

Review of Advances in Digital Recognition of Indian Language Manuscripts

Bhavesh Kataria^{1*}, Dr. Harikrishna B. Jethva²

¹Research Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India

²Associate Professor, Department of Computer Engineering, Government Engineering College, Patan, Gujarat, India

ABSTRACT

Digital content creation and document management in Indian languages are in progressing stage. OCR has become an administrative requirement for effective governance and daily activities. Scripts including those from medieval to contemporary time are of literary and political importance. The present research initiatives highlights the importance and needs of efforts in recognition of printed and handwritten documents written in languages of Indian origin. This paper is aims at reviewing the state of various scripts in use including those from medieval to present era and explores the prospective of digital recognition of handwritten and printed texts and thereby pointing towards futuristic trends in developing restoration software for Indic scripts. While OCRs for Indic scripts like Devanagari has attained good results and still improving the accuracy levels, many medieval and ancient scripts have very little attempts. Challenge is due to the number of languages and their diverse scripts. The scarcity of digitized linguistic resources makes the task a tougher one. The paper also highlights on the characteristics and challenges of recognition of scripts of Indic origin. Largely the digital recognition is limited to simple numerals and isolated characters. The paper enumerates the highest known performance of OCR attempts for important Indic scripts and suggests possibilities of using various approaches including statistical and soft computing for recognizing scripts of medieval times in particular.

Keywords : OCR, Indic Scripts, Pattern Recognition. Character Recognition, Devanagari

I. INTRODUCTION

Preservation and restoration of textual knowledge that has been distributed in the contemporary and archival documents and making it available for the academics and administration purposes has been a primary goal of many document-digitization and technology development initiatives including those by government-funded projects carried out at TDIL and C-DAC [73, 74, 75].

Sophistication in Optical Character Recognition (OCR) systems controls the requirement of automatic processing of the digitized text. OCR refers to the mechanical or electronic conversion of images of typed, handwritten or printed text into

machine-encoded text. It is a specialized field emerging from the amalgamation of pattern recognition, image processing, and natural language processing. In 1951 GISMO, the first OCR of the modern world was presented by David Shepard that would convert printed text into machine language [45]. Software products like those from Adobe, ABBYY FineReader, ReadIris are a few leading examples of OCR systems for recognizing printed and handwritten texts but largely limited to Roman scripts.

This digitization has led to a variety of applications with potential for digital communication in local Indian languages. This motivates researchers to work for developing OCR for Indian scripts including

those that are less widely used. OCR for Indian scripts such as Devanagari and Telugu have achieved a convincing level of digital recognition of texts while some others are still in their introductory stages. Notably, the OCR for Indic scripts has a wide scope for improvement with correct and efficient outcomes. There is a need for an initiative to focus more on OCRs for less widely used scripts but important from a historical perspective. This would not only help the preservation of the script but also reviving it for current digital communication.

lists 22 major Indian languages [74, 75] currently used by Indian population. With hundreds of dialects to these languages used in small geographical pockets, India is seen as a multi-lingual region. Over the last few centuries, millions of documents are created that contain great literary values using different scripts. Over the timeline, these some scripts died while some others have evolved and mark their presence even today, with some of their modern forms with more defined rules and style. Therefore investing in OCR focused on scripts of Indic origin is rationally important and justified.

Table 1: Status of Endangered Language in India [46,49]

Total	Vulnerable	Definitely Endangered	Severely endangered	Critically endangered	extinct	
199	Gondi, Kumaoni, Kurux (India), Tulu, Meithei, Tamang, Kui, Khasi, Bodo, Mundari, Angika, Kokborok, Mizo, Karbi, Ho, Garhwali, Sora, Konyak, Ao, Irula, Kharia, Korku, Tshangla, Thado, Adi, Lhota, Nyishi, Rabha, Sherpa, Tangkhul, Angami, Phom, Dimasa, Ladakhi, Simi, Kabui, Yimchungru, Chokri, Sangtam, Mao, Bishnupriya Manipuri Creole, Chang, Nruanghmei, Rengma, Cuona Menba, Hmar, Paite, Wancho, Bhumji, Kheza, Gutob, Minyong, Tangsa, Khiamngan, Maram, Apatani, Galo, Korwa, Liangmai, Zeme, Nocte, Tagin, Mzieme, Koda, Anal, Maring, Bangni, Khoirao, Manchad, Padam, Hrangkhoh, Pochuri, Khampti, Taruang, Rongpo, Miju, Bokar, Sherdukpen, Balti, Padri, Purik, Spiti	Kangdi, Mandeali, Mising, Mahasui, Kurru, Kuvi, Limbu, Malto, Kului, Kodagu, Badaga, Chambeali, Kolami, Konda, Jaunsari, Bhadravahi, Kinnauri, Churahi, Kachari, Koch, Lepcha, Deori, Juang, Tiwa, Mara, Biete, Gangte, Nahali, Bawm, Hill Miri, Idu, Motuo Menba, Asur, Sulung, Gorum, Kom, Miji, Singpho, Turi, Aka, Bunan, Bharmauri, Moyon, Brokshat, Bori, Jangshung, Milang, Tinan, Darma, Byangsi, Kanashi, Khamba, Lishpa, Dakpa, Howa, Pasi, Zaiwa, Kundal Shahi, Bhalesi, Jad, Khasali, Koro, Zangskari	Geta, Remo, Aiton, Tai Phake, Mech, A'tong	Parji, Sirmaudi, Gadaba, Koraga, Pangvali, Kuruba, Bangani, Lamgang, Muot, Naiki, Pu, Baghati, Takahanyilang, Aimol, Birhor, Kota, Ahom, Luro, Nihali, Andro, Sanenyo, Pengo, Chairel, Koireng, Toda, Toto, Rangkas, Tarao, Purum, Sengmai, Lamongse, Mra, Na, Tolcha Aka, Handuri, Ruga, Shompen, Tai Nora, Tai Rong, Tangam, Onge, Sentilese, Jarawa, Great andamanese, Bellari, Langrong, Manda		

Creation of digital library content and generic OCR including those for multilingual [75] scripts has raised need to invest in digitizing document written using ancient, medieval and modern scripts of the Indian subcontinent. The state of the art in OCR for Indic script developments has not reached the level of maturity compared to that of Roman scripts and still is far from being able to reliably recognize the texts. The Constitution of India in its eighth schedule

II. INDIC WRITING SYSTEM AND SCRIPTS

Indian states attributing to rich heritage and culture has people speaking at least 780-850 different languages [47] and 1,683 dialects or their “mother tongues” using about 86 different scripts (People's Linguistic Survey of India, 2015). In India, there are officially approved 22 regional languages, which primarily follow 14 scripts to write them. For

various reasons, India has lost nearly 250 languages [76, 77] in the last 50 years which raises alarm for increased speed in resorting literary heritage in spoken and written forms. Indic writing systems embrace the syllabic Kharosthi and semi-alphabetic Brahmi scripts of ancient and medieval India. Kharosthi script did not develop into writing system subsequently, However, Brahmi succeeded to become the forerunner for all major scripts used across Southeast Asia, India, Indonesia, and the Tibetan region for writing the languages (See Figure 1).

The descent graph shows that Brahmi in the northern region has evolved into the Gupta scripts, which derives many scripts marking their presence even today in contemporary writing styles. early 75% of the population and attributed to Indo-Aryan linguistic family. Scripts of medieval and contemporary Indian scripts including Devanagari, Gurmukhi, Gujarati, and Bengali further evolved from it.

