic ways to accomplish such lexicon reduction task: knowledge of the application environment, characteristics of the input pattern, and clustering of similar lexicon entries.

A. Lexicon reduction with knowledge of the application environment

Unquestionably, the application environment is the main source of information in limiting the lexicon size. For exam-

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.50

ple, in postal automation system the lexicon size has been reduced by recognizing the ZIP code first. Conventionally, the ZIP code allows the system to reduce the lexicons of thousands entries to a few hundred words [9], [10], [11], [12], [13]. However, when no additional source of knowledge like ZIP code is available, other alternatives may be used to reduce the lexicon. In the case of generic content recognition, where the words are associated to form phrases and sentences, the linguistic knowledge plays an important role in limiting the lexicon [14], [15].

B. Lexicon reduction with characteristics of the input pattern

Other methods attempt to perform a pre-classification of the lexicon entries to evaluate how likely is the matching with the input image. These methods basically look at two aspects: word length and word shape.

The length is a very simple criterion for lexicon reduction. Short words can be easily distinguished from long words by comparing their lengths only. The length of the observation sequence extracted from the input image has intrinsically a hint about the length of the word from which the sequence was extracted. Many lexicon reduction methods make use of such information to reduce the number of lexicon entries to be matched during the recognition process [16], [12], [17], [18], [19], [20], [21]. Kaufmann et al. [17] use a length classifier to eliminate from the lexicon the models that differ significantly from the unknown pattern in the number of symbols. For each model, a minimal and a maximal length are determined. Based on this range, a distance between a word and the model class is defined and used during the recognition process to select only the pertinent models. Kaltenmeier et al. [12] use the word length information given by a statistical classifier adapted to features derived from Fourier descriptors for the outer contours to reduce the number of entries in vocabulary of city names. Kimura et al. [18] estimate the length of the possible word candidates using the segments resulting from the segmentation of the word image. Such estimation provides a confidence interval for the candidate words, and the entries outside of such an interval are eliminated from the lexicon. An underestimation of the interval leads to more classification errors and an over-estimation leads to more computational complexity. Powalka et al. [20] estimate the length of cursive words based on the number of times an imaginary horizontal line drawn through the middle of the word intersects the trace of the pen in its densest area. A similar approach is used by Guillevic et al. [16] to estimate the word length and reduce the lexicon size. The number of characters is estimated using the counts of stroke crossings within the main body of a word.

The shape of the handwritten words is another clue to lexicon reduction. Zimmerman and Mao [22] use key characters in conjunction with word length estimation to limit the size of the lexicon. They attempt to identify some key characters in cursive handwritten words and use them to generate a search string. This search string is matched against the lexicon entries to select the best few matches. A similar approach is proposed by Guillevic et al. [16], but instead of cursive script, they consider only uppercase words. First they attempt to locate isolated characters that are further preprocessed and passed to a character recognizer. The character recognition results are used along with the relative position of the spotted characters to form a grammar. An HMM module is used to implement the grammar and generate some entries that are used to dynamically reduce the lexicon. Kaufmann et al. [17] proposed a method of reducing the size of vocabulary based on the combination of four classifiers: length classifier, profile range, average profile, and transition classifier. All the classifiers use as input the same feature vector used by the recognition system. Madhvanath and Govindaraju [23] present a holistic lexicon filter that takes as input a chain code of a word image and a lexicon and returns a ranked lexicon. First the chain code is corrected for the slant and the skew and features such as natural length, ascenders, and descenders are extracted as well as assertions about the existence of certain features in certain specific parts of the word. The same features are extracted from lexicon entries (ASCII words) by using heuristic rules to combine the expected features of the constituent characters. A graph based framework is used to represent the word image, the lexicon entries and their holistic features and for computing three different distance measures (confidence of match, closeness, and degree of mismatch) between them. These three measures are computed for each lexicon entry and used to rank the hypotheses. A 50% reduction in the size of the lexicon with 1.8% error is reported for a set of 768 lowercase images of city names.

C. Lexicon reduction with clustering of similar lexicon entries

Other approaches [24], [25] attempt to find similarities between lexicon entries and organize them into clusters. So, during the recognition process, the search is carried out only on words that belong to more likely clusters.

