An Aid to Assamese Language Processing by Constructing an Offline Assamese Handwritten Dataset

Debabrata Khargharia and **Samir Kumar Borgohain**
Department of Computer Science & Engineering
National Institute of Technology Silchar, Assam, India, 788010
debabrata21_rs@cse.nits.ac.in, samir@cse.nits.ac.in

Abstract

Recent years have seen a growing interest in analyzing Indian handwritten documents. In pattern recognition, particularly handwritten document recognition, the availability of standard databases is essential for assessing algorithm efficacy and facilitating result comparisons among research groups. However, there is a notable scarcity of standardized databases for handwritten texts in Indian languages. This paper presents a comprehensive methodology for the development of a novel, unconstrained dataset named OAHTD (Offline Assamese Handwritten Text Dataset) for the Assamese language, derived from offline handwritten documents. The dataset, which represents a significant contribution to the field of Optical Character Recognition (OCR) for handwritten Assamese, is the first of its kind in this domain. The corpus comprises 410 document images, each containing a diverse array of linguistic elements including words, numerals, individual characters, and various symbols. These documents were collected from a demographically diverse cohort of 300 contributors, spanning an age range of 10 to 76 years and representing varied educational backgrounds and genders. This meticulously curated collection aims to provide a robust foundation for developing and evaluating OCR algorithms specifically tailored to the Assamese script, addressing a critical gap in the existing literature and resources for this language.

1 Introduction

The recognition of language and characters has gained importance as a crucial field of research across the globe. The study of Optical Character Recognition (OCR) systems is essential to enable automated visual perception and reading. OCR involves a document image analysis process where machine-printed or handwritten documents are subjected to a learning process to convert them into machine-readable and editable format (Smith, 2007). The study of OCR revolves around the different modes of acquisition and modes of writing of the document/texts. Based on the mode of writing, OCR is categorized into *Handwritten and Printed*, while based on the mode of acquisition, it is categorized in *Online and Offline*. Thus, OCR recognizes both handwritten and printed text that is obtained online as well as offline mode (Plamondon and Srihari, 2000). In the online recognition mode, the temporal information (position of the pen, trajectory, etc.) is considered. It requires direct interaction with the user as its handwriting is considered. Offline recognition refers to optically capturing the handwritten text as an image and further analyzing and processing it for recognition. It does not require direct interaction with the user whose handwriting is taken under consideration (Figure 1).

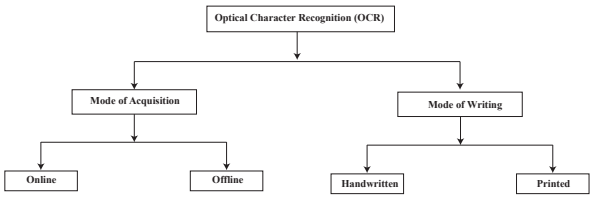

Figure 1: Different Types of Optical Character Recognition

The initial stage in OCR is scanning the documents if they are in a physical form. Preprocessing such as cropping, noise removal, skew removal, binarization, etc., are performed as required. The next important step is the segmentation process, which incorporates segmenting or extracting or isolating the text lines in the document images (handwritten or printed). Segmentation of handwritten documents is a challenging task compared to the printed. The next phase in segmentation is the extraction or isolation of the constituent words and

characters from the extracted text lines. The final stage incorporates a classification model that is able to classify and recognize the final characters obtained after segmentation. A general pipeline of this process is shown in Figure 2.

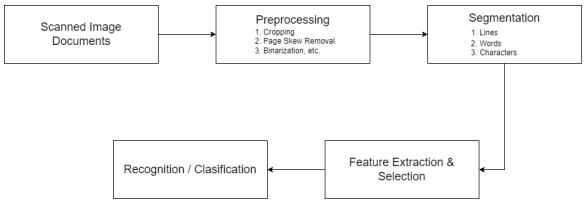

Figure 2: Various stages in character recognition from handwritten documents

There are many challenges associated with the segmentation or recognition of characters in handwritten OCR. Some of them are:

1. Variation in the writing style across different individuals.
2. Variation in writing style is also seen for the same writer at different times depending on the mode, and other situations, etc.
3. Presence of slants or skews in the lines while writing.
4. Inadequate spacing of characters leading to the difficulty in distinguishing the starting and ending points of the characters.
5. The quality of the image obtained for processing also affects the recognition process in OCR.
6. Availability of the dataset in the required form and language.

