UHTelHwCC: A Dataset for Telugu Off-line Handwritten Character Recognition

Rakesh Kummari(✉) and Chakravarthy Bhagvati(✉)

School of Computer and Information Sciences, University of Hyderabad, Hyderabad 500046, India
rakeshkummarics@gmail.com, chakcs@uohyd.ernet.in

Abstract. This paper describes the creation of UHTelHwCC, a labelled dataset for Telugu off-line handwritten character recognition (HCR), and its characteristics. The form images were scanned at 300dpi and digitised into TIFF images. This paper contains two major aspects: the first, processing the forms and creating a connected component dataset and its statistics; and the second, analysis of the dataset. Preliminary results on this dataset using convolutional neural network (CNN) are presented. UHTelHwCC dataset contains samples written by 84 writers. There are a total of 75K samples of 376 classes, and these samples are divided as 60K, 5K, and 10K into training, validation, and test sets respectively. A large amount of metadata pertaining to the writer's characteristics, class distributions, and variances is also provided. It is hoped that the dataset provides the basis for developing practical Telugu off-line HCR systems and other applications. UHTelHwCC is designed to provide a standard benchmark for comparing different algorithms for Telugu HCR and help in the research and development of Telugu HCR systems. It is expected that UHTelHwCC would be like the extremely well-known handwritten numeral dataset, MNIST but for Telugu script. Our initial experiments using convolutional neural networks show performance accuracies between 84% and 90%.

Keywords: Handwritten character recognition · HCR · Off-line Telugu HCR · UHTelHwCC · Telugu dataset · HCR dataset

1 Introduction

Handwritten character recognition (HCR) is the process of converting handwritten character images into machine editable form. HCR is a well-known problem in the pattern recognition community. The research on Indic HCR has started very late compared to other scripts such Latin, Chinese, and Arabic. The reason for the above is non-availability of standard datasets and the complexity of the scripts. The research progress on Telugu scripts has got less attention than the other Indic scripts due to non-availability of standard datasets. The first and

KC Santosh et al. (Eds.): RTIP2R 2021, CCIS 1576, pp. 249–262, 2022.
https://doi.org/10.1007/978-3-031-07005-1_22

foremost step towards building HCR systems is the creation of the standard dataset. Standard datasets are essential to compare the different algorithms for HCR. In this paper, we are creating a Telugu off-line handwritten character dataset. The dataset serves as a stepping-stone in building HCR systems for Telugu and the comparison of different methods on HCR becomes easier. Telugu is an Indic script. Over 80M people are speaking in Telugu from different parts of the world. The details on the Telugu script can be found in [11,17].

To the best of our knowledge, very few handwritten character datasets for Telugu are available in the literature. The dataset used in [19] contains samples of base character only. The dataset used in [7,21] contains samples from 166 classes. But, the above dataset was originally created in on-line mode, then the obtained character images are made available in off-line mode. The character images obtained using on-line digitising device are very different from the character images which are written on a paper. These 160 classes do not cover all the characters in the Telugu script. To overcome these challenges, we are creating a new dataset with 376 classes which are sufficient to cover the entire script. The details regarding dataset creation and its characteristics are explained in the subsequent sections.

A brief study of related work is mentioned in Sect. 2. Section 3 presents how the dataset is created. Section 4 explains the characteristics of the dataset. The experiments on the dataset are presented in Sect. 5.

2 Related Work

The following handwritten character datasets for non-Indic scripts such as Arabic [2,3,8], Chinese [14,24], Korean [10], Latin [6,13,23], and Parsian [9,16,20] scripts are available in the literature. The Al-ISRA [8] dataset is an Arabic dataset that contains handwritten 10000 digits, 37000 words, 500 sentences, and 2500 signatures. It was developed at the University of British Columbia in Canada. The CENPARMI [3] dataset is an Arabic dataset that contains 13439 digits, 21426 characters, and 11375 words. Three hundred twenty-eight writers wrote these samples. The PE92 [10] dataset is a Korean handwritten character dataset that contains 235000 characters. The ETL-9 dataset is a Japanese handwritten character dataset that contains 607200 characters.