III. Proposed Lexicon Reduction Technique

For the purpose of our experiment we have developed a lexicon reduction technique which would be applied for reducing the lexicon size during recognition process. This lexicon reduction technique will determine the number of word HMMs in which the input word image will be compared. For the purpose of the lexicon reduction we have mainly considered two aspects of a word image; one is word-length and the other is shape of the word. The word length of an word image has been represented here with the number of vertical strokes associated with the middle zone [26] of the word image and the number of times an imaginary horizontal line (passing through the middle of the middle zone of the

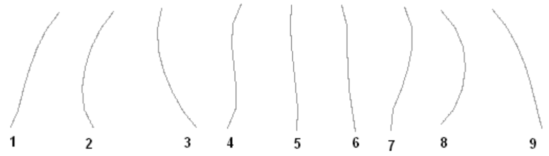


Figure 1. Number of primitive vertical strokes present in the database

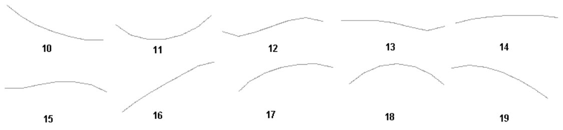


Figure 2. Number of primitive horizontal strokes present in the database

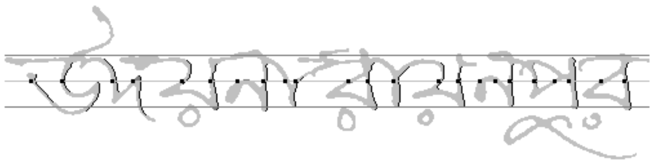
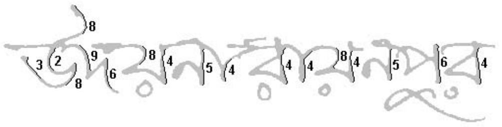


Figure 3. Word image *UDAYNARAYANPUR* indicating vertical strokes and intersecting points by the imaginary line drawn through the middle of the middle zone of the image.

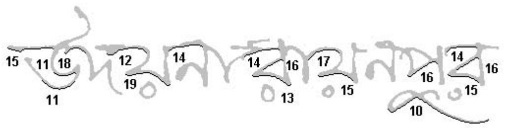
word image) intersects the contour of the word image. For example, in Figure 3, the number of times the imaginary line intersects the contour points is 23 whereas the number of vertical strokes associated with middle zone is 16. Mostly, these two parameters are sufficient to represent the length of the word image. For the shape of the word image we have analyzed the number of vertical and horizontal stokes of the word images. It has been observed that occurrence of vertical and horizontal strokes in a word image depends upon the nature of its overall shape. In other words, the shape of a word image is the combination of several horizontal and vertical strokes. In order to do that the gray level image of a Bangla word is first median filtered and then thresholded into a binary images. Let A be a binary Bangla word image. The aim now is to identify the vertical and horizontal strokes that are present in A. Such a stroke is represented here, as a digital curve which is one-pixel thick and in which all the pixels except two have exactly two 8-neighbours, the other two pixels being the end pixels. In order to get the digital curves representing the vertical and horizontal strokes, two directional view based binary images from A are created. Let E be a binary image consisting of object pixels in A whose right or east neighbour is in the background. In other words, E is formed by the object pixels of A that are visible from the east. Similarly, S is a binary image consisting of object pixels in A whose bottom or south neighbour is in the background [27]. To represent such a digital curve as a feature vector, it has been divided into a certain number of segments and then extracted direction feature (angle) from each segment [27]. The EM-algorithm based Gaussian mixture models and then followed by MMDL [28] model selection criteria are used to find the total number of distinct vertical and horizontal strokes exist in our database. The number of different types of vertical and horizontal strokes are found to be 9 and 10 respectively (see in Figures 1, 2). The vertical strokes are labeled with numbers starting from 1 to 9 and the 10 horizontal strokes are labeled from 10 to 19. So in total we get 19 strokes labeled from 1 to 19 from our database and hence we prepare a global codebook containing the horizontal and vertical strokes with particular label numbers varying from 1 to 19.