Among the various challenges witnessed in literature for the segmentation or recognition of languages and characters (Memon et al., 2020), the availability of the dataset in the desired form and language is a challenging one. Literature has witnessed a surge in OCR technologies for various languages such as Arabic, Chinese, Roman, etc., but OCR techniques for regional languages such as Assamese are comparatively less. The structure of various languages across the globe is different from one another. Thus techniques developed for the segmentation and recognition of one language are different from the other. Due to this, OCR techniques for different languages will also be different. The main motivation of the present work delves significantly into the development of an offline dataset for the Assamese language, as there is no offline public database of Assamese script. Assamese is one of the most dominant languages of North-East India. So creating a standardised database will greatly help the people of the region for improved accessibility.

2 Background

2.1 The Assamese Script

Assamese is the easternmost Indo-Aryan language spoken in the Assam Valley districts, with Lakhimpur in the extreme east and Goalpara in the extreme west (Kakati, 1941). Based on the census report of 2011, Assamese is spoken by 1.26% of the total Indian population (of India, 2011). It is also spoken in some parts of Arunachal Pradesh and Nagaland. Though the language originated in the 7^{th} century, its literature came only in the early 14^{th} century (Kakati, 1941). But history says that Assamese literature existed even before the 14^{th} century, for there is evidence of a rich heritage of oral traditions, including folk songs, religious hymns, pastoral ballads, festival songs, and even children's stories.

2.2 Characteristics of Assamese script

- Unlike the English language, the Assamese language does not have the Upper and Lower Case letters or characters.
- It comprises of 11 vowels and 41 consonants which are known as *Sworobornomala* and *Byonjonbornomala*, respectively.
- The vowels and consonants are called as the **Basic Characters**.
- In the Assamese language, a combination of the basic characters and a combination of various consonants form a different structure and utility. They are known as the **Compound and Composite characters.**
- Over 300 compound characters are present which are also known as *Juktakhor*.
- The language has 10 unique digits known as *Sankyha*.
- It also comprises different Zones and Lines as shown in Figure 3.

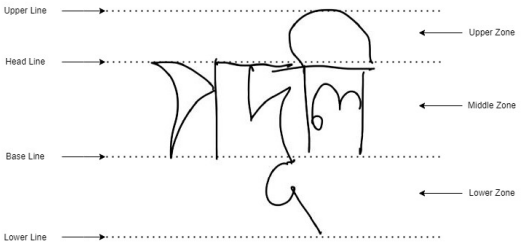

Figure 3: Different zones an Assamese word may have

Figure 4 shows this language's basic characters and digits.

