Study for Theoretical Analysis of Handwritten MODI Script – A Recognition Perspective

D. N. Besekar
Dept. of Computer science and Information technology
Shri. Shivaji College of Science, Akola (M.S.) India

R. J. Ramteke, PhD.
Head, Dept. of Information Technology
School of Computer science
North Maharashtra University, Jalgaon (M.S.) India,

ABSTRACT
During the last three decades, numerous handwriting character recognition systems have been proposed.Many of them presented their limitation when the handwriting character is cursive type and it has some deformation. However this type of cursive character is easily recognized by the human being. The ambient study had been performed on foregion language like Arebic, Chinese, Japanees etc. and South Asian Languages such as Sanskrit, Hindi, Marathi, Nepali etc. are written using Devnagari script. OCR of Modi script language is still not available as it is cursive type and old language i.e.Shivkalin and Peshwekalin. The challenge of recognition of character is even high in Handwritten Domain, due to the varying writing style of each individual.

This paper explain the theoretical analysis of MODI script according to recognition point of view. some issues involved with character recognition of MODI script are also recorded allong with differentiation between Devnagari characters and MODI characters. some complicacies involved with the recognition of handwritten MODI characters using structural features are also reported. With this basic classification of MODI characters on simple basic rules are also reported

Keywords
Hand-written character recognition, MODI script, Roman script, structural features, topological features.

1. INTRODUCTION
Offline handwritten recognition has been popular field of research for many years. it is still remains an open problem.The challenging nature of handwritten recognition and segmentation has attracted the attention of researchers from industry and academic peoples. recogniition, segmentation and classification of MODI script is a difficult task because the MODI handwritten characters are naturally both cursive and unconstrained. simillarly, high similarity between character and distorted and broken characters. Hence extreme variation is observed in the collected samples. An optical character recognition works in several stages such as scanning, preprocessing, feature extraction, classification and post-processing. Each stage has large number of optional techniques which have been used in differrent recognition systems. Selecting a best technique for a particular application is a daunting task. Onehasto exhaustively study the literature, implement them and observe their performance. As far as feature extraction stage is concerned, Govindan et al[1] classified the various features in three categaries i.e. statistical, structural and global transforms and series expansion. Each categori has its pros and cons in terms of comptutional speed, computational complicacy and accuracy. In structural based approach, a character is recognized on the basis of structural primitives from which it is build up and these primitives are also known as character strokes. one advantage of using structural features for character recognition is that they are invarient to character distortion and writing styles to a large extant and the feature extraction process for these techniques is fast [2]. Some authors [3,4,5,6] have used a combination of statistical and structural features for handwritten numerals or character recognition. Skeletal graph based approaches for handwritten recognition has been used by Li et al [7] and Xue et al [8]. An efficient method for extracting curvature features based on curve fitting is proposed by Miura etal and polygonal approximation of a thinned character is proposed by Rocha et al [9].

In thhis paper, an analysis of alphabet set of MODI script is being made in respect of hand-written character recognition. The various issues involved with character recognition of MODI script are reported in next section. Next section also explain about comparision between MODI script and Devnagari script. The various complicacies involved with the recognition of hand-written MODI characters in general are reported. Some points of differentiation between MODI and Roman are also covered.

2. MODI SCRIPT
Modi is one of the scripts used to write the Marathi language, which is the primary language spoken in the state of Maharashtra in western India. There are several theories about the origin of this script. One of them claims that it was developed by Hemadpant (or Hemadri Pandit) during the reign of Mahadev Yadav and Ramdev Yadav (1260–1309). Others claim that it was brought by Hemandpant from Sri Lanka. It is a popular notion that only Marathi is written in Modi. Other languages that have also been found to have been written in Modi are Urdu, Kannada, Gujarati, Hindi and Tamil.

The Modi alphabet was invented during the 17th century to write the Marathi language of Maharashtra. It is a variant of the Devanāgarī alphabet. The Modi alphabet was used until 1950 when it was replaced by the Devanāgarī alphabet. it is syllabic in nature and alphabets are classified into vowels and consonants and numerals. Notable features are that each letter has an inherent vowel (a). Other vowels are indicated using a variety of diacritics which appear above, below, in front of or after the main letter. Some vowels are indicated by modifying the consonant letter itself. Marathi, an Indo-Aryan language spoken by about 71 million people mainly in the Indian state of Maharashtra.