From an unknown input image we extract the horizontal and vertical stokes and arrange them from left to right according to their appearance. Each of these strokes has a unique label in the global codebook. Then we label each and every horizontal and vertical stroke present in the sequence. So for any unknown input word image, we get a sequence of labels based on the appearance and nature of the strokes. Though both the number of different strokes and the sequence in which they occur, are important characteristics to distinguish one type of shape from another, for simplicity and to reduce the size of the feature vector, we have considered only the frequency of the occurrence of different strokes. Since the number of different strokes is 19, the size of the feature vector for a word image is 19 in which a feature represents the frequency of the corresponding stroke appearing in the word image. Two more features are added to the feature vector to achieve a better distinguishing capability. Thus the feature vector for a word image a length of 21(19+2). For example, the feature vector $\langle 0, 1, 1, 6, 2, 2, 0, 4, 1, 1, 2, 1, 1, 3, 3, 3, 1, 1, 1, 23, 16\rangle$ is extracted from the word image in Figure 4. Here the first component being zero means no vertical stroke with label number 1, is found in the word image in Figure 5. Similarly, the value of the 10th component being 1 indicates that there exists only one horizontal stroke with label number 10. The last two components of the feature vector are for the number of times an imaginary horizontal line drawn through the middle of middle zone of the word image intersects the trace of the pen in its densest area and the number of vertical strokes associated to the middle zone of the word image respectively (in Figure 3).



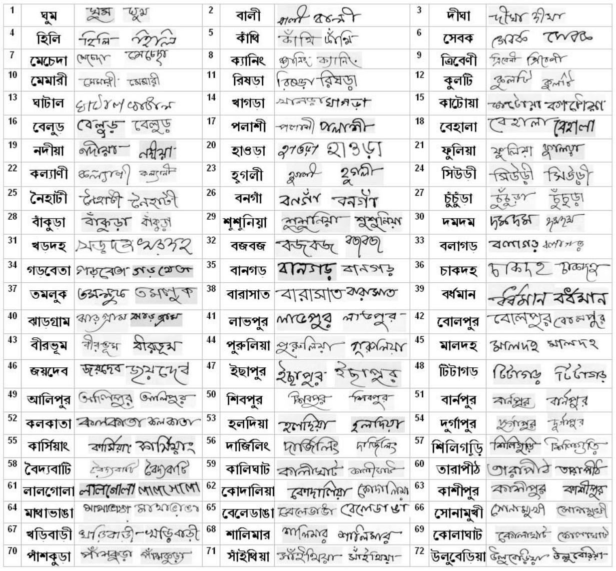
(a)

Figure 4. Vertical strokes along with label numbers.



(b)

Figure 5. Horizontal strokes along with label numbers.



1	ঘুম	2	বালী	3	দীঘা
4	হিলি	5	কাঁথি	6	সেবক
7	মেচেদা	8	ক্যানিং	9	ত্রিবেণী
10	মেমারী	11	রিষড়া	12	কুলটি
13	ঘাটাল	14	খাগড়া	15	কাটোয়া
16	বেলুড়	17	পলাশী	18	বেহালা
19	নদীয়া	20	হাওড়া	21	ফুলিয়া
22	কল্যাণী	23	হুগলী	24	সিউড়ী
25	নৈহাটী	26	বনগাঁ	27	চুঁচুড়া
28	বাঁকুড়া	29	শুশুনিয়া	30	দমদম
31	খড়দহ	32	বজবজ	33	বলাগড়
34	গড়বেতা	35	বানগড়	36	চাকদহ
37	তমলুক	38	বারাসাত	39	বর্ধমান
40	ঝাড়গ্রাম	41	লাভপুর	42	বোলপুর
43	বীরভূম	44	পুরুলিয়া	45	মালদহ
46	জয়দেব	47	ইছাপুর	48	টিটাগড়
49	আলিপুর	50	শিবপুর	51	বার্নপুর
52	কলকাতা	53	হলদিয়া	54	দুর্গাপুর
55	কার্সিয়াং	56	দার্জিলিং	57	শিলিগুড়ি
58	বৈদ্যবাটি	59	কালিঘাট	60	তারাপীঠ
61	লালগোলা	62	কোদালিয়া	63	কাশীপুর
64	মাথাভাঙা	65	বেলেডাঙা	66	সোনামুখী
67	খড়িবাড়ী	68	শালিমার	69	কোলাঘাট
70	পাঁশকুড়া	71	সাঁইথিয়া	72	উলুবেড়িয়া