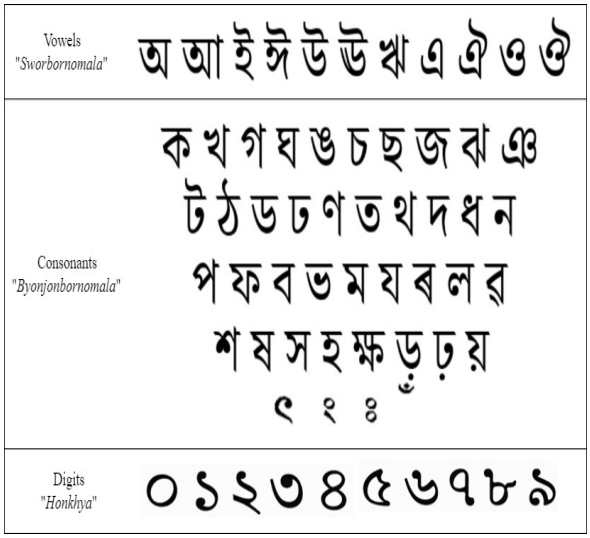

Figure 4: Assamese Character Set

3 Related work

This section discusses the related work for different handwritten datasets of Indic Languages. The dataset selection carries significant weight due to the inherent diversity in language patterns, resulting in variations in learning outcomes. This variability is especially evident in tasks such as Optical Character Recognition (OCR), where different datasets have been extensively examined and experimented with. For instance, languages like Assamese and Hindi incorporate diacritical marks (matras), which affect character recognition differently than languages like Telugu and English, which lack such components. Consequently, researchers have explored a range of datasets to understand these variations and optimize OCR systems for various linguistic contexts. Good datasets are essential for understanding and interpreting handwriting in any language. Table 1 reviews some popular datasets used in studies over the last ten years.

The field of handwritten document analysis for Indian languages has witnessed significant advancements in recent years, largely facilitated by the development of diverse and comprehensive datasets. These datasets have played a crucial role in enabling researchers to develop, evaluate, and benchmark various algorithms and methodologies for optical character recognition (OCR) and related tasks. A chronological examination of the literature reveals a progression in both the scope and sophistication of these datasets. In 2008, (Bhattacharya and Chaudhuri, 2008) made a notable contribution with the ISI-HDND dataset, which comprised 22,556 Devanagari and 23,392 Bangla numerals, addressing the need for numeral recognition in two major Indian scripts. The following year, (de Campos et al., 2009) introduced the Chars 74k dataset, a substantial collection of 74,000 English and Kannada characters, which proved invaluable for cross-lingual character recognition studies. (Baruah and Hazarika, 2011) focused on online handwriting recognition with their dataset of 8,235 Assamese characters, while (Alaei et al., 2011) contributed significantly to Kannada script research with the KHTD dataset, encompassing 4,298 text lines and 26,115 words. This period also saw (Sarkar et al., 2012) introduce the CMATERdb1.1, a collection of 100 pages of Bengali text lines, further enriching the resources for Bengali script analysis. (Acharya et al., 2015) made a substantial contribution with the DHCD dataset, featuring 92,000 Devanagari characters, which has since become a benchmark for Devanagari character recognition tasks. (Singh et al., 2018) expanded the CMATER database with two new additions: CMATERdb2.2.3, containing 15,528 Devanagari words, and CMATERdb2.1.3, comprising 18,931 Bengali words, both of which have proven instrumental in word-level recognition studies. A significant leap towards multi-script analysis came with (Obaidullah et al., 2018)'s PHDIndic_11 dataset, which included 1,458 pages of text lines from 11 different scripts, including Bangla, Oriya, Roman, and Urdu. This dataset has been particularly valuable for developing script-independent recognition techniques and cross-script studies. More recent contributions have focused on addressing gaps in specific language resources. (Dutta and Muppalaneni, 2021) developed a dataset of 512 Assamese handwritten digits, addressing the scarcity of resources for this less-studied language. Similarly,

Table 1: Different handwritten datasets in OCR for Indic Scripts

Paper	Author	Year	Dataset	Language / Script	Type	Dataset Size
(de Campos et al., 2009)	De *et al.*	2009	Chars 74k	English and Kannada	Characters	74,000
(Bhattacharya and Chaudhuri, 2008)	Bhattacharya *et al.*	2008	ISI-HDND	Devanagari and Bangla	Numerals	22,556 and 23,392
(Baruah and Hazarika, 2011)	Baruah *et al.*	2011	Online Handwritten Assamese Characters	Assamese	Characters (Online)	8,235
(Alaei et al., 2011)	Alaei *et al.*	2012	KHTD	Kannada	Text Lines and Words	4,298 and 26,115
(Sarkar et al., 2012)	Sarkar *et al.*	2012	CMATERdb1.1	Bengali	Text Lines	100 pages
(Acharya et al., 2015)	Acharya *et al.*	2015	DHCD	Devanagari	Characters	92,000
(Singh et al., 2018)	Singh *et al.*	2017	CMATERdb2.2.3	Devanagari	Words	15,528
(Singh et al., 2018)	Singh *et al.*	2017	CMATERdb2.1.3	Bengali	Words	18,931
(Obaidullah et al., 2018)	Obaidullah *et al.*	2018	PHDIndic_11	11 scripts such as Bangla, Oriya, Roman, Urdu,etc.	Text Lines	1,458 Pages
(Dutta and Muppalaneni, 2021)	Dutta *et al.*	2021	Assamese Handwritten Digits	Assamese	Digits	512
(Goel and Ganatra, 2023)	Goel *et al.*	2023	Gujarati Dataset	Gujarati	Digits	8,000