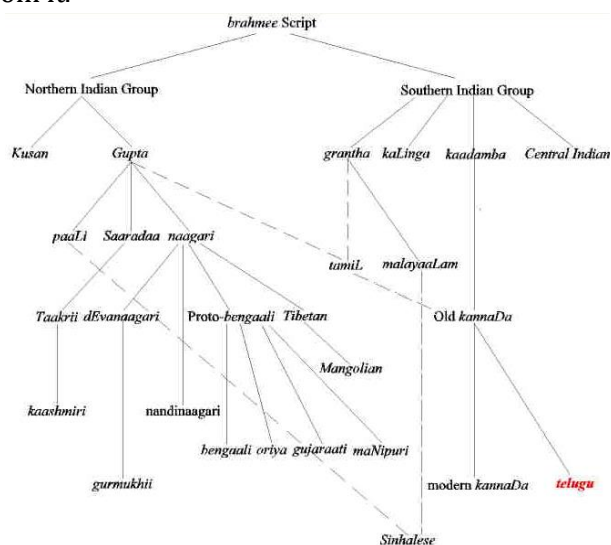


Figure 1. Descent of various Indic scripts from the Braahmee

Some scripts like Tibetan script derived the writing system of the Lepcha (Rong) and the Passepa of the Chinese Yuan dynasty (1206–1368). Brahmi also developed following different structural script in the southern regions of India especially into the Grantha alphabet, which further formed the writing systems of the Dravidian languages in the region (e.g., Tamil, Malayalam, Telugu, and Kannada). It is spoken by about 25% of the population. The development in the Indic scripts [48] also saw the evolution of the

writing systems of the Sinhalese language used in some part of Indian and Sri Lanka. Scripts for Khmer and Mon languages of Southeast Asian regions and Kavi in the region of Indonesia were also influenced by it. The ancient Cham inscriptions of Malayo-Polynesian (Austronesian) speakers from southern Vietnam also used script influenced by South Indic origin. Also, traces in the history suggests Thai writing system was derived from that of the Khmer, and the endangered systems of Buginese and Batak of Indonesia were further derived from Kavi.

Research studies [49] highlight that while many scripts survived the timeline, many became extinct or are on the verge of extinction. Those which survived were also influenced by other scripts developing parallel to them or derived themselves and transformed into new ones in specific regions. The digital recognition of scripts and OCR research is motivated by the facts revealed in the recent study on the status of endangered and extinct languages across Indian states [46].

III. STATUS OF OCR FOR INDIC SCRIPTS

Table 1 (also see Figure 2) highlights that a total of 199 Indic languages are placed in the endangered category, 82 being venerable, 63 under definitely endangered, 06 is severely endangered, 42 in critically endangered and 06 in the already extinct category. This motivates digital preservation initiatives for such scripts and forms the basis of modern research and development in computation linguistics and content creation in digital form in endangered scripts.

Last two decades have seen several attempts at digitizing texts and developing OCR for many Indic scripts. Table 2 summarizes the status of present developments of OCR for various scripts used across Indian states. It clearly suggests handwritten OCR lags behind printed text OCRs. Widely used scripts like Devanagari, Bangla, Telugu, Tamil and Urdu (Perso-Arabic) have significantly improved OCRs both for printed and handwritten texts. Nearly half of the scripts (especially those from medieval period) have negligible attempts to script

digitization or have been explored for OCR application to a little extent. It is also evident that the languages or the scripts which have a smaller share in the population have higher chances of being left behind. The primary reason for the negligence of the OCR attempts for these scripts is due to the adaptation of widely used scripts like Devanagari, Bangla and others for in lieu of the scripts for scripts that have less share in the population. Urdu and Roman English also forms a considerable section of the language used across India and other countries.

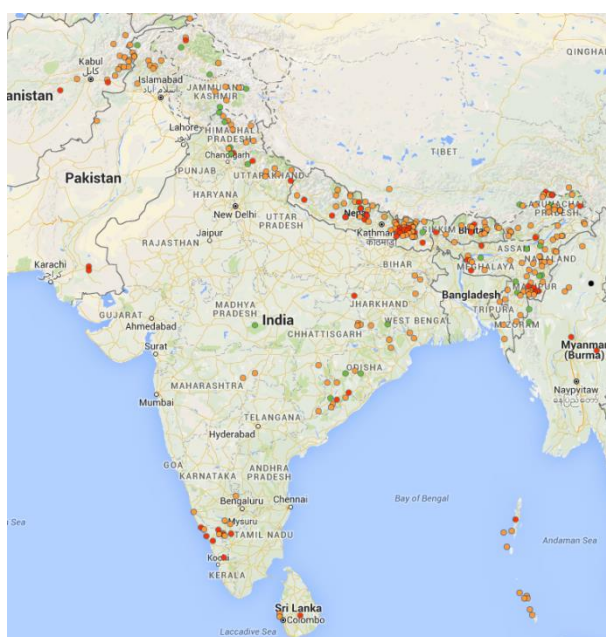


Figure2. Regionwise endangered Languages in Indian region (source: <http://www.endangeredlanguages.com/>)

Clearly, the table reveals that scripts like Modi, Kaithi, Sharada, Bodo and Brahmi needs more focused attempts for developing digitizers not only would enable preservation of contents already developed in these scripts but also help them to reclaim its place for writing texts for communication. These scripts have high literary and cultural value. Largely these scripts have been replaced officially with other widely used scripts or have no digital presence. Some languages have survived this timeline by shifting their scripts to other popular and/or simpler scripts. This includes language like Marathi which officially shifted from Modi to Devanagari script for all its documentation purpose. Punjabi uses Gurumukhi rather than using Shahmukhi, Maithili uses Devanagari rather than

Kaithi, Manipuri uses variant of Bengali rather than Meithei scripts respectively.

Separate scripts for languages like Bodo, Dogri, Konkani, and Sindhi has not been in existence. They rather use forms of scripts used by more popular languages. Bodo uses Devanagari, Dogri uses Devanagari and Perso-Arabic scripts. Konkani is written in forms of Devanagari, Perso-Arabic, Kannada and Malayalam scripts. Similarly Kashmiri used Perso-Arabic based writing styles. Occasional use of Devanagari is also seen for Kashmiri but Sharada script is rarely practiced.

Devanagari script for all its documentation purpose.

A. Reasons for endangered status of Indic Scripts

Some of the noted reasons for death of scripts are listed below

- Absolute number with knowledge of writing
- Intergenerational script transmission
- Shift of language and script use
- Lack of availability of materials for language and script education
- Governmental and institutional language policies including official status and use
- Lack of amount and quality of documentation and digital preservation techniques

Some historical document were written in related scripts but not in wide use today. For example, present Marathi language uses Devanagari Script while historical writings followed related by a structurally different form of a script called Modi especially in the medieval period. A Perso-Arabic Nastaliq script was used to write Urdu. Although present contents are produced digitally and available online and offline, the digital access to materials of historical and cultural heritage in Indic scripts is limited. Developing OCR for contemporary and historical Indic scripts, face several challenges which are different from the issues faced with Latin and Oriental scripts. Heritage materials available in archives used many scripts that are not in use today and were written on media including palm leaves. Digitizing such documents poses great challenges due to their deteriorating storage condition and availability of the only handful of domain experts.

Table 2: Linguistic Composition of India (as per Census report 2011) and Status of OCR

