A Survey on Offline Handwritten North Indian Script Symbol Recognition

Jenis J. Macwan
Information Technology
Dharamsinh Desai Institute of Technology
Nadiad, India
jennis.mekwan3@gmail.com

Mukesh M. Goswami
Faculty of Technology
Dharamsinh Desai Institute of Technology
Nadiad, India
mukesh.goswami@gmail.com

Archana N. Vyas
Faculty of Technology
Dharamsinh Desai Institute of Technology
Nadiad, India
archu0685@gmail.com

***Abstract*—North Indian scripts are used in 70% of the regions of India. Devanagari, Bangla, Gujarati, Oriya, and Gurumukhi are the major North Indian scripts listed in decreasing order of their usage. There are ample amount of work done on printed symbol recognition for various North Indian scripts, but handwritten symbol recognition still needs further attention. This paper represents the thorough study of the work done so far for handwritten symbol recognition for major North Indian script. Furthermore, the challenges of each script and various feature extraction techniques and classifiers used are also discussed. The paper will provide a proper startup and almost complete scenario of the state of the art in offline handwritten symbol recognition from North Indian script.**

Keywords—Gujarati; Handwritten character recognition; Bangla; Devanagari; Gurumukhi; Oriy; OCR.

I. INTRODUCTION

Optical Character Recognition (OCR) is a technology for converting the scanned images of machine-printed or handwritten text (numerals, letters or symbols) into a computer processable format. OCR is one of the emerging technologies, having a great importance in digital image processing. It has many applications [2] like data entry for business documents, government records, bill processing, etc. OCR is categorized into two parts known as handwritten character recognition and printed character recognition [1]. Figure 1 shows the components of OCR. We can find works on Indian languages like Hindi [22], Kannada, Devanagari [23], Gurumukhi [24], Bangla [25], Malayalam, Oriya [26], etc., in the literature.

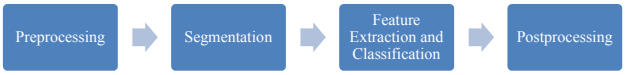

Fig. 1. Components of OCR

Handwritten character recognition could be further divided into on-line and off-line character recognition. In on-line, handwriting is captured and stored in digital form via different means. A special pen is used to write on an electronic surface. The pen's movement is captured by the surface to form the symbols. In off-line, symbols are handwritten on a surface (such as a sheet of paper) and are scanned and digitally stored in grayscale format. The advantage of on-line handwritten character recognition over offline handwritten character recognition is that more information can be captured, like direction, speed, and order of strokes of handwriting. Handwriting is more versatile than printed characters as it may differ in shape, context, size, and style from person to person. This paper focuses on work done for classification of handwritten symbols of North Indian script as there are many challenges and aspects to be covered. All scripts have their own individuality in its shape, style and in amount of consonants, vowels, numerals and modifiers. This paper focuses on different aspects of handwritten symbols, such as:

- Symbol set complexity.
- Availability of dataset.
- Features used for extracting distinct properties of characters.
- Different classifiers used for recognition.

The paper is divided into as follows: section II and III will discuss Introduction to the scripts and Related works respectively. Furthermore, the section IV includes Analysis of classifiers and features followed by Conclusion in section V.

II. INTRODUCTION TO THE SCRIPTS

Almost all Indian languages have a phonetic base and are derived from ancient Brahmi script. India has ten major scripts namely, Devanagari, Bangla, Gujarati, Oriya, Gurumukhi, Kannada, Telugu, Tamil, Malayalam, and Urdu. From these major scripts 18 Indian official languages are derived. The North Indian scripts dominate most of the usage that are namely Devanagari, Bangla, Gujarati, Oriya, and Gurumukhi. Devanagari script is used by over 500 million people and adopted for around 120 languages like, Hindi, Marathi, Gujarati, Sanskrit, and Nepali etc. It has 47 primary characters from which the 33 are consonants and 14 are vowels. One more

978-1-4673-9939-5/16/$31.00 ©2016 IEEE

feature of the Devanagari script is the presence of "Shirolekha" or "Matra" (the horizontal line on the head). Devanagari characters are decomposed into three types: a core strip that contains main characters, an upper modifier symbol strip and a lower modifier symbol strip.