(Goel and Ganatra, 2023) contributed to Gujarati script research with their dataset of 8,000 handwritten digits. These datasets have significantly contributed to the advancement of handwritten document analysis for various Indian languages and scripts. The progression from character-level to word-level and eventually to multi-script datasets reflects the evolving challenges and ambitions in the field of handwritten document analysis for Indian languages.

It is apparent from Table 1 that, for the Assamese Language, there are only two public datasets available, which are an Online Handwritten Characters Dataset and an Offline Handwritten Digits Dataset. The datasets used in (Choudhury and Sarma, 2021; Choudhury et al., 2015; Chourasia et al., 2019; Bania and Khan, 2018; Yadav et al., 2022; Borgohain et al., 2023; Singh et al., 2021) are the author's customized dataset of characters and digits, which was taken either online or offline and performed different classification algorithms for the character and digit recognition of the Assamese language and they have not made it public.

4 Dataset Description

Due to the lack of an existing Assamese dataset suitable for our specific task, we have undertaken the initiative to create our own custom dataset. The novel dataset developed for this research has been designated as **Offline Assamese Handwritten Text Dataset (OAHTD)**. This nomenclature reflects the nature of the data collected, emphasizing its focus on offline handwritten text in the Assamese language. The process of data collec-

tion and compilation for the **OAHTD** followed a structured workflow, which is visually represented in Figure 5. This workflow diagram delineates the systematic approach employed in the dataset's creation, encompassing the various stages from initial data gathering to final dataset compilation.

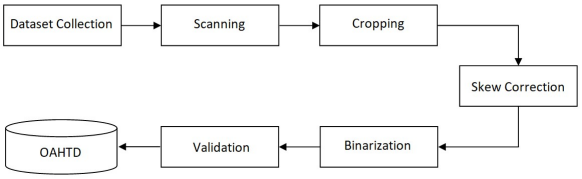

Figure 5: Workflow of the OAHTD creation

4.1 Dataset collection

Even though we have a lot of technology, gathering good data can take time and effort. Some people may not want to share information, and there are worries about privacy. Also, the natural world constantly changes, making it harder to collect data smoothly. These factors make the process of data collection even more challenging. The whole process of our data collection took one year. While collecting a dataset, most researchers assume that the handwritings are flawlessly written, i.e., in a constrained way. In our process, handwritten notes were collected from the Assamese essay writings of school students and other volunteers upon request. We gave the school students a rule page to write an essay on some topics, where each page consists of 17 ruled lines. The remaining volunteers were assigned to write freely on any topic they wanted and any paper they were comfortable with. The remaining volunteers were not given any restrictions regarding the type of page they used or the style of writing they chose. Figure 6 shows the various statistics of the dataset we collected.

Table 2: Statistics of OAHTD

No. of Samples Collected	410
Average no. of Text Lines (per Sample)	12 Lines
Average no. of Words (per Line)	7 words
Total no. of Text Lines	~4,920
Total no. of Words	~34,440
Total no. of Words	~43,680
Format of the scanned images	.jpg

