Online Telugu Hand written Character Recognition: A Review

Goda Srinivasa Rao, Dr.Rajeswara Rao Ramisetty

Research Scholar, Professor & HOD, Department of Computer Science & Engineering, JNTUA, Ananthapuramu, Andhra Pradesh, JNTUCEV, Vijayanagaram, JNTUK-Kakinada

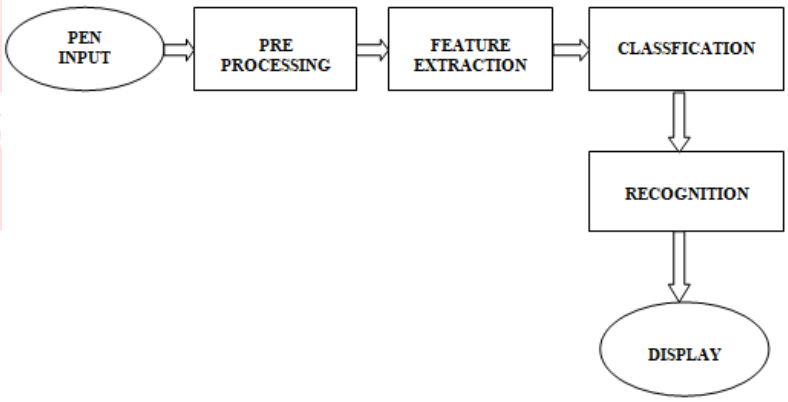
Abstract- This paper presents on different online character recognition systems for telugu language. Based on the existing research literature, recognition accuracies of different character recognition systems have been studied in languages like Devanagari, Tamil and Assamese on numerical datasets. The databases used are HP Tablet PC and UNIPEN, which confines with particular datasets. Based on the classifiers and features, the recognition accuracies will be depends.

Keywords— discrete Fourier transform (DFT), discrete cosine transform (DCT), wavelet transform, Hidden Markov Models (HMMs) and Support Vector Machine (SVM).

1 Introduction

Handwriting recognition can be categorized into online and offline based on data acquisition. The data written on an electronic digitizer using an electronic pen or stylus is captured in online handwriting recognition, where as in offline, the character to be recognized is captured by the scanner. With the advancements in technology, online handwritten input is gaining popularity due to its convenient and effective mode of communication between man and machine. The different features that have been used for extraction are global features (discrete Fourier transform (DFT), discrete cosine transform (DCT), wavelet transform), local features((x; y) coordinates, nth derivative, curvature and angular features), stroke, position, shape, directional string, VPP-HPP, ZONAL DCT,CCH AND PIXEL LEVEL FEATURE observed in the earlier works. Hidden Markov Models (HMMs) and Support Vector Machine (SVM) are the most commonly used classifiers.
The basic steps involved in character recognition are:

- Input capturing
- Pre processing
- Feature extraction and
- Classification



2 Steps Involved In Recognition

2.1. Input

To capture handwritten characters or brush strokes of a users input device called digital pen is used. The data written on electronic surface, which is in analog form is converted to digital form and utilized in several applications.

2.2. Pre Processing

To retain relevant information from the input data, pre processing techniques are used. Pre processing usually consists of normalization, sampling, smoothing, interpolation and resampling.

2.3. Normalization

By applying various normalization techniques on the pattern, in general the recognition rate is high by normalizing character with respect to the height and width along with the starting point.

2.4. Smoothing

The movement of pen may cause jitter in handwriting, which results in noise. Smoothing can be applied to remove this type of noise.
Interpolation:
Interpolation is the prerequisite for applying Re-sampling. Interpolation generates missing points, usually with the constraint that distance cannot be more than a certain threshold.

2.5. Resampling :

When the character has multiple, each stroke is resampled to preserve its ratio with the character by normalizing the character to a fixed number of points.