The NIST [23] dataset is an English handwritten text dataset that contains 810000 characters, digits, and 91500 text phrases. Two thousand one hundred writers have written these samples. The MNIST [13] dataset contains 70000 handwritten digits. The CEDAR [6] dataset is a collection of 14000 city and state names, 5000 zip codes, and 49000 characters and digits. The IRONOFF [22] is an on-line and off-line French and English handwritten dataset. It was developed at the University of Nantes, France. It contains 50000 cursive words and 32000 isolated characters.

The CASIA-HWDB [14], is a collection of Chinese handwritten character and text datasets, was built by the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences (CASIA). It is available in both on-line and off-line modes. The character datasets contain 3.9M samples of 7356 classes includes 7185 characters and 171 symbols. The test dataset contains 5090 pages and 1.35M character samples. The HIT-OR3C [24], is a Chinese handwritten character dataset, was built by Harbin Institute of Technology. It contains 832650 characters from 6825 classes. It is available in both on-line and off-line modes.

The following handwritten character datasets for Indic scripts such as Bangla [2,4], Devanagari [21] Oriya [2], Kannada [2,19], Malayalam [15], Telugu [7,19,21], and Tamil [19] are available in the literature. ISI Indic script handwritten databases are a collection of databases for Bangla, Devanagari, and Oriya scripts. Bangla databases contain numerals, base characters, vowel modifiers, and compound characters. Devanagari databases contain numerals and base characters. The Oriya database contains only numerals. The CMATERdb (Center for Microprocessor Applications for Training Education and Research database) [4] consists of unconstrained handwritten document images of pages, lines, words, and characters for Bangla, Devanagari, Arabic, and Telugu scripts. The PHDIndic_11 [18] is a page-level handwritten document image dataset of eleven official Indic scripts for script identification.

The PBOK dataset [2] is a handwritten text (test-line, word, and character) dataset for four different scripts such as Persian, Bangla, Oriya, and Kannada. The PBOK dataset was collected from a total of 707 text pages written by 436 writers in four different scripts. The PBOK dataset contains a total of 12565 text-lines, 104541 words/sub-words, and 553536 characters. The number of text pages for Persian, Bangla, Oriya, and Kannada scripts is 140, 199, 140, and 228 respectively. These text pages are written by 40, 199, 140, and 57 number of writers for Persian, Bangla, Oriya, and Kannada scripts respectively. The PBOK dataset contains 1787, 2820, 3108, and 4850 number of text-lines from Persian, Bangla, Oriya, and Kannada scripts respectively. The PBOK dataset includes 27073, 21255, 27007, and 29206 number of words/sub-words from Persian, Bangla, Oriya, and Kannada scripts respectively. The PBOK dataset contains 106643, 104190, 129556, and 213147 number of characters from Persian, Bangla, Oriya, and Kannada scripts respectively.

The DHCD [1] is a Devanagari handwritten character dataset. It contains 92000 samples from 46 classes and it is a balanced dataset where each class contains 2000 samples. The Amrita-MalCharDb [15], is a Malayalam handwritten character database, was developed at Amrita university. It includes 29302 samples from 85 classes includes vowels, consonants, half-consonants, vowel and consonant modifiers, and conjunct characters. Seventy-seven native Malayalam writers produced these samples.

HPL India has published a collection of handwritten character [7,21] and word datasets in on-line and off-line mode. There are three isolated character datasets for Telugu, Tamil, and Devanagari scripts. Telugu dataset contains samples from 166 classes where ≈ 270 samples per class. Tamil dataset contains samples from 156 classes where ≈ 500 samples per class. The devanagari dataset contains samples from 111 classes where ≈ 270 samples per class.

Pal et al. [19] used handwritten character datasets for Kannada, Telugu, and Tamil scripts. There are 10779 Kannada handwritten character samples of 48 classes. There are 10872 handwritten Telugu characters of 48 classes. There are 10216 Tamil handwritten character samples of 36 classes.

3 UHTelHwCC Creation

The UHTelHwCC is created from forms designed by Lakshmi et al. [12], and 84 writers fill these forms. There are 11 unique forms, numbered from 0 to 10, were written by a writer. Figure 1 represents a sample form with basic characters. Each form consists of a solid circle, grid, solid rectangles, and writer information such as name, age, and occupation. The solid circle at the top in the forms indicates the form number. The solid rectangles at the bottom of the forms are used to find the skew of the document. Each form grid consists of 180 cells arranged in 18 rows and ten columns. Rows 1,3,.., and 17 corresponds to printed characters, and rows 2,4,.., and 18 corresponds to handwritten characters. Each form contains 90 printed characters and 90 handwritten characters except the final form. Figure 2 shows the final form containing a few empty cells and a few words. The form contents include base characters, *gunitams*, compound characters (*vottus* combined with consonants), Hindu-Arabic numerals, and a few words. Base characters and a few *gunithams* can be seen in Fig. 1, whereas a few *vottus* combined with consonants, Hindu-Arabic numerals, and a few words can be seen in Fig. 2. The form images were scanned at 300dpi and digitised into TIFF images.