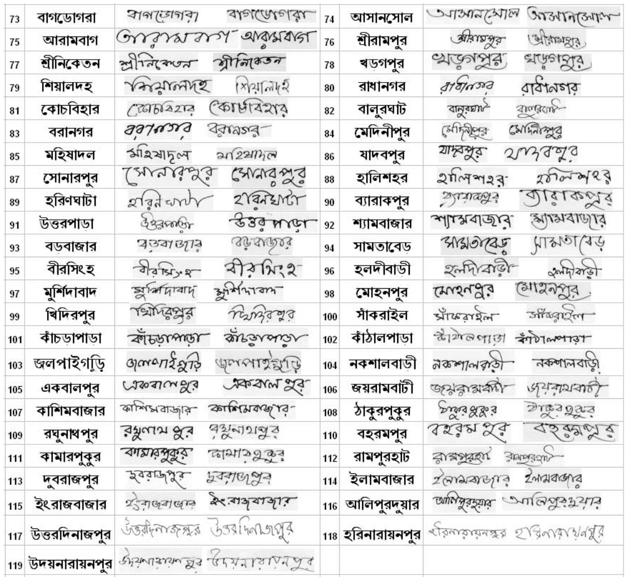
Figure 6. Lexicon words with reference numbers and two handwritten samples (1-72)

IV. Experimental results

For the present experiment, 35700 (300 × 119) feature vectors have been extracted from 119 classes of word images with each containing 300 different writing samples. All the words in the lexicon along with their reference numbers and two handwritten samples in each lexicon are shown in Figures 6 and 7. Out of them, 100 writing samples of each class i.e, 11900 features have been clustered with K-means algorithm and rest of them (200 × 119) have been used to identify which clusters contain which word classes. It has been observed from the clusters that words having almost the same length and a similar kind of shape have been grouped into a cluster. The number of clusters has been determined by trial and error. In this experiment eight such clusters and their corresponding word classes (with error 3.8%) are shown in Table I. It is clear that the clusters are overlapped indicating that the handwritten form of one single word may have different lengths and/or different shapes.

V. Conclusion

In this article, we have described a lexicon reduction technique for reducing the lexicon size during recognition process. We have analyzed the vertical and horizontal strokes of the word images. The number of strokes indicates the length of the word image and at the same time the strokes represent the overall shape of the word images. Though our proposed lexicon reduction technique is applied for recognition of Bangla handwritten words, its robustness properties can easily be exploited for the recognition of handwriting in other scripts also.



73	বাগডোগরা	74	আসানসোল
75	আরামবাগ	76	শ্রীরামপুর
77	শ্রীনিকেতন	78	খড়গপুর
79	শিয়ালদহ	80	রাধানগর
81	কোচবিহার	82	বালুরঘাট
83	বরানগর	84	মেদিনীপুর
85	মহিষাদল	86	যাদবপুর
87	সোনারপুর	88	হালিশহর
89	হরিণঘাটা	90	ব্যারাকপুর
91	উত্তরপাড়া	92	শ্যামবাজার
93	বড়বাজার	94	সামতাবেড়
95	বীরসিংহ	96	হলদীবাড়ী
97	মুর্শিদাবাদ	98	মোহনপুর
99	খিদিরপুর	100	সাঁকরাইল
101	কাঁচড়াপাড়া	102	কাঁঠালপাড়া
103	জলপাইগুড়ি	104	নকশালবাড়ী
105	একবালপুর	106	জয়রামবাটী
107	কাশিমবাজার	108	ঠাকুরপুকুর
109	রঘুনাথপুর	110	বহরমপুর
111	কামারপুকুর	112	রামপুরহাট
113	দুবরাজপুর	114	ইলামবাজার
115	ইংরাজবাজার	116	আলিপুরদুয়ার
117	উত্তরদিনাজপুর	118	হরিনারায়নপুর
119	উদয়নারায়নপুর		