2.6. Feature Extraction:

Feature extraction is extracting the relevant features from the pattern , which are informative and non-redundant. These features extracted should maximize inter-class similarity and minimize intra-class similarity.

2.7. Classification:

In this step, test features are tested with several models to identify the recognized model .thus by identifying the characters that the features represent. The most widely used classifiers for character recognition are SVM and HMM. Even in languages they were used together for better accuracy of character recognition. HMM and SVM are more useful methods because of their basic functionality like using hidden networks and mapping to high dimensional data respectively.

2.8. Unipen:

Large volumes of data are challenge in handwriting recognition system. n the year 1999, a consortium of 40 companies and institutes collected a large database of on-line handwritten samples with the name international Unipen Foundation It is a **large cost free data** having more than 5 million characters, from nearly 2200 writers. The database has been utilized by many researchers at various labs and universities in all over the world. It is also used as testing samples for many pattern recognition competitions like IWFHR.

3 TELUGU CHARACTER DATABASE

Large volumes of data is challenge in handwriting recognition system. n the year 1999, a consortium of 40 companies and institutes collected a large database of on-line handwritten samples with the name international Unipen Foundation It is a **large cost free data** having more than 5 million characters, from nearly 2200 writers. The database has been utilized by many researchers at various labs and universities in all over the world. It is also used as testing samples for many pattern recognition competitions like IWFHR. The Telugu Script Consists Of 14 Vowels, 36 Consonants, and 3 Special Characters.

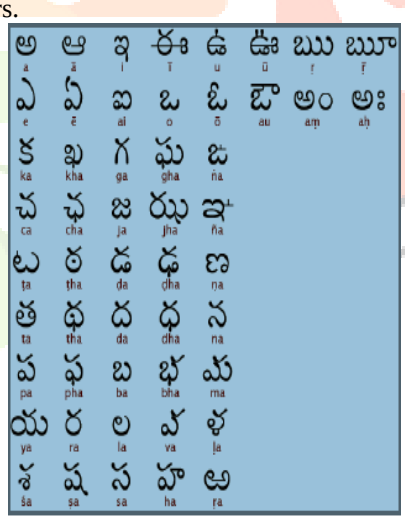


Fig. 1. Telugu character set

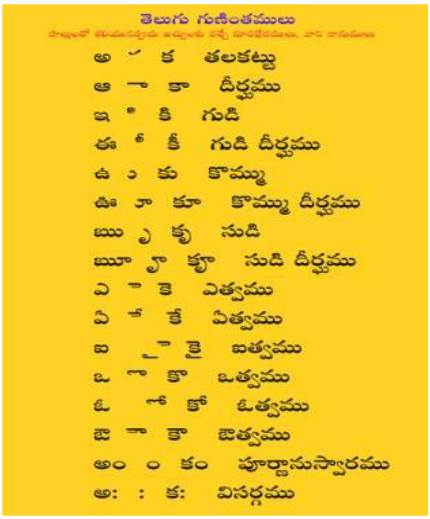


Fig. 2: Telugu Symbol dataset

4 LITERATURE SURVEY

Representation of Tamil characters and numerals is by A.G. Ramakrishna and BhargavaUrala K (2013). Here, the authors reported that the performance of cosine transform with local features is superior than Fourier transform with global features with SVM RBF classifier. As shown in the Table 1, The best result in literature on online handwritten numeral data who have obtained >95% on the UNIPEN numeral database using SVM classifier and drawbacks are DFT with LOCAL feature accuracy is greater than DCT with LOCAL feature, only UNIPEN database must be used.

Table: 1. Tamil Character Recognition Performance

System	Feature	Classsifier	Database	Recognition rate
Online	Local(x,ycord), Global feature (dft,dct)	SVM with rbf kernel	Unipen	98.2%

:

Banditasarma, KapilMehrotra, R. Krishna Naik, S. R. M. Prasanna, SwapnilBelhe, and ChitralekhaMahanta(2013) use normalized (x,y) features for recognition of Assamese numerals using HMM , SVM and its combination as classifier with the accuracy of 98.3%.The performance of the combined system through confusion matrices gave more accuracy than combined system through normalization & addition of scores as shown in Table 2.. The drawback is confusion between numeral 6 and 3 in $Comb-2$ needs further reduction.