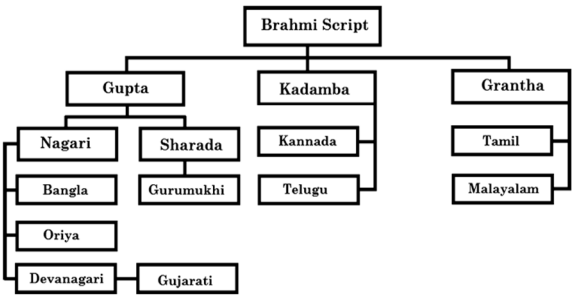

Fig. 2. Evolution of Indian scripts

Bangla script is used for Bengali and Assamese languages and is 6th widely used writing system in the world. There is no letter case in Bangla script and also consist the "Shirolekha". The analysis says that approximately 22.02% of the postal documents in West Bengal are written in Bangla script [3]. Bangla script consists 36 consonants and 21 vowels. The 10 numerals (0 to 9) of Bangla script do not have any horizontal line. Gujarati is third widely used script in India used to generate the Gujarati language. One feature that makes Gujarati different from another northern script is a lack of "Shirolekha," which makes the segmentation work considerably hard. The conjunct characters are also more in the Gujarati language from which some are used very frequently while some are not used at all. Apart from consonants and vowels, Gujarati also has some modifiers that can be in upper, lower or middle part of the characters. Each modifiers corresponding to each vowel and vowels are used to change the sound of constants. Oriya script is used for Oriya language and is syllabic in nature. The Oriya language is written with the consonant-vowel sequence as a unit. Oriya script is very curvy in writing. The conjunct characters of Oriya script are divided into two types. First is the "northern" type and the other is "southern" type. Just like Gujarati script, Oriya also does not have a horizontal line on its head [28]. Considering the decreasing order in the usage of North Indian scripts, Gurumukhi comes last with 29 million users and is used to derive the Punjabi language. Compared to other scripts, Gurumukhi, and Oriya have less amount of conjunct characters. All these scripts have their distinct properties that make them different with the evolution of time even if they are derived from the same script. Table 1 denotes all the information regarding the North Indian scripts which includes the number of consonants, vowels and modifier existence and how many speakers are there of each script.

III. Related work

This section discusses the work already done on these five scripts describing each script work.

A. *Devanagari:*

For offline handwritten recognition generally, structural features give better results. Table 2 shows the results gained so far in Devanagari handwritten character recognition. For numeral the maximum accuracy gained is 98.62% [34] and 98.19% [36] is maximum accuracy gained in handwritten character recognition. Mostly for Devanagari handwritten symbol recognition, classifiers used are SVM and neural networks.

B. *Bangla:*

The analysis says that approximately 22.02% of the postal documents are of handwritten in Bangla script in West Bengal [3]. As the work done in Bangla handwritten character is vast many segmentation techniques has been found to decompose the characters [4]. Directional features, gradient features, etc. [12][7][6] has been used for Bangla script. Liu and Suen [6] have gained maximum accuracy for handwritten numeral with 99.4% accuracy. Table 3 shows the related work for Bangla script.

C. *Gujarati:*

For Gujarati the work on offline handwritten is less, and the accuracy level also requires attention. A good attempt is made by Baheti M.J. and Kale K.V. [29] for numeral recognition using affine invariants moments as a feature, which is the general transform of space coordinates of an image, has gained 92.28% accuracy. Apurva A. Desai [30][32] has done some work for both Gujarati numerals and characters gaining 81.66% and 86.66% respectively. Table 4 shows the maximum results gained for offline Gujarati character recognition.

D. *Oriya:*

The script written on palm leaves is because the curvy nature of it [28] and does not contain many straight lines. For Oriya script work is not done much but accuracy level is good as seen in table 5. As per the analysis it is seen that mostly directional features [18][19] is used for offline numeral recognition. Water reservoir based technique is one of the segmentation technique used for Oriya script segmentation [8].

E. *Gurumukhi:*

Gita Sinha [16] has gained maximum accuracy for handwritten numeral for Gurumukhi script gaining 99.73% accuracy. Table 6 denotes some of the works done on Gurumukhi script where mostly zoning features [14][16][17] are used which has gained good amount of accuracy. Segmentation becomes easier based on the presence of the horizontal line [5].