Figure 7. Lexicon words with reference numbers and two handwritten samples (73-119)

Table I
Clusters and its associated word classes

Clusters	Clusters with word references
1	{1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 16, 19, 21, 22, 23, 24, 26, 27, 28, 29, 30, 32, 37, 43, 47, 50, 51, 54}
2	{9, 10, 15, 17, 21, 22, 32, 34, 37, 41, 42, 43, 46, 49, 50, 53, 58, 63, 70, 78, 82, 83, 86, 89}
3	{29, 60, 64, 68, 76, 81, 84, 89, 96, 109}
4	{7, 14, 15, 33, 38, 40, 42, 48, 52, 55, 56, 57, 58, 59, 60, 61, 65, 67, 68, 69, 71, 78, 79, 80, 81, 83, 84, 85, 87, 89, 90, 91, 93, 95, 96, 97, 98, 99, 113}
5	{74, 77, 92, 94, 101, 102, 104, 106, 107, 111, 114, 115, 116, 117, 118, 119}
6	{4, 5, 6, 7, 8, 9, 10, 13, 14, 16, 18, 19, 20, 21, 22, 24, 25, 28, 29, 30, 31, 32, 33, 35, 36, 37, 39, 45, 46, 48, 53}
7	{44, 49, 57, 59, 62, 66, 69, 70, 72, 76, 79, 82, 84, 88, 89, 91, 96, 97, 98, 99, 100, 103, 105, 108, 109, 110, 111, 112, 113, 114}
8	{52, 61, 64, 73, 74, 75, 94, 101, 102}

References

[1] M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, “Unconstrained farsi (arabic) handwritten word recognition using fuzzy vector quantization and hidden markov models,” *Pattern Recognition Letters*, vol. 22, pp. 209–214, 2001.

[2] D. Guillevic and C. Suen, “Hmm word recognition engine,” in *Proceedings of the* 4^{th} *International Conference on Document Analysis and Recognition*. Ulm, Germany, 1997, pp. 544–547.

[3] T. Paquet and Y. Lecourtie, "Recognition of handwitten sentences using restricted lexicon," *Pattern Recognition*, vol. 26, no. 3, pp. 391–407, 1993.

[4] V. Govindaraju and G. Kim, "A lexicon driven approach to handwritten word recognition for real-time applications," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 19, no. 4, pp. 366 –379, 1997.

[5] C. Olivier, T. Paquet, M. Avila, and Y. Lecourtier, "Recognition of handwitten word using stochastic models," in *Proceedings of the International Conference on Document Analysis and Recognition*, 1995, pp. 19–23.

[6] S. Madhvanath and V. Govindaraju, "The role of holistic paradigms in handwritten word recognition," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 23, no. 2, pp. 149–164, 2001.

[7] L. Heutte, A. Nosary, and T. Paquet, "A multiple agent architecture for handwritten text recognition," *Pattern Recognition*, vol. 37, no. 4, pp. 665–674, 2004.

[8] C. Olivier, T. Paquet, M. Avila, and Y. Lecourtier, "Optimal order of marcov models applied to bankchecks," *Int. Jour. of Pattern Recognition and Artificial Intelligence*, vol. 11, no. 5, pp. 789–800, 1997.

[9] A. El-Yacoubi, M. Gilloux, R. Sabourin, and C. Y. Suen, "Unconstrained handwritten word recognition using hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 752–760, 1999.

[10] C. Farouz, *Reconnaissance de Mots Manuscrits Hors-Ligne dans un Vocabulaire Ouvert par Modelisation Markovienne.* PhD thesis, Universite de Nantes, Nantes, France, August 1999.

[11] P. D. Gader, M. A. Mohamed, and J. H. Chiang, "Handwritten word recognition with character and inter-character neural networks," *IEEE Transactions on Systems, Man and Cybernetics - Part B*, vol. 27, pp. 158–164, 1994.

[12] A. Kaltenmeier, T. Caesar, J. M. Gloger, and E. Mandler, "Sophisticated topology of hidden markov models for cursive script recognition," in *Proceedings of the International Conference on Document Analysis and Recognition*. Tsukuba, Japan, 1993, pp. 139–142.