The steps in UHTelHwCC creation include 1. Removal of background lines, 2. Textline extraction, 3. Character extraction, and 4. CC extraction and labelling. Figure 3 shows steps involved in creating UHTelHwCC.

First, the input forms are binarized using global thresholding with the threshold value as 220. The binary forms are inverted, where ones represent the foreground. Background lines need to be removed to extract and label Telugu handwritten characters from the forms. The connected component analysis is used to remove background lines. The CCs are extracted from binary and inverted forms. It is observed that the largest CC corresponds to the background and the second-largest CC is corresponds to background lines. All the pixel values of the second-largest CC are replaced with zeros to remove background lines. Figure 4

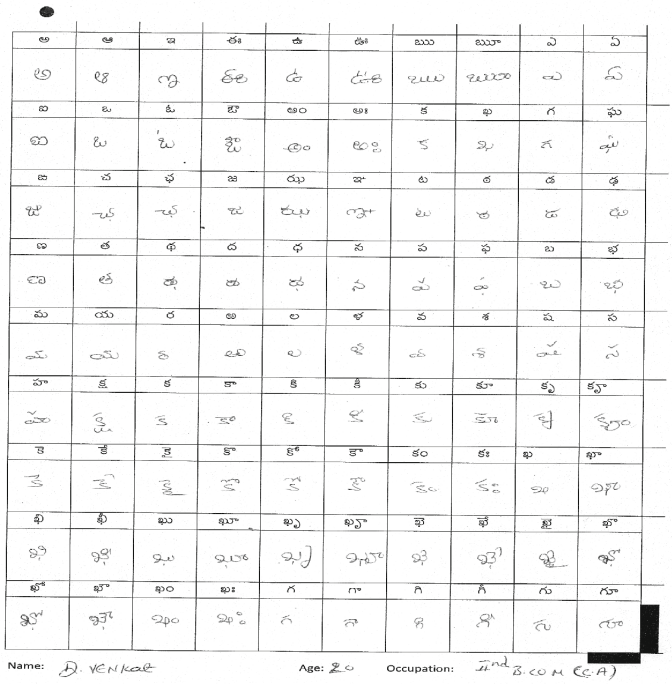

Fig. 1. Sample form with basic characters

is a form without background lines, and it is obtained by removing background lines from the form shown in Fig. 1.

The next step is the extraction of text-lines. Horizontal Projection Profile (HPP) is used for the above task. HPP shows the number of foreground pixels in each row. HPP is applied to the image that is obtained in the previous step. The peaks in HPP correspond to text-lines, and valleys indicate the gap between the text-lines. Figure 5 is HPP of the image shown in Fig. 4. In Fig. 5, X-axis represents the number of pixels, and Y-axis represents row numbers. Figure 6 and 7 show sample printed and handwritten text-lines respectively.

After extraction of text-lines, Vertical Projection Profile (VPP) is applied on each text-line to get individual Telugu characters. VPP shows the number of foreground pixels in each column. Peaks in VPP correspond to the characters, and valleys indicate the gap between characters. Then, extracted characters are labelled based on character position in the text-line. Corresponding handwritten text-lines follow Printed text-lines. Figure 8 shows randomly selected samples from extracted handwritten characters from all the forms.

Fig. 2. Sample form with words and numerals

The final step in dataset creation is the extraction and the labelling of connected components from characters obtained in the previous step. A character consists of one or more connected components. All vowels except *ee*, *aM*, and *aH* consist of a single connected component. All consonants except *gha*, *pa*, *pha*, *Sha*, *sa*, *ha*, *tha*, and *Ta* consist of single connected component. Consonants with single CC that are modified by all vowel modifiers except *R*, *RR*, *ai*, *aM*, and *aH* consist of the single connected component. Consonants with two CC that are modified by all vowel modifiers except *aa*, *o*, *oo*, and *au* consist of multiple connected components. Compound characters consist of multiple connected components. Characters with multiple components are segmented and labelled based on the component position. All the CCs are scaled to 32×32, and labels are manually verified. For example, *ka vottu* always presents to the right side of the character. Figure 9 shows sample CCs extracted from different forms.