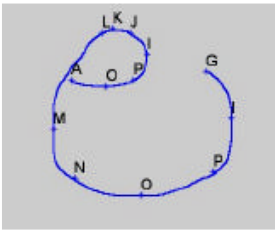
Table 2. Comparison of classifiers for character Recognition

System	Feature	Classifier	Database	Recognition rate
Online	X,Y coordinates	HMM	Hp tablet pc	96.5%
		SVM	HP Tablet PC	96.8%
		Hmm+svm(normalised+addition of scores)	HP Tablet PC	98%
		Hmm+svm(analysis of confusion matrix)	HP Tablet PC	98.3%

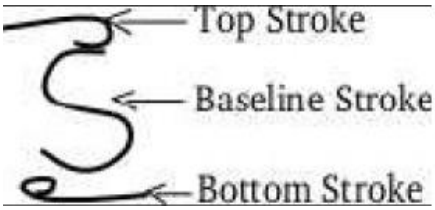
Lajish V.L and Sunil Kumar Kopparapu deals with Devanagari handwriting recognition by using storke and curvature points as features using fuzzy directional feature and directional feature as classifier and fuzzy directional feature has more accuracy than directional feature. They applied this feature for writer independent unconstrained stroke level data set. The results show that the Fuzzy Directional Features (FDF) out performed commonly used Directional Features (DF).

Table 3: Comparison of Various Databases

System	Feature	Classifier	Database	Recognition rate
Online	Stroke,curvature points	Fuzzy directional feature (fdf)	Trained data,tested data	Fdf>df
		DIREACTIONAL FEATURE(DF)	TRAINED DATA,TESTED DATA	DF<FDF