S. No	Languages	Scripts	% of Total population	Printed		Handwritten	
				Method	Accuracy	Method	Accuracy
1	Hindi	Devanagari	41.03	Structural feature based tree classifier [1]	96%	Modified quadratic classifier on directional features[4]	94.24%
2	Bengali	Bangla	8.11	Feed forward neural network based recognition on freeman chain code[5]	98% (isolated character) 96% (continuous characters)	k-curvature feature extraction, multi-layer perceptron classifiers[9]	96.25%
3	Telugu	Telugu	7.19	Principal Component Analysis followed by support vector classification [11]	96.56%	Convolutional neural networks Principal Component Analysis, Support vector machines[10]	92.26% (consonants) 92% (vowel modifier)
4	Marathi	Devanagari,	6.99	Minimum distance classifier technique[14]	90%	modified wavelet based kernels[12]	97.94%
		*Modi,		-data not available	-	structure similarity approach for isolated symbols[6]	
5	Tamil	Tamil	5.91	Nearest neighborhood classifier on Geometric moments and discrete cosine transforms[17]	98%	fuzzy approach [18]	94% (character) 95% (numerals) 76-94%
6	Urdu	Perso- Arabic	5.01	Feed Forward Back Propagation neural network [19]	100%	Tree based dictionary search [22] Principal component analysis,[23]	96.2%
		Devanagari		##	-	##	-
7	Gujara ti	Gujarati	4.48	K-Nearest Neighbour[24]	97.78%	Neural Network[25]	88.6%
8	Kannad a	Kannada	3.69	RBFN using Haar wavelets and structural features[26]	99.1%	SVM classifier[27]	95%
9	Malayala m	Malayalam	3.21	Singular value decomposition [30]	97%	HMM and SVM [29]	97.97% and 95.24%
10	Oriya	Oriya	3.21	stroke and run-number based features[31]	96.3%	Curvature feature and principal component analysis[32]	94.60%
11	Punjabi	Gurumukhi	2.83	Hybrid classification using binary decision trees and nearest neighbours [34]	97%	SVM classifier used with RBF kernel[33]	
		Shahmukhi		\$Data Not Available		\$Data Not Available	
12	Assame se	Assamese	1.28	Mathematical morphology [36]	100% (only numerals)	Feed Forward Backpropagation ANN[35]	96%
13	Maithil i	Devanagari	1.18	@Devanagari printed OCR techniques can be used	\$Data Not Available	@Devanagari printed HOCR techniques can be used	\$Data Not Available
		Kaithi		\$Data Not Available	\$Data Not Available	\$Data Not Available	\$Data Not Available
14	Santali	Devanagari	0.63	@Devanagari printed OCR techniques can be used	\$Data Not Available	@Devanagari printed HOCR techniques can be used	\$Data Not Available
		Oriya, Bengali		\$Data Not Available	\$Data Not Available	\$Data Not Available	\$Data Not Available
15	Kashmiri	Perso-Arab ic	0.54	@urdu printed OCR techniques can be used	\$Data Not Available	@Devanagari printed OCR techniques can be used	\$Data Not Available
		Sharada,		\$Data Not Available	\$Data Not Available	\$Data Not Available	\$Data Not Available
		Devanagari		@Devanagari printed OCR techniques can be used	\$Data Not Available	@Devanagari HOCR techniques can be used	\$Data Not Available
16	Nepal i	Devanagari	0.28	Line, word and character Fragmenter with Tessera ct [45]	97%	DTW algorithm[44]	89%
17	Sindi	Devanagari	0.25	@Devanagari printed OCR techniques can be used	\$Data Not Available	@Devanagari printed OCR techniques can be used	\$Data Not Available
		Arabic		Support vector Machines, fuzzy neural network, HMM [42]	97%	@Perso-Arab ic HOCR techniques can be used	\$Data Not Available
		Gurumukhi		@Gurumukhi printed OCR techniques can be used	\$Data Not Available	@Gurumukhi printed HOCR techniques can be used	\$Data Not Available
18	Konkani	Devanagari	0.24	@Devanagari printed OCR techniques can be used	\$Data Not Available	@Devnagari HOCR techniques can be used	\$Data Not Available
		Kannada		@Kannada printed OCR techniques can be used	\$Data Not Available	@Kannada HOCR techniques can be used	\$Data Not Available
		Perso-Arab ic		@Perso-Arab ic printed OCR techniques can be used		@Perso-Arab ic printed HOCR techniques can be used	
		Malayalam		@Malayalam printed OCR techniques can be used		@Malayalam printed HOCR techniques can be used	
19	Dogri	Devanagari,	0.22	@Devanagari printed OCR techniques can be used	\$Data Not Available	@Devnagari HOCR techniques can be used	\$Data Not Available
		Perso-Arab ic		@ Perso-Arab ic printed OCR techniques can be used	\$Data Not Available	@Perso-Arab ic HOCR techniques can be used	\$Data Not Available
20	Manipu ri	Bengali	0.14	Syllable Based Model[40]	94.57	@Bengali HOCR techniques can be used	\$Data Not Available
		Meithe i Mayek		Support Vector Machines[39]	98.45%	Support Vector Machine classifier, with RBF Kernel [39]	95.16%
21	Bod o	Devanagari	0.13	\$Data Not Available		\$Data Not Available	
22	Sanskrit	Devanagari	<0.1		98%	ough Transform and Minimum Distance classifier[72] ANN[38]	94% 98%
		Brahmi		\$Data Not Available		\$Data Not Available	

IV. DEVELOPING OCR FOR INDIAN SCRIPTS

Developing OCR for Indian scripts is relatively challenging for the fact that these are structurally and semantically different in composition to their counterpart non-Indian scripts in several ways. Nearly all Indian scripts are composed of basic symbols that include consonants and modifiers. Contrary to non-Indian Scripts, Indian scripts have no issue of case sensitivity but pose challenges in digital recognition due to its substantial cursive structure. Almost all Indian scripts can be described using three horizontal zones as shown in figure 3.



Figure 3. Different zones of Devanagari text

Serval attempts have been made in last few decades and some interesting results have been obtained for recognizing scripts including Devanagari, Gurumukhi, Bangla, Tamil, Telugu, and Oriya (table 2). As compared to non-Indic scripts, the research on OCR for handwritten Indic scripts has not achieved much perfection. Indic scripts have a lot of scope for improving recognition results especially for handwritten samples and text written in different styles. Interestingly till date in general, commercially acceptable complete OCR system for handwritten text in Indian script is not available which provides motives to invest in research into OCR for Indic Scripts.

The complex nature of the scripts used for some important languages is visible from the Table 3. The text is written in different languages used across India exhibit several similar and dissimilar properties which allow them to use same or variants of the script with them. Devanagari and Urdu (Perso-Arabic based) are such scripts which are used for more than one languages.

Table I: Example of Texts in Indian Scripts

Sample Text	Language and Script
सूरज की किरण	Hindi (Devanagari)
সূর্যের দল	Bangali (Bengali)
சூரிய கதிர்கள்	Tamil (Tamil)

സൂര്യ കിരണങ്ങൾ	Malyalam (Malyalam)
ಸೂರ್ಯನ ಕಿರಣಗಳು	Kannada (Kannada)
સૂર્યની કિરણો	Gujrati (Gujrati)
ਸੂਰਜ ਦੇ ਐਕਸਰੇ	Punjabi (Gurumukhi)
సూర్యుడు యొక్క కిరణాలు	Telugu (Telugu)
सूर्य किरण	Marathi (Devanagari)
सूरज किरण	Medieval Marathi (Modi)
سورج کی کرن	Urdu
नमो नमः, नमस्कारः	Sanskrit

Observations show that some scrips are characterized by 'shirorekha' (viz. Hindi, Marathi, Punjabi, and Bengali). Nearly all Indic scripts have a cursive style of writing characters with some relatively more cursive in nature and some having self-occluding cursive structures like in medieval Modi script and most Dravidian languages.

The long history of civilization and political evolution has led to the development of numerous scripts along the timeline from ancient to medieval to the modern time. Some of them have evolved and modified its form and can be seen in use even today. Some scripts such as forms of classic Brahmi, cursive Kaithi, and Modi had been used in past but not used today. They have great potential to contribute to a knowledge base and literary assets. Such scripts lack availability of OCR system at large and have been less explored. Developing OCR for such scripts can contribute to the digital restoration of archived documents from history and help to revive it and possibly bring them back to use.

V. RECOGNITION OF MEDIEVAL AND ANCIENT SCRIPTS

Recent decades have developed the keen interest of researchers and institutions like CDAC, TIFR, ISI, IIT, and TDIL in development of OCR especially focused towards Indic scripts. Many research groups at universities in India have come up with recognition system for Indian scripts that are in use today. Some convincing results with the increased level of acceptance have been achieved for printed text for

scripts including Devanagari, Gurumukhi, Bangla, Tamil, Telugu, and Oriya. Standardization of OCR for handwritten characters is been taken and many researchers are on to it. While some scripts are neglected for the purpose of OCR due to their restricted population base, others have seen some developments carried out by researchers and research institutions in last few decades. Table 2 quantifies efforts in recognizing printed and handwritten texts for scripts. Important to note is that results are based on best results using specific method rather than overall recognition rates as the data samples and level of complexities are different. Most of the reported work is at the isolated character level while very few are recognizing continuous character texts [5] of varying complexities. This makes possibilities for improving the OCR in specific dimension for different scripts.

B. Recognition of Devanagari Script

The Devanagari script has the largest share in writing Indic script with its use for Sanskrit, Hindi, Marathi and Pali languages. It has been substantially researched for the purpose of developing OCR as early as in the 1970s with the work of R. M. K. Sinha et al. [59]. They suggested syntactic pattern analysis technique for the recognizing handwritten and machine printed Devanagari characters. U. Pal et al. in their paper [1] obtained up to 96% accuracy for characters and demonstrated how zonal information and shape characteristics can be used to derive basic, modified and compound characters for the classification convenience. They exhibited how characters are recognized by a structural feature based tree classifier is suitable for modified characters and also use of hybrid approach for compound characters.