Table I. Information regarding North Indian scripts

Script	Languages	Consonants	Vowels	Existence of modifiers	Speakers
Devanagari	Many	33	14	Yes	500 million
Bangla	Assamese, Bengali	36	21	Yes	96 million
Gujarati	Gujarati	34	12	Yes	46 million
Oriya	Oriya	34	14	Yes	33 million
Gurumukhi	Punjabi	35	9	Yes	29 million

Table II. Work done on offline handwritten Devanagari script

References	Data type	Features	Classifier	Dataset	Accuracy
Mahesh Jangid, Renu Dhir, Rajneesh Rani, Kartar Singh [34]	Numeral	Zoning Density, Background Directional Distribution (BDD)	SVM	22546	98.62%
U. Bhattacharya, S.K. Parui, B. Shaw, K. Bhattacharya [35]	Numeral	Scalar features	1) MLP 2)HMM	22535	1)92.83%, 2)87.69%
Kapil Mehrotra, Saumya Jetley, Akash Deshmukh, Swapnil Belhe [36]	Character	Unspecified	Convolution Neural Networks	41359	98.19%

Table III. Work done on offline handwritten Bangla Script

References	Types of dataset	Features	Classifier	Dataset	Accuracy
Liu and Suen [6]	Numeral	Gradient	CFPC, DLQDF	23,392	99.4%
Purkait and Chanda [8]	Numeral	Morphological	MLP	23,392	97.75%
Lu et al.[7]	Numeral	Directional & Density	SOM	16000	97.28%
Chaudhuri and Majumdar [9]	Basic	Curvelet	SVM	3,900	95.5%
Roy [11]	Basic	DCCH	Quadratic	14,879	93.9%
Bhattacharya [10]	Basic	LCCH	MLP	21,725	92.14%
Pal [12]	Compound	Directional	MQDF	20,543	85.90%
Das [13]	Compound	Shadow, LR, QT	SVM	19,765	80.51%

Table IV. Work done on offline handwritten Gujarati script

References	Types of dataset	Features	Classifier	Dataset	Accuracy
Ravi Nagar, Suman K. Mitra [31]	Numeral	Orientation estimation	SVM	12889	98.93%
Baheti M. J., Kale K. V. [29]	Numeral	Affine invariant moments	KNN, PCA, SVM, Gaussian distribution function	1600	SVM=92.28%,Gaussian=87.2%, KNN=90.04%, PCA=84.1
Apurva A.Desai [30]	Numeral	Profile vector	Feed Forward Back propagation neural network	3000	81.66%

Apurva A.Desai [32]	Character	Aspect ratio and extent method	SVM, KNN	7960	86.66%
Jayashree R. Prasad, Dr. U.V.Kulkarni, Rajesh S. Prasad [33]	Character	Grid features	Neural Network	Unspecified	71.66%

Table V. Work done on offline handwritten Oriya script

Author	Types of dataset	Features	Classifier	Dataset	Accuracy
Pal [19]	Numerals	Directional	MQC	5,638	98.40%
Roy [18]	Numeral	Directional	Quadratic	3,850	94.81%
Bhowmick [20]	Numeral	Scalar	HMM	5,970	90.50%
Pal [21]	Character	Curvature	Quadratic	18,190	94.6%

Table VI. Work done on offline handwritten Gurumukhi script

References	Types of dataset	Features	Classifier	Dataset	Accuracy
Gita Sinha [16]	Numeral	Zone based	SVM with RBF kernels	1500	99.73%
Garg [14]	Character	Structural	Neural network	6,900	96%
Kartar Singh Siddharth, Renu Dhir [17]	Character	Zoning density and background directional distribution features	SVM with RBF kernel	7000	95.04%
Sharma and Jhajj [15]	Character	Zone	KNN and SVM	5,125	73.02%

IV. ANALYSIS OF CLASSIFIERS AND FEATURES

Out of many components of OCR, feature extraction and classification is very crucial. Features are the distinct properties of characters based on which the classifiers recognizes the characters. Many feature extraction and classification techniques exist for character recognition. There are different kinds of features based on a character's shape, size and context etc like statistical or global features, series expansion co-efficient, and shape based features etc. As the handwritten may differ from person to person the more prominent features for it could be shape based features. Shape based features has also two types, which are contour based and region based features [37]. As per the analysis it is seen that many features have been used for handwritten script recognition in North Indian languages like moment based features [29], gradient features [6], zoning features [34], directional features [18], geometrical features, statistical features, morphological features, hybrid features [27] and many others. Apart from feature extraction many classification techniques are used for offline handwritten character recognition like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Multi Layer Perceptron (MLP), and Artificial Neural Networks (ANN) etc. KNN is known as the most basic and benchmark classifier being the simplest in the method. SVM is used for both classification and regression and can work on both linear and non-linear classification. ANN is more complex of a classifier having many hidden layers and SVM advantages on it as SVM uses maximum margin which gives the best boundary for classification and separation of a dataset while ANN uses gradient decent which may or may not provide the best boundary for separating the dataset. The analysis shows that mostly work of classification is done using KNN [32] and SVM [9].