CURVATURE POINTS Fig. 3: illustration of Curvature Points

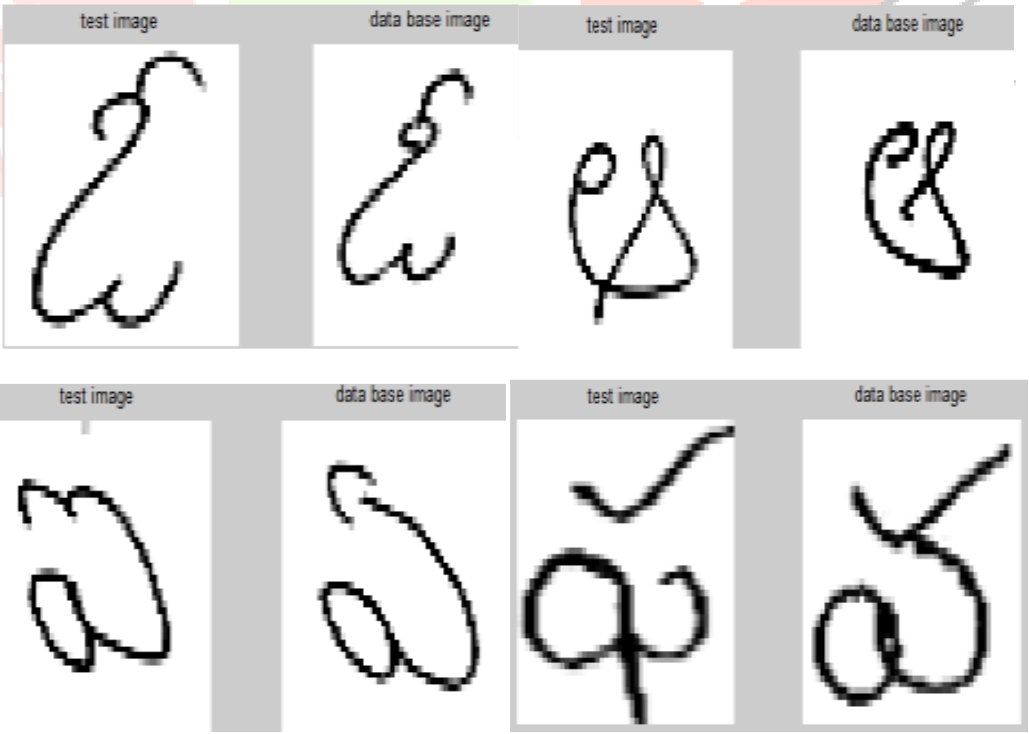


STROKE

Fig. 4: Illustration of various strokes

Minoru Mori, Seiichi Uchida, and Hitoshi Sakano(2013) deals with local and global features for character recognition and Adaboost as classifier. In this work the main advantages are that Global features give better recognition accuracy(97.3%) not only for training samples but also for test samples and global features are good discriminators power than local features. The drawback is 5-9,1-7,5-8,1-4 PATTERNS were misrecognised.

System	Feature	Classsifier	Data Base	Recognition rate
ONLINE	LOCAL (X,YCORD), GLOBAL FEATURE (RELATIVE VECTOR)	AdaBoost	UNIPEN	97.3%



S.DuttaChowdhury, U. Bhattacharya, S. K. Pauri(2013) works with shape and position as features for handwriting recognition and proposed system as classifier. They proposed a approach for online handwritten character recognition using levenshtein distance metric for UNIPEN database that had given good accuracy rates for different languages bangla numerals(98.43%) bangla characters(86.24%),devanagari(83.95%),telugu(87.10%),tamil(85%). The proposed method improves existing recognition results on the **same** databases was the shortcoming of this work

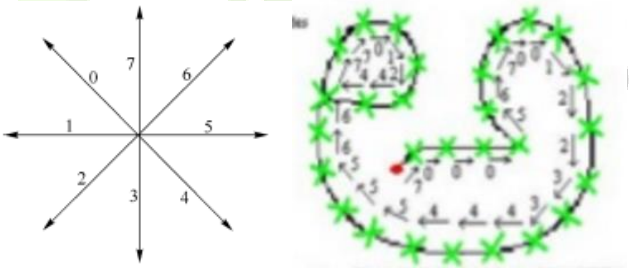
System	Feature	Classsifier	Database	Recognition rate
Online	Shape,position(levenshtein distance metric)	Proposed scheme	Unipen	Bangla numerals,basic characters-98.43%,86.24%, Telugu87.1% Tamil85%, Devanagari-83.95%,

System	Feature	Classsifier	Database	Recognition rate
Online	Shape,position(levenshtein distance metric)	Proposed scheme	Unipen	Bangla numerals,basic characters-98.43%,86.24%, Telugu87.1% Tamil85%, Devanagari-83.95%,

character differentiated by shape and position

LI Lei, ZHANG Li-liang, and SU Jing-fei(2012) deals with direction string and nearest neighbour matching as feature for character recognition and proposed scheme as classifier. The main advantage is that this scheme will transfer pulse data. The advantage is implementation of proposed scheme using Direction string using symbol strokes and distance between the direction symbols. The drawback is different threshold to be used for different symbol patterns and it could resolve the problem of symbol size variance as a whole, but it cannot handle symbol size variance in part. The recognition accuracy depends on threshold and pattern.