[13] C. K. Lee and G. Leedham, "Rapid analytical verification of handwritten alphanumeric address fields," in *Proceedings of the 7^{th} International Workshop on Frontiers in Handwriting Recognition*. Amsterdam, Netherlands, 2000, pp. 571–576.

[14] A. Brakensiek, J. Rottland, A. Kosmala, and G. Rigoll, "Offline handwriting recognition using various hybrid modeling techniques and character n-grams," in *Proceedings of the 7^{th} International Workshop on Frontiers in Handwriting Recognition*. Amsterdam, Netherlands, 2000, pp. 343–352.

[15] U. Marti and H. Bunke, "Towards general cursive script recognition," in *Proceedings of the 6^{th} International Workshop on Frontiers in Handwriting Recognition*. Taejon, Korea, 1998, pp. 379–388.

[16] D. Guillevic, D. Nishiwaki, and K. Yamada, "Word lexicon reduction by character spotting," in *Proceedings of the 7^{th} International Workshop on Frontiers in Handwriting Recognition*. Amsterdam, Netherlands, 2000, pp. 373–382.

[17] G. Kaufmann, H. Bunke, and M. Hadorn, "Lexicon reduction in an hmm-framework based on quantized feature vectors," in *Proceedings of the 4^{th} International Conference on Document Analysis and Recognition*. Ulm, Germany, 1997, pp. 1097–1101.

[18] F. Kimura, M. Shridhar, and Z. Chen, "Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words," in *Proceedings of the International Conference on Document Analysis and Recognition*. Tsukuba, Japan, 1993, pp. 18–22.

[19] A. L. Koerich, R. Sabourin, and C. Y. Suen, "A time-length constrained level building algorithm for large vocabulary handwritten word recognition," in *Proceedings of the 2^{nd} International Conference on Advances in Pattern Recognition*. Rio de Janeiro, Brazil, 2001, pp. 127–136.

[20] R. K. Powalka, N. Sherkat, and R. J. Whitrow, "Word shape analysis for a hybrid recognition system," *Pattern Recognition*, vol. 30, no. 3, pp. 412–445, 1997.

[21] R. Bertlolami, C. Gutmann, H. Bunke, and A. L. Spitz, "Shape code based lexicon reduction for offline handwritten word recognition," in *Proceedings of the 8^{th} IAPR International Workshop on Document Analysis Systems*, 2008, pp. 151–157.

[22] M. Zimmermann and J. Mao, "Lexicon reduction using key characters in cursive handwritten words," *Pattern Recognition Letters*, vol. 20, pp. 1297–1304, 1999.

[23] S. Madhvanath and V. Govindaraju, "Holistic lexicon reduction," in *Proceedings of the 3^{rd} International Workshop on Frontiers in Handwriting Recognition*. Buffalo, USA, 1993, pp. 71–78.

[24] M. Gilloux, "Real-time handwritten word recognition within large lexicons," in *Proceedings of the 5^{th} International Workshop on Frontiers in Handwriting Recognition*. Essex, UK, 1996, pp. 301–304.

[25] Z. Wimmer, B. Dorizzi, and P. Gallinari, "Dictionary preselection in a neuro-markovian word recognition system," in *Proceedings of the 5^{th} International Conference on Document Analysis and Recognition*. Bangalore, India, 1999, pp. 539–542.

[26] T. K. Bhowmik, S. K. Parui, and U. Roy, "Discriminative hmm training with ga for handwritten word recognition," in *Proceedings of the 19^{th} International Conference on Pattern Recognition (ICPR)*. IEEE, Brazil, 2008, pp. 1–4.

[27] T. K. Bhowmik, U. Bhattacharya, and S. K. Parui, "Recognition of bangla handwritten characters using an mlp classifier based on stroke features," in *Proceedings of the 11^{th} International Conference on Neural Information Processing (ICONIP)*. Springer-Verlag, 2004, pp. 814–819.

[28] M. A. T. Figueiredo, J. M. N. Leitao, and A. K. Jain, "On fitting mixture models," in *EMMCVPR*. Springer-Verlag, 1999, pp. 54–69.