V. MAJOR CHALLENGES AND ISSUES

Apart from features, many problems occur like rotation problem, scaling problem, and shifting problem etc. One of the major problems is the lack of dataset as it is not readily available from any source and one has to generate it manually. Furthermore the lack of the global platform and lack of the co-ordination between researchers also affects the progress of character recognition. The work on Gujarati, Oriya and Gurumukhi is limited. Major work is done on numerals only. In Gujarati, no work for handwritten conjunct character is done. Furthermore the amount of conjunct characters of some scripts like Bangla, Devanagari, and Gujarati are more, like Gujarati has more than 990 and Bangla has 280 around conjunct characters.

VI. CONCLUSION

This paper included the discussion of the properties of all major North Indian scripts and the work done on the offline handwritten script of North India. The challenges and issues of the scripts have been discussed which helps the researcher working in this field. Furthermore it also includes the survey and analysis of different feature extraction and classification techniques applied on the major five North Indian scripts namely Devanagari, Bangla, Gujarati, Oriya and Gurumukhi. For offline handwritten character recognition, apart from Bangla, other mentioned scripts here still require further attention and many problems need to be resolved.

References

[1] Moro, Kamal, et al. "New approach of feature extraction method based on the raw form and his skeleton for gujarati handwritten digits using neural networks classifier." *Annals of the University Dunarea de Jos of Galati Fascicle III: Electrotechnics, Electronics, Automatic Control & Informatics* 37.1, 2014.

[2] Patel, Chhaya, and Apurva Desai. "Gujarati handwritten character recognition using a hybrid method based on binary tree-classifier and k-nearest neighbour." *International Journal of Engineering Research and Technology*. Vol. 2. No. 6. ESRSA Publications, 2013.

[3] Pal, Umapada, Ramachandran Jayadevan, and Nabin Sharma. "Handwriting recognition in indian regional scripts: a survey of offline techniques." *ACM Transactions on Asian Language Information Processing (TALIP)* Vol 11.1, 2012.

[4] Bishnu, A. And Chaudhuri, B. B., "Segmentation of Bangla handwritten text into characters by recursive contour following," *Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR'99)*, pp. 402–405, 1999.

[5] Kumar, R. And Singh, A., "Detection and segmentation of lines and words in Gurmukhi handwritten text," *Proceedings of the 2nd IACC'10*, pp. 353–356, 2010.

[6] Liu, C. L. And Suen, C. Y." A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters." *Patt. Recog.Vol 42*, 12, pp/3287–3295,2009.

[7] Lu, S., Tu, X., And Lu, Y." An improved two-layer SOM classifier for handwritten numeral recognition".In *Proceedings of the International Conference on Intelligent Information Technology pp.* 367–371,2008.

[8] Purkait, P. And Chanda, B. "Off-line recognition of handwritten Bengali numerals using morphological features." In *Proceedings of the 12th International Conference on the Frontiers of Handwriting Recognition* pp. 363–368,2010.

[9] Chaudhuri, B. B. And Majumdar, "A. Curvelet-based multi SVM recognizer for offline handwritten bangla: A major Indian script". In *Proceedings of the 9th International Conference on Document Analysis and Recognition* pp. 491–495,2007.

[10] Bhattacharya, U., Shridhar, M., And Parui, S. K." On recognition of handwritten Bangla characters",In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing* pp. 817–828,2006.

[11] Roy, A., Bhowmik, T. K., Parui, S. K., And Roy, U." A novel approach to skew detection and character segmentation for handwritten Bangla words." In *Proceedings of the Conference on Digital Image Computing: Techniques and Applications* pp. 30–37,2005.

[12] Pal, U., Wakabayashi, T., And Kimura, F. "Handwritten Bangla compound character recognition using gradient feature". In *Proceedings of the 10th Information Technology Conference pp.*208–213, 2007.

[13] Das, N., Das, B., Sarkar, R., Basu, S., Kundu, M., And Nasipuri, M., Handwritten "Bangla basic and compound character recognition using MLP and SVM classifier". *J. Comput.Vol 2*, 2, pp.109–115, 2010.

[14] Garg, N. "Handwritten Gurumukhi character recognition using neural networks". Master's thesis, Thapar University, Patiala, 2009.