System	Feature	Classsifier	Database	Recognition Rate
Online	Direction string	Nearest neighbour based classifier	Numbers, Symbols	It depends on threshold and pattern



Example for direction string

G. Siva Reddy et al.,(2012) worked with (x,y) coordinates, VPP-HPP, DCT, CCH, Pixel level information as features for Assamese numeral recognition and HMM, VQ and proposed as classifiers. The recognition accuracy rate was 99.3% for combined system.The advantage is that combined proposed scheme reduced misclassification rate at the individual system level with use of feature set. The drawback was Confusion is seen to arise in patterns having similar directional information

System	Feature	Classsifier	Database	Recognition rate
Online	X,Y coordinates	HMM	HPTablet PC	96.6
Offline	Vpp-hpp,zonal dct,**cch** and pixel level feature	VQ	HPTablet PC	97.6
Combined	Min euclidean distance(offline),max likelihoods(online)	Proposed	HPTablet PC	99.3

Wei Zeng, XiangXuMeng, ChengLei Yang, Lei Haung(2006) worked with local, global and shape as features for handwritten character recognition and HMM as classifier. The recognition accuracy 92.06% can be achieved using Delaunay triangle feature, which increases stability and robustness compare with other features implemented.The drawback is Delaunay Triangle Descriptor increases the size of feature vector.

System	Feature	Classifier	Database	Recognition rate
Online	Local, Global Shape	HMM		92.06%

Deepjoy Das, Rituparna Devi, SRM Prasanna, Subhankar Ghosh, Krishna Naik(2014) worked with features (x,y) coordinates and their first and second derivatives and classifiers HMM and SVM for Assamese language. The confusion matrix shows for the first 10 strokes of Assamese language using SVM classifier got a little more accuracy than confusion percentage matrix of the 10 strokes using HMM classifier. The drawback was that a larger set of database has to be used in SVM case for improvement.

System	Feature	Classifier	Database	Recognition Rate
Online	(x,y) Coordinates and their first , second derivatives	HMM & SVM	HP Tablet PC	94% in HMM and 96% in SVM

5 Conclusion

In this article database, feature extraction and classifiers of different online handwriting characters are exposed. In online handwriting, recognition can be done using strokes, and also by using dominant points of stroke and direction between these strokes A brief introduction of popular classifiers and a few recommendation ways are given, and this can be the main content for the future work.

References

[1] Deepjoy Das, Rituparna Devi, SRM Prasanna, Subhankar Ghosh, Krishna Naik "performance comparison of online handwriting recognition system for assamese language based on hmm and svm modelling" in iccsea, sppr, vlsi, wimoa, scai, cnsa, west – 2014 pp. 87–95, 2014. © cs & it-cscp 2014.

[2] S. Dutta Chowdhury, U. Bhattacharya, S. K. Parui "Online Handwriting Recognition Using Levenshtein Distance Metric" in 2013 12th International Conference on Document Analysis and Recognition.

[3] Bandita Sarma1, Kapil Mehrotra2, R. Krishna Naik3, S. R. M. Prasanna4, Swapnil Belhe5, Chitralekha Mahanta6 "Handwritten Assamese Numeral Recognizer Using HMM & SVM Classifiers" in 2013 IEEE.

[4] G. Siva Reddy, Puspanjali Sharma, S. R. M. Prasanna, C. Mahanta and L N Sharma "Combined Online and Offline Assamese Handwritten Numeral Recognizer"in 2012 IEEE.

[5] Lajish V.L., Sunil Kumar Kopparapu "Fuzzy Directional Features for Unconstrained On-line Devanagari Handwriting Recognition" .

[6] A. G. Ramakrishnan, Bhargava Urala K "Global and Local Features for Recognition of Online Handwritten Numerals and Tamil Characters"in MOCR '13, August 24, 2013.

[7] Minoru Mori a,⇑ , Seiichi Uchida b, Hitoshi Sakano c "Global feature for online character recognition" in 2013

[8] LI Lei (_), ZHANG Li-liang, SU Jing-fei "Handwritten character recognition via direction string and nearest neighbor matching" in 2012.

[9] Wei Zenga,b,c,_, XiangXu Mengb, ChengLei Yangb, Lei Huangd "Feature extraction for online handwritten characters using Delaunay triangulation" 2006.