Fig.1 : MODI Vowels and diacritics

Fig.2: MODI Consonants

3. REVIEW OF MODI SCRIPT RECOGNITION

Modi script is being used by Shivkalin and PeshweKalin kingdoms and upto the begining of English rular. with this, other languages was also available such as sanskrit, Nepali etc. the work done for the recognition of MODI Script is still not reported in literature research. But some papers are also available for Devnagari script, not in Modi script. it is found that still not good ICR available for Devnagari script. much more work is done on similar languages like urdu, snihali, tamil, telgu etc. MODI script is kept away due to its cursive and complex type nature and available only in old document and on stone in a carving form.

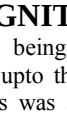
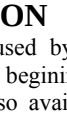
4. STRUCTURAL FEATURES OF MODI SCRIPT CHARACTERS

The conventional handwritten characters recognition system is based on the feature extraction from the character shape such as a line inclination, relative position of such line, the wide of the different parts from the line etc [10,11,12]. These character features can be used for non-cursive handwritten characters or printed characters. however, these features are not enough for the cursive character recognition. Structural features are based on the geometrical and topological properties of a character. A character is composed of number of components in the form of strokes. These strokes may be line, arcs, curves, loops etc. each character component is called as a character primitive. These are extracted either from a skeleton or a contour of a character image. The character is recognized on the basis of number of such primitives, the kind of primitives and the relation between these primitives in some order. Some commonly used topological and geometrical features in character recognition are end points, T-points, cross points, extrema(top, bottom, left, right), direction of strokes, loops, convex and concave arcs,straight lines(Horizontal, vertical, slanted), directional points, bend points, arbitrary shape etc. Some authors have used binary feature vector obtained by using topological or geomotrical features. In this case, a character image is divided into number of regions and the presence or absence of a particular feature in a region is detected and represented using binary features. The variations in handwritten versions of a character due to different writers are so high that this representation gives a lot of variation in computed features. in some studies, structural features have been used to perform preliminary classification where whole alphabet set is divided into groups or subgroups [13,14] based on topological features.

5. COMPARISION BETWEEN DEVANAGARI AND MODI SCRIPT

Some points of differentiation between these script are given as:

1] some letters of modi script are just like as devnagari script having only little bit changes. These are as follows.

Devnagari	ग	घ	भ	ष	त
MODI					
	ga	gha	bha	sha	ta
Devnagari	ण	श	ड	छ	ळ
MODI					
	na	sa	da	cha	ala

ga, gha, bha, sha, cha, ala shows veryless difference. ta has curved to the left. Na has solid note to the right. sha also has curved shape to the left. Da has all parts in curved shape.

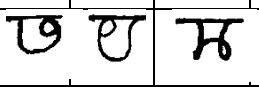
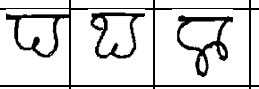
2] Some another group of letters have following differences:

Devnagari	ह	इ / ई	ध	द	ख
MODI					
	ha	e	dha	da	kha
Devnagari	र	ट	ठ	ज	न
MODI					
	ra	ta	tha	ja	na

here, if ‘ha’ have given upper half semi circle towards right then it became ‘e’. if ‘ha’ have given small circle solid or hollow then it became ‘dha’. if ‘ha’ have given upper half semi circle towards left then it became ‘kha’.

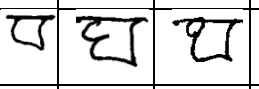
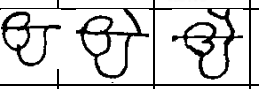
In the second row it has been found that if vertical bar turned towards left then it became ‘ra’. if the same vertical bar turned to semi cricle then it became ‘ta’. if this semi circle contains dot ‘.’ then it became ‘tha’. in this way if ‘ja’s top left circle become solid then it became ‘na’.