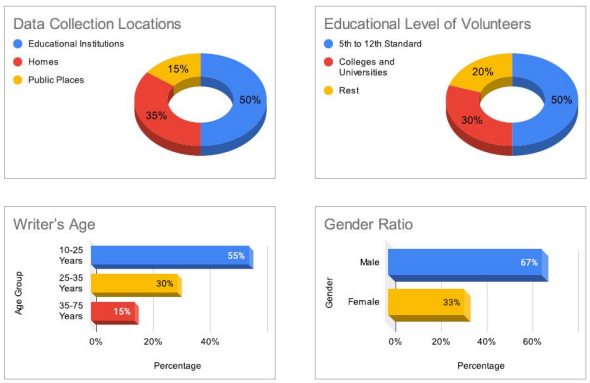

Figure 6: Volunteer Information for Dataset Collection

4.2 Scanning

All documents were scanned using an HP Scanjet 8270 flatbed scanner at a resolution of 300 DPI in RGB color mode. To minimize page skew, careful attention was given during the document feeding process. The scanned images were saved in .jpg format, following a naming convention of A***.jpg, where *** represents a unique integer assigned sequentially to each image. A sample scanned image is shown in Figure 7.

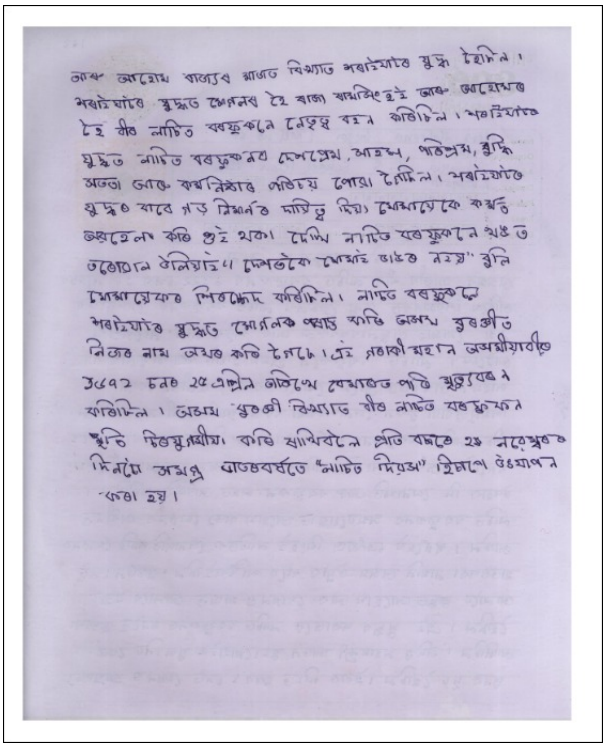

Figure 7: Sample of a scanned document image

4.3 Cropping and Resizing

The images were manually cropped to isolate the desired region of interest (ROI). Following this,

the cropped images were resized with the output dimensions set to *output_size = (800,800)* to ensure uniformity across all images.

4.4 Skew Correction and Binarization

Skew correction of the images was performed using the straight-line fitting method (Cao et al., 2003). The corrected images were then binarized using Gaussian Blur and Adaptive Thresholding with the following parameters: *max_value=255, block_size=11, C=2, and blur_ksize=(21, 21)*. Subsequently, high pass filter was applied to eliminate noise caused by factors such as ink smudging and dust on the scanner bed. A sample binarized image is shown in Figure 8.