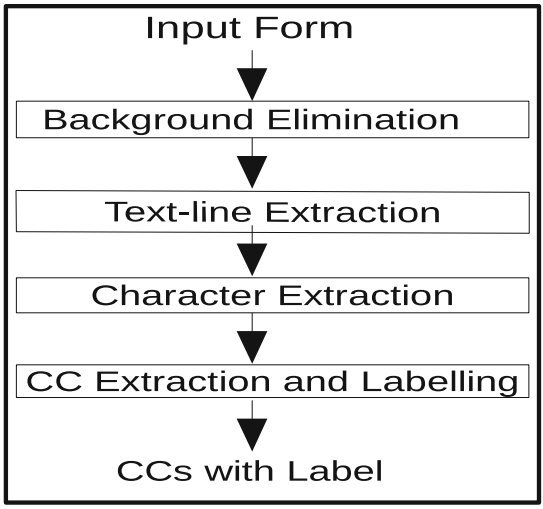

Fig. 3. Steps in UHTelHwCC creation

4 UHTelHwCC Characteristics

This section describes the dataset characteristics such as CC image type and size, writing instrument used, sample distribution, data splits, and writers' details such as the number of writers, histogram of writers' age, educational qualification. UHTelHwCC is a Telugu off-line handwritten connected component dataset, and it can be used for character recognition and writer identification. All the connected components are, in binary format, of size 32×32 and labelled. Black color sketch pen is used as a writing instrument for filling forms.

4.1 Sample Distribution

UHTelHwCC contains 75K samples (connected components) of 376 classes (distinct connected components). The number of samples in each class ranges from 1 to 3329. Figure 10 is a horizontal bar graph with the number of classes on the X-axis and number of samples in range on the Y-axis. In UHTelHwCC, 9, 6, 11, 34, 79, 163, 20, 11, 21, 19, and 3 classes are having samples in the range 1–10, 11–20, 21–30, 31–50, 51–75, 76–100, 101–200, 201–500, 501–1000, 1001–2000, and 2001–3500 respectively.