Efforts by Veena Bansal et al. in their paper [2] at ICDAR 2001 demonstrated the accuracy of 93% for a complete OCR for printed Devanagari texts. Accepted attempts in recognizing handwritten Devanagari characters include work by U. Pal et al [4] with an accuracy of 94.24% for offline handwritten characters primarily based on features like directional information with the modified quadratic classifier. Mahesh Jangid [60] in his paper reported 95.89% accuracy for handwritten characters for a dataset of 12240 samples. Arora et al. [3]

experimented with Combining Multiple Feature Extraction techniques like intersection, shadow feature, chain code histogram and straight line fitting features for handwritten characters and classified characters using four multi-layer perceptrons (MLP) based classifier with an accuracy of 92.80%.

Segmentation of characters is challenging for Indian scripts especially handwritten ones. V. M. Ladwani et al. [62] have experimented and proposed segmentation in hierarchical order for handwritten Devanagari words. It uses morphological image processing and neighborhood tracing algorithms for zoning and achieved accuracy around 57%. Other efforts in OCR for Devanagari include Recent effort in recognizing printed Devanagari characters include work by Ankush A.Mohod et al. [61] using Neuro-Fuzzy Integrated System for classification with instances of accuracy up to 100% of test images against databases characters. Although OCRs for printed texts in Devanagari scripts has reached to acceptable accuracy handwritten text still lacks good accuracy due to variation in writing style and different ways of writings.

C. Recognition of Gurmukhi Script

Gurumukhi is a script for writing the Punjabi language. Pioneer efforts for developing OCR for Gurumukhi has been initiated by Lehal and his group [34, 63]. They used segmentation process which breaks a word into sub-characters and used a hybrid classification scheme that uses binary decision trees and nearest neighbors to classify sub-characters and combine them to form Gurmukhi characters. The impressive recognition rate of 97% is achieved with speed of 175 characters/second. Jindal, M.K et al in [64] highlighted that performance of standard OCRs that are designed for fine printed documents declines when tested on degraded document samples. It described structural features useful for degraded printed Gurumukhi documents.

Sukhpreet Singh et al. described use of Gabor Filter based method and SVM classifier with RBF kernel[33] for feature extraction for handwritten characters and achieved 94.29% accuracy. The database was limited to 200 samples each of basic 35 Gurumukhi characters for different writers.

D. Recognition of Bangla Script

Some initial recognized efforts for developing OCRs for printed Bangla (Bengali) script document was reported by Mahmud et al. in their paper [5] which described the use of Feedforward neural network based recognition on a feature such as Freeman chain code. It achieved 98% accuracy on isolated characters while 96% accuracy was reported for continuous characters. OCRs for Handwritten characters are attempted by other researchers using different techniques [6, 7, 8, 9] and obtained a different level of accuracy. Purkait et al. [9] obtained accuracy of 96.25% and described novel Morphological features and k-curvature feature extraction technique to recognize handwritten scripts. The feature space was trained using multi-layer perceptron (MLP) classifiers and then fused them using modified 'Naive'-Bayes combination to obtain the increases recognition accuracy. U. Bhattacharya et al. in their paper [65] reported accuracy up to 92.14% for handwritten characters. It demonstrated results by downsampling the histogram feature by applying a Gaussian filter and Multilayer MLP trained by backpropagation (BP) algorithm for classification.

E. Recognition of Tamil Script

Tamil is one of the popular Dravidian languages spoken in and around Tamil Nadu with its own 'Tamil' script. Several attempts for printed and handwritten script recognition is already undertaken for printed and handwritten scripts as well. Recognition rate as high as 98% for printed text characters is obtained by K. G. Aparna et al. in their paper [17] which describes the use of Nearest neighbourhood classifier on Geometric moments and discrete cosine transforms.

Seethalakshmi R et al. [66] described the translation of printed Tamil characters into Unicode characters. It used vertical and horizontal histogram projections for segmentation and passed features to Support Vector Machine (SVM) for supervised learning. Suresh, R.M. et al. in their paper [18] presented results of recognition of printed and handwritten Tamil characters using the fuzzy approach with accuracy about 76% for handwritten characters.

F. Recognition of Malayalam Script

Anil R et al. have worked for recognizing Printed Malayalam characters and presented their paper [30] with an accuracy up to 97%. It was based on Singular Value Decomposition and used Euclidean distance measure for determining the nearest machining class of characters among all. Another attempt for printed Malayalam script character is by Bindu Philip et al. [67] which identified distinctive structural features of machine-printed text lines and recognized characters using SVM technique with accuracy between 90.22% and 95.31%. Interesting results were produced by Primekumar K.P et al. in their report [29] which demonstrated the comparative performance of character recognition using HMM and SVM on on-line Malayalam handwritten characters. The accuracy of 97.97% for SVM and 95.24% for HMM was obtained on a test sample of 1279 characters.

G. Recognition of Kannada Script

One of the finest results in printed text recognition in Kannada script recognition is by B. Vijaykumar et al. [26] accuracy as high as 99.1% which describes the use of RBF and subspace approach for text recognition. R Sanjeev Kunte et al. experimented with vowels and consonants [68] for printed Kannada text. They used Hu's invariant moments and Zernike moments to extract features of characters followed by neural classification and achieved a recognition rate of 96.8%. Initiatives in recognizing handwritten Kannada scripts includes work by Rajput et al. [27] who exercised shape descriptors based handwritten character recognition engine for recognition. It included use of Fourier descriptor and chain codes followed by SVM classifier for obtaining final results with performance of 98.45% and 93.92%, for numeral characters and vowels respectively.

H. Recognition of Oriya Script

Character recognition of Oriya has also been looked into by researchers in the recent decade, B. B. Chaudhuri B.B et al. [31] presented work on automatic recognition of printed Oriya script which reported an accuracy of 96.3%. It has used techniques of skew correction, line segmentation, zone detection, word and character segmentation and applied a mixture of stroke and run-number

based features, together with features obtained from the concept of a water reservoir. U Pal et al. [32] reported method of off-Line Oriya handwritten character recognition that used curvature feature and principal component analysis to achieve recognition rate up to 94.60%.

I. Recognition of Gujarati Script

Gujarati contrary to Devanagari is a class of cursive script without 'shirorekaha'. With people speaking Gujarati spread across India has a large base of printed contents in circulation including magazines and newspapers. It has a large base of speakers in the state of Gujrat, Maharashtra, and parts of Rajasthan. Research attempts at recognizing printed and handwritten characters of Gujarati script are in progress. Brijesh Sojitra et al. used Neural Network based implementations for recognition of Gujrati characters and obtained recognition rate of 97.78%. They used 10 data samples from 5 different fonts with 50 samples for each character. Another Neural network based experiment by A R Vasant et al. [25] was done on handwritten digits of Gujarati and demonstrated accuracy around 88.70% for digits written in various sizes.

J. Recognition of Urdu Script

Urdu is one of the widely used scripts used by Indian states since Mugal rulings. It is Nastaliq script which is written from right to left. Attempt to recognize printed texts of Urdu script is reported by in [19, 21] based on ligatured based approach with feed forward back propagation neural networks and showed accuracy up to 98.3%. H. Malik et al. experimented and reported character recognition based on segmentation of printed Urdu Scripts Using Structural Features and obtained accuracy up to 99.4%. Handwritten attempts for Urdu script recognition report accuracy of 96.3% using tree-based dictionary search [22] and Principal Component Analysis [23] respectively. Sohail Abdul et al. described [70] use of Finite State Model for Urdu Nastalique OCR, Khalil Khan et al. proposed Method for Urdu Language Text Search in Image-Based Urdu Text [71] with varying level of accuracy from 78% to 100% for 2 to 5 character ligatures respectively.

K. Recognition of Assamese Script

Assamese scripts is used in the region around present Assam in north east India. Aydin M. et al. presented work [35] on Assamese character recognition using feedforward Artificial Neural Networks with a reported accuracy of 96% on handwritten characters. Medhi K. et al. described recognition of handwritten numerals of Assamese script using mathematical morphology reported accuracy of 100% on their data set.

L. Recognition of Language Using adapted Scripts

Many Indian languages either do not have separate scripts to write or have moved to other widely and well-established scripts. This keeps them live while speaking in native language writing in script obtained from other languages. For example, Marathi, Sanskrit, Maithili use Devanagari script rather using their original Modi Kaithi and Brahmi scripts respectively. Similarly, other lesser-known scripts including Santali, Sindhi, Kashmiri, Konkani, Dogri, Nepali, and Bodo also used an occasionally Devanagari script to write texts. Manipuri has started using Bengali and its variants to write texts rather using original Meithei script. Konkani has traces of using Malayalam, Kannada and even Perso-Arabic script for its writing while Sindhi also used back in time Perso-Arabic based writing styles. Sindhi also find some writings in Gurumukhi and Devnagari. Kashmiri moved from Sharada to Devanagari and Perso-Arabic writing scripts. The study shows that attempts in developing OCRs for native scripts will provide chances of reviving the lost and endangered scripts and bring them back to use.