3] here we consider another group of 10 Modi letters:

Devnagari	स	य	म	फ	झ
MODI					
	ca	**ya**	**ma**	**fa**	**za**
Devnagari	प	थ	क	ल	च
MODI					
	pa	**tha**	**ka**	**la**	**cha**

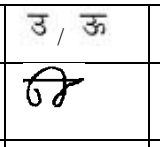
'ca' is just like of devnagari numeral 'seven' . if we take mirror image or circle drawn to the left top of this then it became 'ya'. If semi circle drawn to the top left of 'ma' then it became 'fa'. If semi circle added to the top left of 'pa' then it became 'tha'. 'la' is just like of numeral 'four' written in devnagari.

4] Now consider another group of 10 Modi letters:

Devnagari	व	ब	आ	ओ	औ
MODI					
	va	ba	aa	o	ou
Devnagari	अ	ए	ऐ	अं	अः
MODI					
	a	ae	aee	anm	anh

if we added knot to the middle left top side of 'va' then it became 'ba'. If diagonal bar added to the top of 'aa' then it became 'o'. If two diagonal bar added to the top of 'aa' then it became 'ou'.
similarly, If diagonal bar added to the top of 'a' then it became 'ae'. If two diagonal bar added to the top of 'a' then it became 'aee'. If 'dot' added to the top of 'a' then it became 'anm'. If two 'dot' added to the right side of 'a' then it became 'anh'.

5] Now consider the remaining letters of modi script:

Devnagari	उ / ऊ	क्ष	ज्ञ	ढ
MODI				
	u	ksh	dnh	dha

'u' & 'ksh' have all curvers. 'dha' contains two different letters. 'dha' can be written as forming knot to the right and for external side.

6. DIFFICULTIES IN EXTRACTING STRUCTURAL FEATURES

A set of features to be used for recognition are also selected keeping in view the classification(s) being used for classification purpose. For example, the statistical features are suitable for classification with classifiers such as ANN and SVM. The comparisons of some features using MLP and SVM on Devanagari character set is given in [15]. The structural features are not suitable for hand-written character recognition and some reasons behind this are as follows:

1] Extracting structural features from MODI handwritten characters images are very difficult:

a) Due to complex structure of some of its alphabet letters.

b) The various strokes existing in a character may touch with each other (mingle) due to hasty writing. Some examples are given in figure 3.

Fig 3**:** Some Modi characters with touching strokes.

c) The chances of presence of broken stroke primitives are large as its various letters are built up using multiple strokes. Some examples are given in Figure 4.

Fig 4: Multi-stroke character with broken primitives.

d) There is blurring in some parts of a character image. Extracting skeleton of such characters may loss a lot of information. The situation is illustrated in Fig. 5, where some parts of the characters (enclosed inside circle) are completely lost.

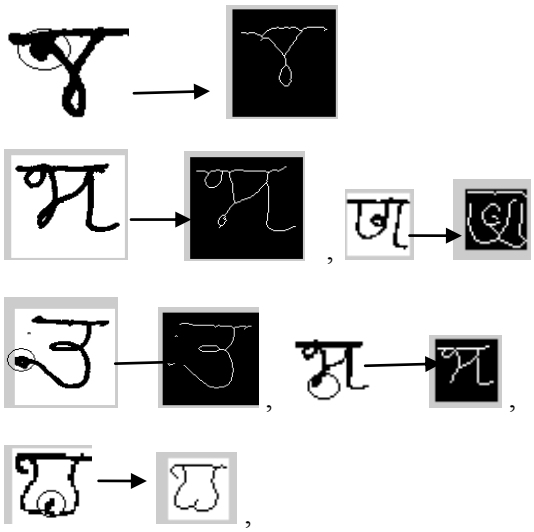


Fig 5**:** Blurred part of some MODI characters (upper row) and loss of strikes of these characters due to skeleton (lower row).

2] In noisy situation the structural features do not work. The noise not only create problem in extracting structural features but also perturbs the topological structure of a character image.

3] MODI script is written in cursive type; hence it creates extra branches in letters, which is shown in following figure 6. Structural features are extracted from the skeleton-zed image of a character. The problem is further complicated as spurious branches and clusters are formed in thinning process as given in Figure 5. Such unpredictable behavior regarding formation of extra branches is common in skeleton-zing algorithms [16]. The spurious branches and clusters of small size are ease to remove but bigger size poses problem.