Fig. 4. A form after removing background lines

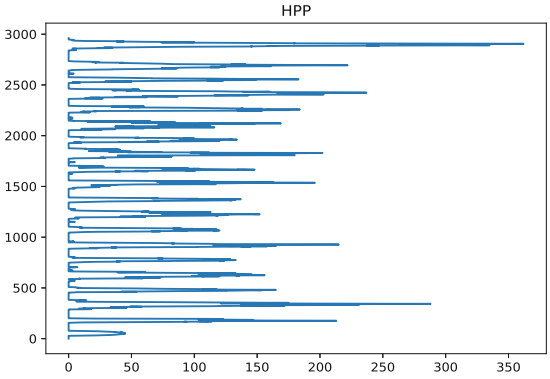

Fig. 5. Horizontal projection profile

Fig. 6. Sample extracted printed text-line

Fig. 7. Sample extracted handwritten text-line

Fig. 8. Sample extracted handwritten characters

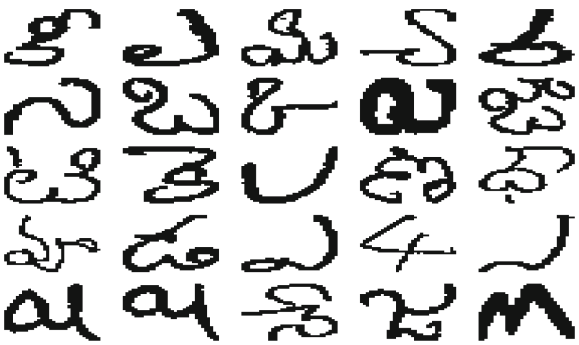

Fig. 9. Sample extracted handwritten connected components

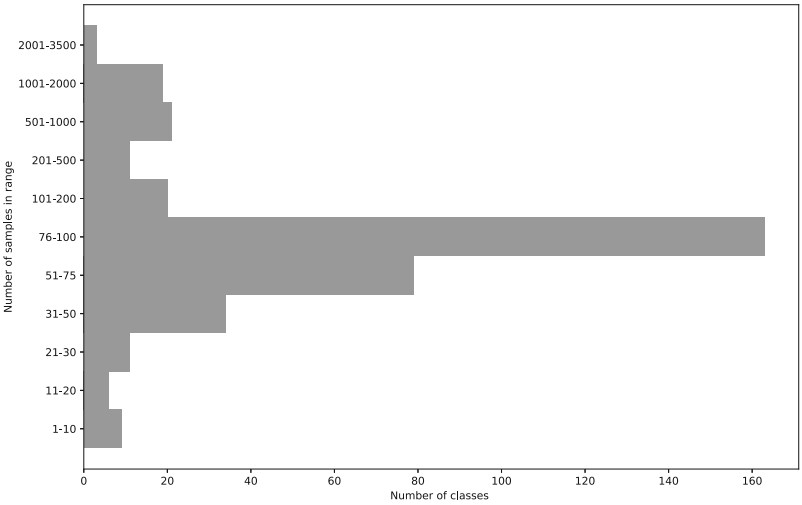

Fig. 10. Sample distribution

4.2 Data Splits

UHTelHwCC consists of 75K samples of size 32×32 in binary format. Each sample can have two values, either zero (background pixel) or one (foreground pixel). There are a total of 75K samples of 376 classes, and these samples are divided as 60K, 5K, and 10K into training, validation, and test sets, respectively. The training, validation, and test sets contain samples of 376, 362, and 367 classes, respectively. The training, validation, and test sets are used for training classifiers, tuning hyper-parameters, and evaluating the performance of the model, respectively. Nine classes have less than eight samples that are neither included in the validation set nor test set. Another five classes have samples in the range 8–15 that are not included in the validation set but the test set. The training, validation, and test sets contain 80%, $\approx$ 7%, and $\approx$ 13% of total samples respectively.

4.3 Writers Details

Eighty-four different writers filled the forms. These writers are from different age groups and educational qualifications. The handwriting of an individual changes with age [5]. Writers are from diverse age groups ranging from 10–80. Figure 11 shows the histogram of the writers' age. In Fig. 11, writes age on the X-axis and number of writers on the Y-axis presented. There are 23, 30, 22, 5, 2, 1, and 1

writers from the age groups 11–20, 21–30, 31–40, 41–50, 51–60, 61–70, and 71–80 respectively. The majority of the writers (≈90%) are from the age 10–40, and the rest of the writers are from age greater than 40.

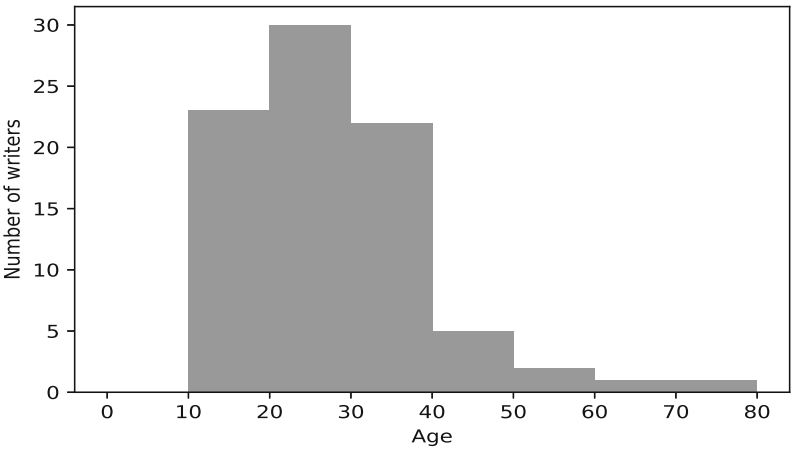

Fig. 11. Histogram of writers' age

The writers are from different educational qualifications ranges from school to master's. Figure 12 shows how many writers are from different educational qualifications. The X-axis represents qualification, and the Y-axis represents the number of writers. 20, 2, 12, 28, and 9 writers are from SSC, intermediate, diploma, bachelors, and master's, respectively.