[15] Sharma, D. And Jhajj. "Recognition of isolated handwritten characters in Gurmukhi script". *Int. J. Comput. Appl. Vol 4*, pp 9–17, 2010.

[16] Rani, Gita Sinha Rajneesh, and Renu Dhir. "Handwritten Gurmukhi numeral recognition using Zone-based hybrid feature extraction techniques.", 2012.

[17] Siddharth, Kartar Singh, Renu Dhir, and Rajneesh Rani. "Handwritten Gurmukhi Character Recognition Using Zoning Density and Background Directional Distribution Features." *International Journal of Computer Science and Information Technologies* 2, no. 3 pp: 1036-1041, 2011.

[18] Roy K., Tandra Pal, Umapada Pal, and Fumitaka Kimura. "Oriya handwritten numeral recognition system." In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pp. 770-774. IEEE, 2005.

[19] Pal, Umapada, Nabin Sharma, Tetsushi Wakabayashi, and Fumitaka Kimura. "Handwritten numeral recognition of six popular Indian scripts." In*Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, pp. 749-753. IEEE, 2007.

[20] Bhowmik, Tapan K., Swapan K. Parui, Ujjwal Bhattacharya, and Bikash Shaw. "An hmm based recognition scheme for handwritten oriya numerals." In *Information Technology, 9th International Conference on*, pp. 105-110. IEEE, 2006.

[21] Pal, Umapada, Tetsushi Wakabayashi, and Fumitaka Kimura. "A system for off-line Oriya handwritten character recognition using curvature feature." In*Information Technology, 10th International Conference on*, pp. 227-229. IEEE, 2007.

[22] Hanmandlu, Madasu, et al. "Fuzzy model based recognition of handwritten hindi numerals using bacterial foraging." *Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference on*. IEEE, 2007.

[23] Jayadevan, R., et al. "Offline recognition of Devanagari script: A survey."*Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 41.6 pp. 782-796, 2011.

[24] Mahto, Manoj Kumar, Karamjit Bhatia, and R. K. Sharma. "Combined horizontal and vertical projection feature extraction technique for Gurmukhi handwritten character recognition." *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*. IEEE, 2015.

[25] Pal, Arghya. "Bengali handwritten numeric character recognition using denoising autoencoders." *Engineering and Technology (ICETECH), 2015 IEEE International Conference on*. IEEE, 2015.

[26] Sarangi, Pradeepta K., Ashok K. Sahoo, and P. Ahmed, "Recognition of Isolated Handwritten Oriya Numerals using Hopfield Neural Network", International Journal of Computer Applications, Vol. 40 No. 8, pp. 36-42, 2012.

[27] GARG, N., "Handwritten Gurumukhi character recognition using neural networks," Master's thesis, Thapar University, Patiala, 2009.

[28] "Writing systems and languages of the world," [online] Available: http://www.omniglot.com, [Accessed: 23-Nov-2015].

[29] MJ, Baheti, and K. V. Kale. "Gujarati Numeral Recognition: Affine Invariant Moments Approach.".

[30] Desai, Apurva A. "Gujarati handwritten numeral optical character reorganization through neural network." *Pattern recognition* Vol 43.7, pp. 2582-2589, 2010.

[31] Nagar, Ravi, and Suman K. Mitra. "Feature extraction based on stroke orientation estimation technique for handwritten numeral." *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*. IEEE, 2015.

[32] Desai, Apurva A. "Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space." *CSI Transactions on ICT*, pp. 1-7, 2015.

[33] Prasad, Jayashree R., U. V. Kulkarni, and Rajesh S. Prasad. "Offline handwritten character recognition of Gujrati script using pattern matching."*Anti-counterfeiting, Security, and Identification in Communication, 2009. ASID 2009. 3rd International Conference on*. IEEE, 2009.

[34] Jangid, Mahesh, et al. "SVM classifier for recognition of handwritten devanagari numeral." *Image Information Processing (ICIIP), 2011 International Conference on*. IEEE, 2011.

[35] Bhattacharya, U., et al. "Neural combination of ANN and HMM for handwritten Devanagari numeral recognition." *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[36] Mehrotra, Kapil, et al. "Unconstrained handwritten Devanagari character recognition using convolutional neural networks." *Proceedings of the 4th International Workshop on Multilingual OCR*. ACM, 2013.

[37] Zhang, Dengsheng, and Guojun Lu. "Review of shape representation and description techniques." *Pattern recognition* Vol. 37.1, pp. 1-19, 2004.