Fig 6**:** Extra Branches produced due to cursive writing.

4] The variations in handwritten characters due to different writing are also high that the structural representation gives high variations in computed features.

5] In structural approach non-metric measures are used for comparison and the chances of optimal results are low.

The above said factors not only pose problems in extracting and estimating stroke primitives but also cause problems even for extracting the topological features which are very useful for performing primary classification.

7. COMPLEXITIES OCCURS IN CHARACTER RECOGNITION

In order to know the performance of a feature extraction method on some scripts, the classifier under consideration is trained and tested on large dataset pertaining to that script. If number of samples is small, the machine may not learn adequately. There is a lot of impact of size of training data set on the performance of a classifier. Until we have any sample set available for training and testing the performance of a recognition system for a script, to carry out research for that script is indeed impossible. MODI character recognition related work is not available at all in literature. The recognition of MODI handwritten script is difficult as compared to many other language handwritten scripts. Some reasons are as:

1) All characters of this script are cursive in nature that gives a lot of variation in individual writing of a character.
2) Very complex structure of some of its member classes such as given in Figure 7.

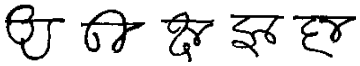


Fig7: Some MODI characters having complex structure.

3) Large within class variations due to different handwritten of different individuals as it is given in Figure 8.

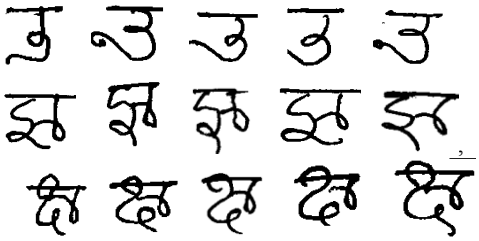


Fig 8: Various within class cariations of abovecharacters.

4) The synoymoous strcture of some of its member classes. The various characters such as ‘ja’ and ‘na’, ‘da’ and ‘ha’ , ‘kha’ and ‘tha’, ‘pa’, ‘ra’ and ’ta’ are almost similar in writing. Such characters, as shown in Figure 9, are very difficult to distinguish.

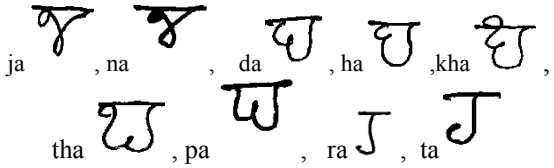


Fig 9: Some classes of MODI script having almost similar structure.

5) Pressence of large number of classes, which is shown in the Figure 4.

8. COMPARISION BETWEEN MODI AND ROMAN SCRIPT

Through out the world Roman script has been adopted whereas MODI script was used in some part of South Asian region only. The widespread use of Roman script is obvious as it is a script used for international communication purpose. The structure of both these scripts is different. some points of differentiation between these scripts are given as:

1)The alphabet set of Roman consists of lower case and upper case letters whereas no such characters exists in case of MODI. The half characters are formed by eliminating some strokes from full size (basic) characters. In addition to this, MODI alphabet set also consists of some compound characters which are formed by combining two characters. The number of letters in MODI alphabet set.is large ascompared to Roman alphabet set.

2)The Roman alphabet letters are made up of vertical, horizontal, slant and curved strokes. The use of slant or horizontal strokes is less as compared to curved or vertical strokes in roman letters. All such strokes are mostly used in MODI letters.

3)Although the use of curved strokes is large in Roman yet its letters do not exhibit complex structure. Apart from compound alphabet letters, even the structure of some basic alphabet letters of MODI are quite complex.

4)The basic, half and compound alphabet letters of MODI contains a horizontal line at their top called as headline. A

MODI word is formed by connecting its letter with each other through a headline. No such headline exist in case of Roman alphabet letters. The letters are connected each other, to form a word, through a connecting stroke in hand-written text.