5 Experiments

We have implemented a Convolutional Neural Network (CNN) to report the baseline accuracies on the UHTelHwCC dataset. The CNN contains 5-layers (two convolution and three fully connected layers) like LeNet-5 [13]. The number of filters in convolutional layers is 32 and 64. There are 1000, 500, and 376 nodes in fully connected layers. $ReLU$ activations are used in all the layers except the output layer. *Softmax* activations are used in output layer. The *adam* optimizer and categorical cross-entropy loss function, L_2 regularization (10^{-5}), and dropout (0.5) are used. Figure 13 shows training, validation accuracies, and training, validation loss for 15 epochs. Table 1 shows the CNN accuracy and loss computed on training, validation, and test sets. We have obtained 89.19%

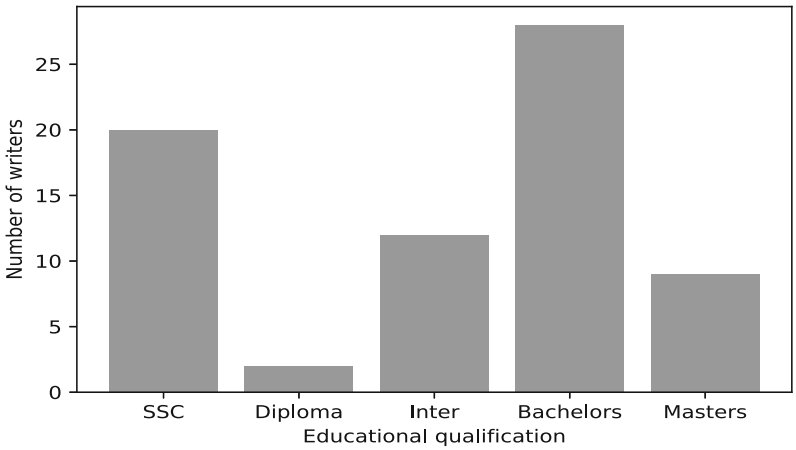

Fig. 12. Histogram of writers' educational qualification

average accuracy and 0.6 average loss over ten runs on the test set. UHTelHwCC dataset can be used to develop Telugu off-line HCR systems and compare the performance of different HCR algorithms.

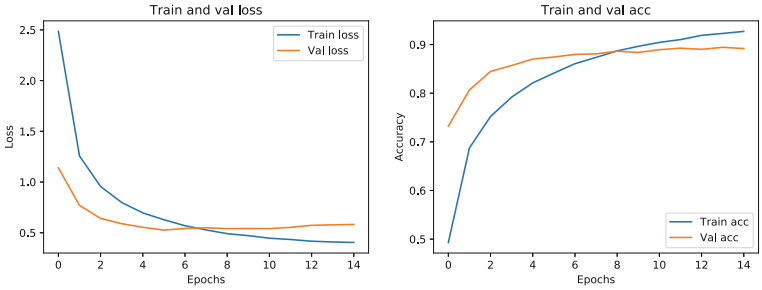

Fig. 13. CNN accuracy and loss

Table 1. CNN accuracy and loss

Data split	Size	Accuracy(%)	Loss
Training set	60K	99.097 ± 0.053	0.216 ± 0.002
Validation set	5K	89.07 ± 0.265	0.599 ± 0.016
Test set	10K	89.195 ± 0.229	0.597 ± 0.011

6 Conclusion

The main contribution of this paper is the creation of UHTelHwCC, a dataset for Telugu off-line handwritten character recognition. This dataset contains 75K samples of 376 classes written by 84 writers. This dataset contains sufficient variations in it and a large number of classes making it suitable for current and future research in HCR. A large number of writers, the variation in their educational qualifications and their ages make the dataset a potential benchmark for handwriting analysis. Connected component analysis, HPP, and VPP are used in form processing to automate labelling of CCs. Manually verified the labels. Finally, we have implemented a CNN and obtained 89.19% average accuracy and 0.597 average loss over ten runs on the test set.

Acknowledgment. We thank Prof. Atul Negi for providing scanned copies of handwritten forms. The first author acknowledges the financial support received from the Council of Scientific and Industrial Research (CSIR), Government of India, in the form of a Senior Research Fellowship.