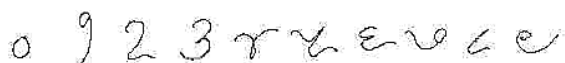
VI. RECOGNITION OF MEDIEVAL AND ANCIENT SCRIPTS

While OCR systems have evolved during past few decades and have been successfully used for widely used scripts, the work on scripts of medieval and ancient India is limited. OCRs for such scripts would help to revive them. Scripts like Modi for Marathi, Shahmukhi for Punjabi, Sharada for Kashmiri, Kaithi for Maithili and Meithei Mayek for Manipuri are some among them. The analysis of the status and

possibilities of availability of OCR for such scripts is discussed next.

M. Modi Script

Modi is a medieval script used primarily for writing the Marathi language till 1950's in the regions around Maharashtra including some parts of central and southern part of India. Historical forms of Modi script are Bahamanikalin, Chitnisi, Peshvekalin, and Anglakalin. Peshvekalin and Anglakalin forms have been considered most beautiful and stable style of writing. In 1950's officially Devanagari has replaced Modi. Modi is rich in its ability to convey language aspects but is considered to be highly complex. It has evolved and changed over last few decades. Theoretical analysis of handwritten Modi Script [50] was presented by D. N. Beseekar et al. describing the issues in recognizing isolated characters of Modi script, comparison with Devanagari and Roman scripts, and also highlighted the importance of structural features in recognition of Modi script. D. N. Beseekar et al. in [51,52]] highlighted used of morphological approach and chain code for recognizing Modi Numerals. Dataset used by D. N Beseekar et al. is presented below



Work by A. S. Ramteke et al. in [53] describe the use of relative spatial covariance and structure similarity approach for recognizing isolated characters of Modi script. The work in [51, 52, 53] is limited to recognizing numerals and selected characters of Modi script written by a single person. The work on multi-character texts and contiguous connected words and sentences are not reported. Fonts like 'Hemadree' are now available for typing Modi text. Unicode for Modi script is also developed [54, 55] recently which motivates and open avenues for digitization and character recognition of printed text.

आपण काय करत आहेत | आपल्या परीक्षा कशी होती |

सूर्य किरण येत आहेत | तो अतिशय चांगले वाटते |

Complete OCR for Modi symbols and text is still a challenging task in the digitization of the scripts as most samples available are in handwritten forms. With Unicode font available now, more intensive efforts in digitizing handwritten and printed Modi script is justified in current time.

N. Kaithi Script

Kaithi script is primarily used for writing Maithili and other forms of Hindi languages including Awadhi, Bhojpuri, Magahi and also to an extent Urdu till a few decades back. Very few sections of the society use it due to various constraints outlined earlier in this paper. It is considered to have developed from the Gupta script around the 16th century. Kaithi is an independent writing system profoundly used throughout northern India in the regions covered by Bihar and Uttar Pradesh. Traces are also found in Mauritius and Trinidad populated by people of this region. It is essentially an independent scribal and printing tradition showing features of Devanagari, Gujarati, and other major North Indic scripts. It was an alternate script to Devanagari, Persian, and other scripts commonly used in northern India for a long time. More formal use of Kaithi script is seen in Bihar region as administrative scripts around an 18th, 19th and 20th century. Kaithi was adopted by also Western missionaries in the region and translated Christian literature in this local script. Last traces of its use in Bihar dates around the 1960s when Devanagari replaced it below

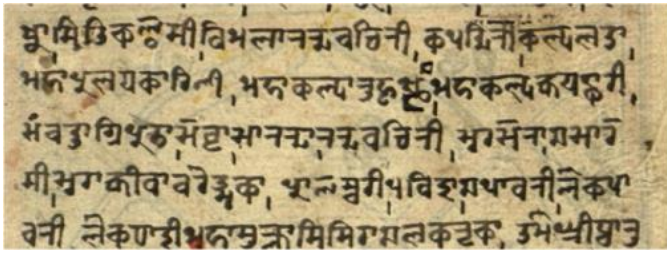
तु का कान ह्व | तहान शहिन कैसन रहे |

सूना के कगिस आवात वा | वृहान अय्युषा भागत वा |

Nearly no recognized work is reported for digitization of Kaithi. Hence research for digital recognition for preservation, representation, and reproduction of written and printed Kaithi documents in digital media is highly recommended.

O. Sharada Script

It is a script to write Kashmiri and Sanskrit, primarily developed from Brahmi script around mid of 8th century and used in the regions of present Kashmir and West Pakistan. Remains of its use are traced with only handful Brahmins of the region who use it for writing and calculating astrological and ritual formulations. Some traces of its use are also found in northwestern regions of India, including Punjab, Himachal Pradesh and in some parts of Central Asia. A sample of text written using Sharada script is shown below.



Today, only a small group of Brahmins continue to use the Sharada alphabets for writing and calculating astrological and ritual formulations. Efforts in developing digital OCRs for Sharada will boost the digital restoration and archival projects.

P. Shahmukhi Script

Shahmukhi script is the Perso-Arabic style of writing Punjabi which was primarily used by Muslims in Punjab, parts of Kashmir and Punjab province of Pakistan. It was profoundly used during the 10th and 11th century after the Mughal conquest and establishments in Indian Subcontinent. Shahmukhi was used until the mid-20th century when Gurumukhi replaced Shahmukhi post-Independence and is used to a little extent. This script has little work towards digitization and OCR. Work by G. S Lehal et al. [56] had worked for Shahmukhi to Gurmukhi Transliteration System and reported an average accuracy around 91%. The sample data used by them is as under

اس گل وچ جنوں امیں بہتے پنجابیوں نوں ویکھدے ہاں تاں پرنسپل تیجا سنگھ دے لیکھ وچ بیاتیاں گنیاں کوڑیاں سچائیاں ہور وی شدت نال محسوس بندیاں ہین۔ امیں دیس نوں پیار کرن دا دعویٰ کردے ہاں پر اپنے صوبے نوں وساری بیٹھے ہاں۔ اس دا سبہ توں وڈا ثبوت ایہہ ہے کہ بھارت دے لگ بھگ بہتے صوبے اپنے اپنے ستھاپنا دوس بڑے اتشہاء تے جذبے نال مناوندے ہین۔ اپنی زبان، اپنے سہیجاچار، اپنے پچھوکر تے اپنے ورثے تے مان کردے ہین۔ اپنی قومی پچھان تے مان کردے ہین۔ پر سادا بایا آدم ہی نرالا ہے۔ سرکاراں توں لے کے عام لوکاں تک پنجابی صوبے دے بنن دن بارے پوری طرحاں اویسلے ہی رہندے ہین۔

The method was based on a bi-gram language model for mapping Shamukhi to Gurmukhi equivalents Following is hand written excerpts from Gurbani text written (from right to left) using Sahnukhi script. Literary archives from the region have an abundance of handwritten documents which needs to be restored, digitized and transliterated for later generations. More investigations and efforts in developing OCRs is advised.

Q. Meithei Mayek script

Meithei Mayek script is used for writing Meithei (or Meetei/Manipuri) language and was considered the official language of the Manipur region of India, until the 18th century, before being replaced by the Bengali script for writing Manipuri. A large section of the region feels the need to revive the

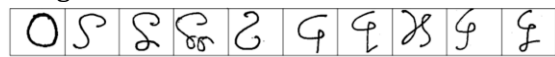
original script which motivates researchers for developing OCRs for Meithei Mayek script.

Keynsham Angphun Maring et al. in their paper [57] have attempted a recognition of handwritten and printed Manipuri numerals based on support vector machine. In this paper, they focused only on numerals and did not attempt to recognize characters and forms of mixed text. The dataset included the following handwritten and printed numerals.



Transliteration between Bengali script and Meithei Mayek scripts was done by Thoudam Doren Singh in his paper [40] for Web-Based Manipuri News Corpus. It was based on monosyllabic characteristics of Manipuri language.