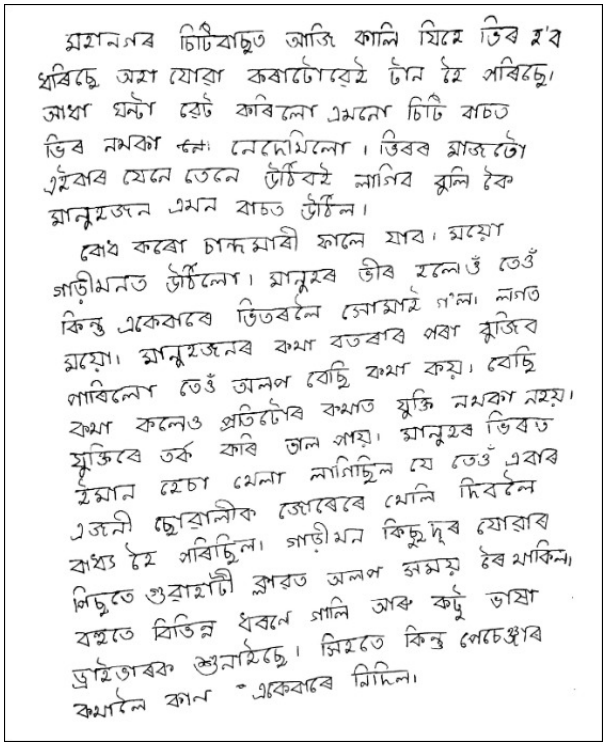

Figure 8: Image after Skew-Correction, Binarization and Noise-removal

4.5 Dataset validation methods

One of the most important parts of creating a dataset is the evaluation of the data by subject experts. For this work, we reached out to different subject experts who are faculties at respectable universities in Assam. They checked these documents based on the following criteria :

- Grammar accuracy: The evaluators assessed the correctness of grammar in the documents, paying attention to sentence structure, punctuation, and adherence to syntactic rules. This ensures that the text is free from grammatical errors that could affect its clarity.
- Subject and predicate structure: This involves analyzing the fundamental sentence construction to ensure the proper relationship between the subject and the predicate, which is crucial for syntactic correctness and logical flow in the writing.
- Verb usage: The experts evaluated whether verbs were used appropriately, focusing on verb tense, agreement with subjects, and consistency throughout the text. Correct verb usage contributes to clear and effective communication.
- Readability: The readability of the text was judged based on how easily a reader could understand it. The experts classified the writing into three categories—well-understood, average, or poor—based on factors such as sentence complexity, vocabulary, and overall flow of ideas.

Table 3: Detailed description of some document images taken from the dataset OAHTD

Document ID	Height	Width	Aspect Ratio	No. of Lines	No. of Words
001	800	800	1	10	125
002	800	800	1	10	120
003	800	800	1	9	110
004	800	800	1	9	110
010	800	800	1	16	160
030	800	800	1	15	175
045	800	800	1	14	170
100	800	800	1	17	180
102	800	800	1	17	184
104	800	800	1	15	160
154	800	800	1	16	165
290	800	800	1	17	188
292	800	800	1	17	198
294	800	800	1	17	212

5 Conclusion and Future Scope

In this paper, we have discussed the different steps in creating an unconstrained dataset for the Assamese Language from offline handwritten documents containing Assamese script pages. This dataset is the first of its kind in this domain, i.e. in the field of OCR for the handwritten Assamese language. Each document contains words, digits, characters and other symbols written by the writers. In this current dataset, 410 document images were

collected from 300 writers of different age groups ranging from the age 10 to 76 years, with different educational backgrounds, and genders.

Creating this dataset is anticipated to aid in developing effective models for Assamese OCR. It will also help the research community to evaluate the state-of-the-art algorithms and instigate more research work in the field of Assamese handwriting character recognition.

Future work will include the development of an additional variant of this dataset, to be designated as the **Offline Assamese Handwritten Word Dataset (OAHWD)**. While the current Offline Assamese Handwritten Text Dataset (OAHTD) comprises a collection of offline handwritten Assamese text images, the proposed OAHWD will focus specifically on offline handwritten Assamese word images. This expansion aims to provide researchers with more granular data for word-level analysis and recognition tasks. In the interest of advancing research in Assamese handwriting recognition and related fields, both the OAHTD and the forthcoming OAHWD are planned to be made publicly accessible to the research community.

Acknowledgements

Many people helped us complete this dataset. We are grateful to all the Assamese students and volunteers who contributed their handwriting samples to make this project successful. We are also thankful to Mrs. Snigdha Rani Bordoloi, who helped us collect the samples from the school students and also in the data annotation. Also, a heartfelt thanks to Sri Sumanta Chaliha, Vice-Chairman, Publication Board Assam and also to Mrs Nibedita Changkakoty, Associate Professor and HOD in the Department of Assamese, Dibrugarh University who helped us in the validation of the Assamese Text.

References

Shailesh Acharya, Ashok Kumar Pant, and Prashnna Kumar Gyawali. 2015. Deep learning based large scale handwritten devanagari character recognition. In *2015 9th International conference on software, knowledge, information management and applications (SKIMA)*, pages 1–6. IEEE.