References

1. Acharya, S., Pant, A.K., Gyawali, P.K.: Deep learning based large scale handwritten Devanagari character recognition. In: 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pp. 1–6. IEEE (2015)
2. Alaei, A., Pal, U., Nagabhushan, P.: Dataset and ground truth for handwritten text in four different scripts. Int. J. Pattern Recogn. Artif. Intell. **26**(04), 1253001 (2012)
3. Alamri, H., Sadri, J., Suen, C.Y., Nobile, N.: A novel comprehensive database for Arabic off-line handwriting recognition. In: Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR, vol. 8, pp. 664–669 (2008)
4. Das, N., Acharya, K., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: A benchmark image database of isolated Bangla handwritten compound characters. Int. J. Doc. Anal. Recogn. (IJDAR) **17**(4), 413–431 (2014)
5. Huber, R.A., Headrick, A.M.: Handwriting Identification: Facts and Fundamentals. CRC Press, Boca Raton (1999)
6. Hull, J.J.: A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. **16**(5), 550–554 (1994)
7. Jayaraman, A., Sekhar, C.C., Chakravarthy, V.S.: Modular approach to recognition of strokes in Telugu script. In: Ninth International Conference on Document Analysis and Recognition, ICDAR 2007, vol. 1, pp. 501–505. IEEE (2007)
8. Kharma, N., Ahmed, M., Ward, R.: A new comprehensive database of handwritten arabic words, numbers, and signatures used for ocr testing. In: Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No. 99TH8411). vol. 2, pp. 766–768. IEEE (1999)
9. Khosravi, H., Kabir, E.: Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern Recogn. Lett. **28**(10), 1133–1141 (2007)

10. Kim, D.H., Hwang, Y.S., Park, S.T., Kim, E.J., Paek, S.H., Bang, S.Y.: Handwritten Korean character image database PE92. IEICE Trans. Inf. Syst. **79**(7), 943–950 (1996)
11. Kummari, R., Bhagvati, C.: UHTelPCC: a dataset for Telugu printed character recognition. In: Santosh, K.C., Hegadi, R.S. (eds.) Recent Trends in Image Processing and Pattern Recognition, RTIP2R 2018. CCIS, vol. 1037, pp. 24–36. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-9187-3_3
12. Lakshmi, A.V., Krishna, T.S., Negi, A.: Dataset generation for OCR. Int. J. Comput. Trends Technol. (IJCTT) **2**(1), 48–51 (2011)
13. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
14. Liu, C.L., Yin, F., Wang, D.H., Wang, Q.F.: CASIA online and offline Chinese handwriting databases. In: 2011 International Conference on Document Analysis and Recognition, pp. 37–41. IEEE (2011)
15. Manjusha, K., Kumar, M.A., Soman, K.: On developing handwritten character image database for Malayalam language script. Eng. Sci. Technol. Int. J. **22**(2), 637–645 (2019)
16. Mozaffari, S., Faez, K., Faradji, F., Ziaratban, M., Golzan, S.M.: A comprehensive isolated Farsi/Arabic character database for handwritten OCR research (2006)
17. Negi, A., Bhagvati, C., Krishna, B.: An OCR system for Telugu. In: Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 1110–1114. IEEE (2001)
18. Obaidullah, S.M., Halder, C., Santosh, K., Das, N., Roy, K.: PHDIndic_11: page-level handwritten document image dataset of 11 official Indic scripts for script identification. Multimedia Tools Appl. **77**(2), 1643–1678 (2018). https://doi.org/10.1007/s11042-017-4373-y
19. Pal, U., Sharma, N., Wakabayashi, T., Kimura, F.: Handwritten character recognition of popular south Indian scripts. In: Doermann, D., Jaeger, S. (eds.) Arabic and Chinese Handwriting Recognition, SACH 2006. LNCS, vol. 4768, pp. 251–264. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78199-8_15
20. Solimanpour, F., Sadri, J., Suen, C.Y.: Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language (2006)
21. Swethalakshmi, H., Jayaraman, A., Chakravarthy, V.S., Sekhar, C.C.: Online handwritten character recognition of Devanagari and Telugu characters using support vector machines. In: Tenth International Workshop on Frontiers in Handwriting Recognition, Suvisoft (2006)
22. Viard-Gaudin, C., Lallican, P.M., Knerr, S., Binter, P.: The IRESTE on/off (IRONOFF) dual handwriting database. In: Proceedings of the Fifth International Conference on Document Analysis and Recognition, ICDAR 1999 (Cat. No. PR00318), pp. 455–458. IEEE (1999)
23. Wilkinson, R., et al.: The first census optical character recognition systems conf.# NISTIR 4912. The US Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD (1992)
24. Zhou, S., Chen, Q., Wang, X.: HIT-OR3C: an opening recognition corpus for Chinese characters. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 223–230. ACM (2010)