Recognition of Handwritten Numerals of Manipuri Script was reported by Chandan Jyoti Kumar et al. [38] using techniques based on support vector Machine with RBF Kernel and obtained 95.16% accuracy. Sample used by them included the following



Complete OCR for all characters of the script is still not developed with acceptable accuracy. With recognition results limited to numerals, need for OCR is strongly felt.

VII. REQUIREMENTS FOR OCR FOCUSED ON MEDIVAL INDIAN SCRIPTS

Character recognition is a kind of pattern recognition problems. Indian scripts especially Medieval and Ancient scripts exhibit complex and variable characteristics which make development of OCRs more challenging. Contrary to modern OCRs for Roman and Chinese scripts OCRs for Indian scripts needs customized pre and post-processing of the input samples for digital recognition. Presence and absence of 'shirorekha', cursive style of writing, the direction of strokes, connectedness and grouping of

symbols are important features for recognition of characters. Skewness is also one another important characteristics, which have to be specially handled differently among different script.

Present Indian scripts either have their origin in medieval period or itself are medieval scripts. Therefore, there is a huge collection of literary contents created in them and needs to be recognized, converted/transliterated to digital forms. Strategy for Indic script OCRs can also follow either statistical, syntactic/ structural or hybrid approach [58] based on the feature set used for the purpose of recognition.

The statistical approach can be used when patterns of script characters are represented as a vector of fixed length containing a list of numeric features. Since Indic scripts have the large possibility of exhibiting natural variants of a character, many samples are used for collecting statistics to be used during the training phase. The statistics based approach is very useful for identifying cases of the unknown character. Statistical features include distribution of points, geometrical moments, and crossing information.

As most Indic scripts are largely cursive in nature and often contains loop structures, structural classification techniques can use them with custom decision rules for character classification. Indic scripts can be easily characterized and differentiated using structural features including character strokes, presence, and absence of character holes, number of endpoints and loops or other character properties such as concavities. Presence or absence of one or more of these features help to recognize the natural variants of a character and also discriminate between similar-looking characters such samples like ॐ ॐ ॐ or ॐ ॐ ॐ or ॐ ॐ ॐ in Modi, Gujarati and Devanagari scripts respectively.

For document images retrieved from historical archives, the statistical information is more useful as they are tolerant to noise in sample space over which the representative training has been performed. It can give better results than structural descriptions. On the other hand, for cases of script samples with variation due writing style are better handled using

structural descriptions. Since document scan is neither perfect nor the writing styles, a hybrid approach can combine the two at appropriate stages for better results. An OCR for any Indic script with focus on medieval and ancient characteristics must also follow the similar stages to that of any standard OCR but will be highly dependent of specifics of the script under recognition. A standard OCR encompasses stages including digitization and preprocessing, segmentation, feature extraction and classification and finally the character recognition.

OCRs for Indic scripts in its digitization stage scan converts input images (intensity maps) containing printed or handwritten text into machine-readable digitally defined streams for symbols and patterns of the script. Later stages would map them once recognized to digitally defined characters or else generate character definitions wherever no digital fonts available. Preprocessing stage for scripts especially those from medieval period is important as most documents are available in historical archives with improper storage conditions. The preprocessing stage comprises of enhancing activities to make them suitable for recognition purpose. Noise removal techniques, bridging gaps, and broken structures skew detection and correction, and skeletonization are important activities. Special attention must be given during noise removal for characters formed with 'dots' which would otherwise be treated as noise and culled away. Skew correction is yet another important task due to improper feeds during scanning or a characteristics of writer induced phenomenon. OCRs would need to identify multiple characters within words, words within lines and lines within the image document. The task becomes more difficult for the character made from more than one independent structures. Segmenting characters for scripts with 'shirorekha' and structural fusion is a difficult task. A segmentation algorithm must be developed to decide the best segmentation point both for handwritten and printed texts. Incorrect segmentation can lead to the incorrect recognition. Some important generic stages of the digital character recognition system for Indic Scripts is describe next.

VIII. REFERENCES

- [1] U. Pal and B.B. Chaudhuri, "Printed Devnagari script OCR system", Knowledge Based Computer Systems : Research and Applications, Ed. K. S. R. Anjaneyulu, M. Sasikumar and S. Ramani, Narosa Publishing House, 1996, pp. 359-371
- [2] Veena Bansal, M. K. Sinha, "A Complete OCR for Printed Hindi Text in Devanagari Script", ICDAR, 2001, 2013 12th International Conference on Document Analysis and Recognition, 2013 12th International Conference on Document Analysis and Recognition 2001, pp. 800-804, doi:10.1109/ICDAR.2001.953898
- [3] Arora, S.; Bhattacharjee, D.; Nasipuri, M.; Basu, D.K.; Kundu, M., "Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition," Industrial and Information Systems, 2008. ICIIS 2008. IEEE Region 10 and the Third international Conference on , vol., no., pp.1,6, 8-10 Dec. 2008 doi: 10.1109/ICIINFS.2008.4798415
- [4] Pal, U.; Sharma, N.; Wakabayashi, T.; Kimura, F., "Off-Line Handwritten Character Recognition of Devnagari Script," Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on , vol.1, no., pp.496,500, 23-26 Sept. 2007 doi: 10.1109/ICDAR.2007.4378759
- [5] Mahmud, J.U.; Raihan, M.F.; Rahman, C.M., "A complete OCR system for continuous Bengali characters," TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region , vol.4, no., pp.1372,1376 Vol.4, 15-17 Oct. 2003 doi: 10.1109/TENCON.2003.1273141
- [6] Mandal, S.; Sur, S.; Dan, A.; Bhowmick, P., "Handwritten Bangla character recognition in machine-printed forms using gradient information and Haar wavelet," Image Information Processing (ICIIP), 2011 International Conference on , vol., no., pp.1,6, 3-5 Nov. 2011 doi: 10.1109/ICIIP.2011.6108911
- [7] T. K. Das, A. Datta, S. K. Parui, and B. B. Chaudhuri , Recognition Of Handprinted Bangla Numerals Using Neural Network Models, U. Bhattacharya, Advances in Soft Computing - AFSS 2002, Springer Verlag, Lecture Notes on Artificial Intelligence, Eds. N.R. Pal and M. Sugeno, LNAI 2275, 2002, pp. 228-235.
- [8] Bhattacharya, N.; Pal, U., "Stroke Segmentation and Recognition from Bangla Online Handwritten Text," Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on , vol., no., pp.740,745, 18-20 Sept. 2012 doi: 10.1109/ICFHR.2012.275
- [9] Purkait, P.; Chanda, B., "Off-line Recognition of Hand-Written Bengali Numerals Using Morphological Features," Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on , vol., no., pp.363,368, 16-18 Nov. 2010 doi: 10.1109/ICFHR.2010.63
- [10] Soman, Soumya T; Nandigam, Ashakranthi; Chakravarthy, V.Srinivasa, "An efficient multiclassifier system based on convolutional neural network for offline handwritten Telugu character recognition," Communications (NCC), 2013 National Conference on , vol., no., pp.1,5, 15-17 Feb. 2013 doi: 10.1109/NCC.2013.6488008
- [11] Jawahar, C.V.; Pavan Kumar, M.N.S.S.K.; Kiran, S S Ravi, "A bilingual OCR for Hindi-Telugu documents and its applications," Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on , vol., no., pp.408,412 vol.1, 3-6 Aug. 2003 doi: 10.1109/ICDAR.2003.1227699
- [12] Shelke, S.; Apte, S., "A novel multistage classification and Wavelet based kernel generation for handwritten Marathi compound character recognition," Communications and Signal Processing (ICCSP), 2011 International Conference on , vol., no., pp.193,197, 10-12 Feb. 2011 doi: 10.1109/ICCSP.2011.5739299
- [13] Urmila Shinde, Vanita Mane, Rajashree Shedje, "Marathi Character Recognition Using Ant Miner Algorithm", International Journal of Advanced Computational Engineering and Networking, ISSN: 2320-2106, Volume-2, Issue-10, Oct.-2014, pp.101-107
- [14] Kiran R Dahake, S R Suralkar and S P Ramteke. Article: Optical Character Recognition for Marathi Text Newsprint. International Journal of Computer Applications 62(16):11-15, January 2013
- [15] S. M. Mali, "Moment And Density Based Handwritten Marathi Numeral Recognition",

- Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166 Vol. 3 No.5 Oct-Nov 2012, pp.707-712
- [16] A S Ramteke, G S Katkar, "Recognition of Off-line Modi Script : A Structure Similarity Approach", International Journal of ICT and Management, February 2013 Vol- I Issue -I, ISSN No. 2026-6839, pp.12-15
- [17] K. G. Aparna, A. G. Ramakrishnan, "A Complete Tamil Optical Character Recognition System", Document Analysis Systems, Lecture Notes in Computer Science Volume 2423, 2002, Aug 2002, pp. 53-57
- [18] Suresh, R.M.; Ganesan, L., "Recognition of printed and handwritten Tamil characters using fuzzy approach," Computational Intelligence and Multimedia Applications, 2005. Sixth International Conference on , vol., no., pp.291,296, 16-18 Aug. 2005 doi: 10.1109/ICCIMA.2005.47
- [19] S. A. Husain and S. H. Amin. "A multi-tier holistic approach for Urdu Nastaliq recognition". In IEEE Int. Multi-topic Conference, Karachi, Pakistan, Dec. 2002.
- [20] Malik, H.; Fahiem, M.A., "Segmentation of Printed Urdu Scripts Using Structural Features Visualisation", 2009. VIZ '09. Second International Conference in doi: 10.1109/VIZ.2009.12 Publication Year: 2009 , PP: 191 – 195
- [21] Inam Shamsher, Zaheer Ahmad, Jehanzeb Khan. Title: "OCR for Printed Urdu Script Using Feed Forward Neural Network". Conference name: "Proceedings of world academy of science, engineering and technology", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:1, No:10, 2007 volume 23, August 2007, pp 508-513.
- [22] Malik, S.; Khan, S.A., "Urdu online handwriting recognition," Emerging Technologies, 2005. Proceedings of the IEEE Symposium on , vol., no., pp.27,31, 18-18 Sept. 2005 doi: 10.1109/ICET.2005.1558849
- [23] Khalil Khan, Rehan Ullah, Nasir Ahmad Khan and Khwaja Naveed. Article: Urdu Character Recognition using Principal Component Analysis. International Journal of Computer Applications 60(11):1-4, December 2012
- [24] Brijesh Sojitra, Vishnukumar Dhakad, "Neural Network In Character Recognition Of Gujarati Script" Journal Of Information, Knowledge And Research In Computer Engineering, ISSN: 0975-6760, Volume – 02, Issue – 02, Pp.269-272
- [25] Vasant, A.R.; Vasant, S.R.; Kulkarni, G.R., "Performance Evaluation of Different Image Sizes for Recognizing Offline Handwritten Gujarati Digits Using Neural Network Approach," Communication Systems and Network Technologies (CSNT), 2012 International Conference on , vol., no., pp.270,273, 11-13 May 2012 doi: 10.1109/CSNT.2012.66
- [26] Vijaykumar, B.; Ramakrishnan, A.G., "Radial basis function and subspace approach for printed Kannada text recognition," Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on , vol.5, no., pp.V,321-4 vol.5, 17-21 May 2004 doi: 10.1109/ICASSP.2004.1327112
- [27] Rajput, G.G.; Horakeri, R., "Shape descriptors based handwritten character recognition engine with application to Kannada characters," Computer and Communication Technology (ICCCT), 2011 2nd International Conference on , vol., no., pp.135,141, 15-17 Sept. 2011 doi: 10.1109/ICCCT.2011.6075175
- [28] Vishwaas, M.; Arjun, M.M.; Dinesh, R., "Handwritten Kannada character recognition based on Kohonen Neural Network," Recent Advances in Computing and Software Systems (RACSS), 2012 International Conference on , vol., no., pp.91,97, 25-27 April 2012 doi: 10.1109/RACSS.2012.6212704
- [29] Primekumar, K.P.; Idiculla, S.M., "On-line Malayalam handwritten character recognition using HMM and SVM," Signal Processing Image Processing & Pattern Recognition (ICSIPR), 2013 International Conference on , vol., no., pp.322,326, 7-8 Feb. 2013 doi: 10.1109/ICSIPR.2013.6497991
- [30] Anil R, Arjun Pradeep, Midhun E M, Manjusha K, "Malayalam Character Recognition using Singular Value Decomposition", International Journal of Computer Applications, ISSN:0975 – 8887, Volume 92 – No.12, April 2014, pp-6-11.
- [31] Chaudhuri, B.B.; Pal, U.; Mitra, M., "Automatic recognition of printed Oriya script," Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on , vol., no.,

- pp.795,799, 2001
doi: 10.1109/ICDAR.2001.953897
- [32] Pal, U.; Wakabayashi, T.; Kimura, F., "A System for Off-Line Oriya Handwritten Character Recognition Using Curvature Feature," Information Technology, (ICIT 2007). 10th International Conference on, vol., no., pp.227,229, 17-20 Dec. 2007
doi: 10.1109/ICIT.2007.63
- [33] Sukhpreet Singh, Ashutosh Aggarwal, Renu Dhir, "Use of Gabor Filters for Recognition of Handwritten Gurmukhi Character", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012, ISSN: 2277 128X, pp.234-239.
- [34] G S Lehal and Chandan Singh, "A Complete OCR System For Gurmukhi Script" Proceedings SPR2002, Windsor, Canada, Lecture Notes in Computer Science, Vol. 2248, Springer- Verlag, Germany, 2002, pp. 344-352
- [35] Aydin, M.; Celik, E., "Assamese character recognition with Artificial Neural Networks," Signal Processing and Communications Applications Conference (SIU), 2013 21st, vol., no., pp.1,4, 24-26 April 2013, doi: 10.1109/SIU.2013.6531488
- [36] Medhi, K.; Kalita, S.K., "Recognition of assamese handwritten numerals using mathematical morphology," Advance Computing Conference (IACC), 2014 IEEE International, vol., no., pp.1076,1080, 21-22 Feb. 2014
doi: 10.1109/IAdCC.2014.6779475
- [37] R. Dineshkumar and J. Suganthi, "Sanskrit Character Recognition System using Neural Network", Indian Journal of Science and Technology, Vol 8(1), 65-69, January 2015, ISSN (Print) : 0974-6846, ISSN (Online) : 0974-5645, pp.65-69
- [38] Chandan Jyoti Kumar, Sanjib Kumar Kalita, "Recognition of Handwritten Numerals of Manipuri Script", International Journal of Computer Applications (0975 - 8887) Volume 84 - No.17, December 2013, pp.1-5
- [39] Romesh Laishram, Angom Umakanta Singh, N.Chandrakumar Singh, A.Suresh Singh, H.James, "Simulation and Modeling of Handwritten Meitei Mayek Digits using Neural Network Approach", Proc. of the Intl. Conf. on Advances in Electronics, Electrical and Computer Science Engineering — EEC 2012, ISBN: 978-981-07-2950-9
doi:10.3850/ 978-981-07-2950-9 769, pp.355-358
- [40] Thoudam Doren Singh, "Bidirectional Bengali Script and Meetei Mayek Transliteration of Web Based Manipuri News Corpus", Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 181-190, COLING 2012, Mumbai, December 2012, pp.181-189
- [41] D. N. Hakro, A. Z. Talib, Z. Bhatti, G. N. Mojai, "A Study Of Sindhi Related And Arabic Script Adapted Languages Recognition", Sindh University Research Journal, Vol. 46 (3) 323-334 (2014), pp.323-333
- [42] Bashir, R.; Quadri, S., "Identification of Kashmiri script in a bilingual document image," in Image Information Processing (ICIIP), 2013 IEEE Second International Conference on, vol., no., pp.575-579, 9-11 Dec. 2013, doi: 10.1109/ICIIP.2013.6707658
- [43] Santosh K.C., Cholwich Nattee, "A Comprehensive Survey On On-Line Handwriting Recognition Technology And Its Real Application To The Nepalese Natural Handwriting", Kathmandu University Journal Of Science, Engineering And Technolgy Vol. 5, No. I, January, 2009, pp 31-55
- [44] Prajwal Rupakheti, Bal Krishna Bal, "Research Report on the Nepali OCR", Madan Puraskar Pustakalaya, 2009
- [45] Fritz E. Froehlich, Allen Kent, The Froehlich/Kent Encyclopedia of Telecommunications: Volume 3, CRC Press, 31-Oct-1991
- [46] ChartsBin statistics collector team 2011, Number of Endangered Languages by Country, ChartsBin.com, August, 2015, <<http://chartsbin.com/view/1339>>.
- [47] Uday Narayan Singh, 'Minor and Minority Languages in India', in Report by G.N. Devy Sub-Group, Protecting Non-Scheduled Languages, 11th five year plan proposal, Ministry of Human Resource Development, 2006.
- [48] Indic writing systems. 2015. Encyclopædia Britannica Online. Retrieved 27 August, 2015, <http://www.britannica.com/topic/Indic-writing-systems>
- [49] UNESCO 2011, Number of endangered languages by country, 2011, United Nations Educational, Scientific and Cultural Organisation Institute for Statistics, Paris, France, 2011, [<http://www.unesco.org/culture/languages-atlas/index.php?hl=en&page=atlasmap>].