Alireza Alaei, P Nagabhushan, and Umapada Pal. 2011. A benchmark kannada handwritten document dataset and its segmentation. In *2011 International Conference on Document Analysis and Recognition*, pages 141–145. IEEE.

Rubul Kumar Bania and R Khan. 2018. Handwritten assamese character recognition using texture and diagonal orientation features with artificial neural network. *Int J Appl Eng Res*, 13(10):7797–7805.

Udayan Baruah and Shyamanta Hazarika. 2011. Online Handwritten Assamese Characters Dataset. UCI Machine Learning Repository.

Ujjwal Bhattacharya and Bidyut Baran Chaudhuri. 2008. Handwritten numeral databases of indian scripts and multistage recognition of mixed numerals. *IEEE transactions on pattern analysis and machine intelligence*, 31(3):444–457.

Olimpia Borgohain, Pramod Kumar, and Saurabh Sutradhar. 2023. Recognition of handwritten assamese characters. In *Proceedings of 3rd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2022*, pages 223–230. Springer.

Yang Cao, Shuhua Wang, and Heng Li. 2003. Skew detection and correction in document images based on straight-line fitting. *Pattern Recognition Letters*, 24(12):1871–1879.

Ananya Choudhury and Kandarpa Kumar Sarma. 2021. A cnn-lstm based ensemble framework for in-air handwritten assamese character recognition. *Multimedia Tools and Applications*, pages 1–36.

Himakshi Choudhury, Subhasis Mandal, Sanjeevan Devnath, SR Mahadeva Prasanna, and Suresh Sundaram. 2015. Combining hmm and svm based stroke classifiers for online assamese handwritten character recognition. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE.

Chandan Kumar Chourasia, Manashjyoti Barman, et al. 2019. Handwritten assamese character recognition. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE.

Teófilo E de Campos, Bodla Rakesh Babu, and Manik Varma. 2009. Character recognition in natural images. In *International conference on computer vision theory and applications*, volume 1, pages 273–280. SCITEPRESS.

Prarthana Dutta and Naresh Babu Muppalaneni. 2021. Diginet: Prediction of assamese handwritten digits using convolutional neural network. *Concurrency and computation: practice and experience*, 33(24):e6451.

Parth Goel and Amit Ganatra. 2023. Handwritten gujarati numerals classification based on deep convolution neural networks using transfer learning scenarios. *IEEE Access*, 11:20202–20215.

Banikanta Kakati. 1941. ASSAMESE,ITS FORMATION AND DEVELOPMENT. Government of Assam.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE access*, 8:142642–142668.

Sk Md Obaidullah, KC Santosh, Nibaran Das, and Kaushik Roy. 2018. Phdindic_11: page-level handwritten document image dataset of 11 official indic scripts for script identification. *Multimedia Tools and Applications*, 77:1643–1678.

Census of India. 2011. Language: India, states and union territories.

Réjean Plamondon and Sargur N Srihari. 2000. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):63–84.

Ram Sarkar, Nibaran Das, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri, and Dipak Kumar Basu. 2012. Cmaterdb1: a database of unconstrained handwritten bangla and bangla–english mixed script document image. *International Journal on Document Analysis and Recognition (IJDAR)*, 15:71–83.

Jaisal Singh, Srinivasan Natesan, Marcin Paprzycki, and Maria Ganzha. 2021. Experimenting with assamese handwritten character recognition. In *International Conference on Big Data Analytics*, pages 219–229. Springer.

Pawan Kumar Singh, Ram Sarkar, Nibaran Das, Subhadip Basu, Mahantapas Kundu, and Mita Nasipuri. 2018. Benchmark databases of handwritten bangla-roman and devanagari-roman mixed-script document images. *Multimedia Tools and Applications*, 77:8441–8473.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Mihir Yadav, Divyansh Mangal, Srinivasan Natesan, Marcin Paprzycki, and Maria Ganzha. 2022. Assamese character recognition using convolutional neural networks. In *Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2021*, pages 851–859. Springer.