5)Some alphabet letters have common characteristics in MODI script. These characteristics are presence or absence of a side bar, presence or absence of a middle bar and a single stroke or two strokes touching with headline. On the basis of these characteristics, it is easy to divide a MODI set into subset. No such characteristics exist in case of Roman alphabet. However, Roman characters may be sub-categorized based on their location and space covered in a word.

9. CONCLUSION

A lot of work hasbeen done for the development of character recognition system for the various languages of the world but good ICR of majority languages are still not available. Among the various Indian scripts, Like Devnagari, MODI is one of the script. The main reasion behind negligible work done for MODI hand-written character recognition may be the complexity of the script, oldest language and non-availability of a MODI hand-written character database. The recognition of hand-written script of a language is more complicated as compared to hand-printed or machine-printed. Also, the recogniton of MODI hand-written script is difficult as compared to many other language hand-written script due to the cursive nature of most of its class members, complex structure of some of its class members, large within class variations due to different writing habits of different individuals, and synonymous structure of some of its member classes. Also it is very difficult to extract geometrical and topological features from MODI hand-written characters.

10. REFERENCES

[1] Govindan v.k, Shivaprasad A.P, "Character Recognition – A Review", Pattern recognition,vol.23, No. 7, 1990.

[2] Khorsheed M.S, "Off-line Arabic Character Recognition – A Review", Pattern Analysis and Application, vol. 5, pp. 31-45, 2002.

[3] Baird H.S, "Feature Identificationfor Hybrid Structural/Statistical Pattern Classification", Computer Vision, Graphics and image Processing, vol. 42, pp. 318-333, 1988.

[4] Foggia P, Sansone F, Vento M., "Combining Statistical and Structural Approaches for Handwritten Character Description",Image and Vision Computing, vol. 17, pp. 701-711, 1999.

[5] Cai J. and Z-Q Liu, "Integration of Statistical and Structural Information for unconstrained handwritten Numeral Recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 21, pp. 263-270, 1999.

[6] Heutte L, Paquet T, Moreau J, Lecourtier Y and Olivier C, "A Structural / Statistical Features Based Vector for Handwritten Character recognition",Pattern Recognition Letter, vol. 19, pp. 629-641, 1998.

[7] Li X, Oh W, Hong J and Gao W, "Recognizing Compnents of Handwritten Characters by Attributed Relational Graphs with Stable Features", Proceeding of International conf. Document Analysis and Recognition, Ulm, Germany, pp. 616-620, 1997.

[8] Xue H and Govindaraju V, "Building Skeletal Graphs for Structural Feature Extraction on Handwriting Images", Proceeding on 6th International Conf. on Document Analysis and Recognition, Seattle, WA, USA, pp. 96-100, 2001.

[9] Rocha J. and Puvlidis T., "A Shape Analysis Model with Applications to a Character Recognition System", IEE Trans. on Pattern Analysis and Machine Intelligence, vol. 16,, No. 4, pp. 393-405, 1994.

[10] Namboodiri A.M., Jain A.K., "Online Handwritten script recognition", IEEE trans. on PAMI vol. 26, No.1, pp 124-130, 2004.

[11] Palmondon R, Srihari S.N, "Online and offline handwritten recognition: acomprehensive survey", IEEE Trans, PAMI, vol22, No.1, pp 63-84, 2000.

[12] Nsakatani Y, Sasaki D, LiguniY, Maeda H, " online recognition of handwritten Hiragana characters based upon complex autoregressive model", Trans on PAMI,vol.21, No.1, pp 73-76, 1999.

[13] Lehal G.S and Chandan singh, "A Gurumukhi script Recognition system", International conf. on Pattern Recognition, Barcelona, Spain, vol.2, pp 557-560, 2000.

[14] B.B. Chaudhari and U.Pal, "A Complete printed Bangla OCR system", Pattern Recognition, vol.31,No.5, pp 531-549, 1998.

[15] Satish Kumar, "Performance Comparison of features on Devnagari Hand-printed Dataset", International Journal of Recent Trendsin Engineering, Vol.1, No.2, pp. 33- 37, 2009.

[16] R.C. Gonzalez and R.E. Woods, "Digital Image Processing", 2002, Pearson Education.