- [50] D. N. Besekar, R. J. Ramteke, "Study for Theoretical Analysis of Handwritten MODI Script – A Recognition Perspective", International Journal of Computer Applications (0975 – 8887) Volume 64– No.3, February 2013, pp-45-49.
- [51] D. N. Besekar, "Recognition Of Numerals Of Modi Script Using Morphological Approach", Shodhsamiksha Aur Mulyankan, ISSN- 0974-2832 RNI-RAJBIL 2009/29954.Vol.III, Issue-27, pp-63-66
- [52] D. N. Besekar, R. J. Ramteke, "A Chain Code Approach for Recognizing Modi Script Numerals", Indian Journal of Applied Research, Vol-I, Issue-3, Dec 2011, ISSN-2249-555X, pp-222-225
- [53] A. S. Ramteke, G S Katkar, "Recognition of Off-line Modi Script : A Structure Similarity Approach", International Journal of ICT and Management, February 2013 Vol- I Issue –I, ISSN No. 2026-6839, pp-12-15
- [54] Pandey, Anshuman. "Proposal to Encode the Modi Script in ISO/IEC 10646". Unicode Consortium. 2011, <http://www.unicode.org/L2/L2011/11212r2-n4034-modi.pdf>
- [55] Unicode Standard 8.0, Copyright © 1991-2015 Unicode. < <http://unicode.org/charts/PDF/U11600.pdf> >
- [56] Tejinder Singh Saini and Gurpreet Singh Lehal, "Shahmukhi to Gurmukhi Transliteration System: A Corpus based Approach", Advances in Natural Language Processing and Applications Research in Computing Science 33, 2008, pp. 151-162
- [57] Kansham Angphun Maring, Dr. Renu Dhir, "Recognition Of Cheising Iyek/Eeyek-Manipuri Digits Using Support Vector Machines" IJCSIT, Vol. 1, Issue 2 (April 2014), e-ISSN: 1694-2329 | p-ISSN: 1694-2345, pp-1-6.
- [58] B.Anuradha Srinivas, Arun Agarwal, And C.Raghavendra Rao, "An Overview Of Ocr Research In Indian Scripts", Ijcses International Journal Of Computer Sciences And Engineering Systems, Vol.2, No.2, April 2008, pp-141-153
- [59] R.M.K.Sinha and H.N.Mahabala (1979), Machine recognition of Devnagari script, IEEE Trans. on Systems, Man and Cybernetics, SMC-9, 435-441.
- [60] Mahesh Jangid, "Devanagari Isolated Character Recognition by using Statistical features", International Journal on Computer Science and Engineering, ISSN : 0975-3397 Vol. 3 No. 6 June 2011, pp-2400-2407
- [61] Ankush A.Mohod, Nilesh N.Kasat, "Optical Character Recognition of Printed Text in Devanagari Using Neuro - Fuzzy Integrated System", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-3, Issue-7, December 2013
- [62] Ladwani, V.M.; Malik, L., "Novel Approach to Segmentation of Handwritten Devnagari Word," in *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on* , vol., no., pp.219-224, 19-21 Nov. 2010
doi: 10.1109/ICETET.2010.143
- [63] Lehal, G.S.; Singh, C., "A Gurmukhi script recognition system," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on* , vol.2, no., pp.557-560 vol.2, 2000
doi: 10.1109/ICPR.2000.906135
- [64] Jindal, M.K.; Sharma, R.K.; Lehal, G.S., "Structural Features for Recognizing Degraded Printed Gurmukhi Script," in *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on* , vol., no., pp.668-673, 7-9 April 2008
doi: 10.1109/ITNG.2008.223
- [65] U. Bhattacharya, M. Shridhar, and S.K. Parui, "On Recognition of Handwritten Bangla Characters", ICVGIP 2006, LNCS 4338, pp. 817–828
- [66] Seethalakshmi R.†, Sreeranjani T.R.†, Balachandar T., "Optical Character Recognition for printed Tamil text using Unicode", Journal of Zhejiang University SCIENCE, ISSN 1009-3095, 2005, pp. 297-1305.
- [67] Bindu Philip and R. D. Sudhaker Samuel, "An Efficient OCR for Printed Malayalam Text using Novel Segmentation Algorithm and SVM Classifiers", International Journal of Recent Trends in Engineering, Issue. 1, Vol. 1, May 2009, pp. 178-182
- [68] R Sanjeev Kunte, R D Sudhaker Samuel, "A simple and efficient optical character recognition system for basic symbols in printed Kannada text", *Sadhana* Vol. 32, Part 5, October 2007, pp. 521–533
- [69] Apurva A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network," Pattern Recognition, Volume 43, Issue 7, July 2010, ISSN 0031-3203, pp. 2582-2589,
- [70] Sohail Abdul, Sattar Shams-ul, Haque Mahmood Khan Pathan, "A Finite State Model for Urdu

- Nastalique Optical Character Recognition”, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.9, September 2009, pp. 116-122
- [71] Khalil Khan, Muhammad Siddique , Muhammad Aamir & Rehanullah Khan, “An Efficient Method for Urdu Language Text Search in Image Based Urdu Text”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3, March 2012 ISSN (Online): pp. 1694-0814.
- [72] Sujata S. Magare, Ratnadeep R. Deshmukh, “Offline Handwritten Sanskrit Character Recognition Using Hough Transform and Euclidean Distance”, International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 10 No. 2 Oct. 2014, pp. 295-302
- [73] National Mission for Manuscripts, [<http://namami.org>]
- [74] Technology Development for Indian Languages (TDIL), Department of Information Technology (DIT), Govt. India, [<http://www.tdil.mit.gov.in>]
- [75] Centre for Development of Advanced Computing, Multilingual Computing & Heritage Computing, [http://www.cdac.in/index.aspx?id=milingual_heritage]
- [76] People’s Linguistic Survey of India (PLSI), <http://peopleslinguisticsurvey.org/>
- [77] Matthias Brenzinger, “Language Diversity Endangered”, ISBN, Walter de Gruyter GmbH & Co, 2007, 978-3-11-017054
- [78] Pal, U.; Chaudhuri, B.B., "OCR in Bangla: an Indo-Bangladeshi language," in *Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision & Image Processing, Proceedings of the 12th IAPR International Conference on* , vol.2, no., pp.269-273 vol.2, 9-13 Oct 1994 doi: 10.1109/ICPR.1994.576917
- [79] Dunn, C.E.; Wang, P.S.P., "Character segmentation techniques for handwritten text-a survey," in *Pattern Recognition, 1992. Vol.II. Conference B: Pattern Recognition Methodology and Systems, Proceedings, 11th IAPR International Conference on* , vol., no., pp.577-580, 30 Aug-3 Sep 1992 doi: 10.1109/ICPR.1992.201844
- [80] Sharma, D.V.; Lehal, G.S., "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* , vol.2, no., pp.1022-1025, doi: 10.1109/ICPR.2006.258
- [81] Chaudhuri, B.B. ; CVPR Unit, Indian Stat. Inst., Kolkata, India ; Bera, S., Handwritten Text Line Identification in Indian Scripts, Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference 2009, ISBN- 978-1-4244-4500-4, 10.1109/ICDAR.2009.69
- [82] Sundaram, S.; Ramakrishan, A.G., "Lexicon-Free, Novel Segmentation of Online Handwritten Indic Words," in Document Analysis and Recognition (ICDAR), 2011 International Conference on , vol., no., pp.1175-1179, 18-21 Sept. 2011 doi: 10.1109/ICDAR.2011.237

Cite this Article

Bhavesh Kataria, Dr. Harikrishna B. Jethva, "Review of Advances in Digital Recognition of Indian Language Manuscripts ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 4 Issue 1, pp. 1302-1318, January-February 2018. Available at doi : <https://doi.org/10.32628/IJSRSET1841215>

Journal :URL : <https://ijsrset.com/IJSRSET1841215>