JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN COMPUTER ENGINEERING

# SIMPLIFYING COMPLEXITIES OF K-NN USING NEURAL NETWORK

[1]*BRIJESH SOJITRA*, [2]*RAM LAL YADAV*

[1, 2] **Department of Computer Science & Engineering,**
**Kautilya Institute of Technology & Engineering, Rajasthan Technical University, Jaipur, Rajasthan, India.**

***brijeshsojitra@gmail.com, ram.bitspilani@gmail.com***

***ABSTRACT :*** *Today large amount of data is available in printed form and large amount of old data is available in printed or in handwritten form. For processing, analyzing and searching purpose this data is not useful. Its very difficult task. To make processing, analyzing and searching easy, data should be available in electronic form. Large amount of work is done for English language. Most of Indian languages lies under Bramhi script, very less research work is carried for recognition of Indian scripts. When recognition is about Gujarati script, it becomes complex task because Guajarati script contains 34 consonants, 12 vowels, 39+ conjuncts, special symbols, Maatras. Number of models are defined by researchers all over the world for regional languages. But this models and its parameters varies with change in language and for a model performance varies with change in script.*

*Here we came up with novel approach of implementation neural network in pattern recognition. A model that learn itself and than can be use for real data recognition, a model that learn that itself. K-nn(K-nearest neighbor) is mostly used for character recognition. But there are some problems with it, consumption of CPU cycles is high and when there is large training set its performance degrades. In k-NN with increase in size of training set. It performance decrease. To resolve this problem we pre processed training data with neural network (SOM). Self Organizing Maps. We have pre processed data using Self Organizing Map and it creates clusters of nearest neighbors in pattern. And that pre processed data is used for K-nn. This preprocessing increases performance of K-nn and recognition rate.*

***KEY WORDS : Neural Network, K-nn, Self- Organizing Map, Gujarati.***

## 1. INTRODUCTION

Today technology is spanned in very aggressive manner all over the world. Cheap price of computer and internet has given acceleration to reach every corner of the world. Today most of documents systems in govt offices or everywhere is text based. Large amount of documents and literature are in printed text format or in scanned format. There is need of some efficient method that can indentify characters from the printed scanned documents. A computer system that recognize characters from a scanned image or document and can process automatically is called Optical Character Recognition (OCR) system. One of the initial technique is Template Matching technique.

Character recognition becomes more challenging when it about handwritten character recognition. In English it all about 26 alphabets, and lots of work is done for English handwritten and printed character recognition. but if we talk about Indian scripts, which contains partial characters, joint characters and lots of similar characters, at that time all these developed methods for English are not useful for Indian scripts. Brahmi scripts are far more complex than English scripts. All the Indian languages lies under brahmi script.

Many different models are designed by researchers focusing on particular problem, for particular script. For example, the simplest one is Template Matching, K- nearest neighbor, Support Vector Machine, HMM and various other models are developed. If Artificial Neural network is used in the field of pattern and character recognition than it will be possible to develop a model from existing models that can learn itself and than recognize row scanned image. Self learning systems are always complex in design and implementation, but have great efficiency. Though design and structure of model may vary according to the structure of the script.

K-nearest neighbor (K-nn) Classifier is widely used technique in OCR field. In template matching each character is compared with set of templates and the best matching result is given in output. Than the post processing phase converts output into standard text format.

Information available is not restricted to a language. It is available in various languages. The scripts of different languages have different characteristics,

hence new models has to be designed which makes use of unique characteristics of local scripts to recognize them easily. [1]

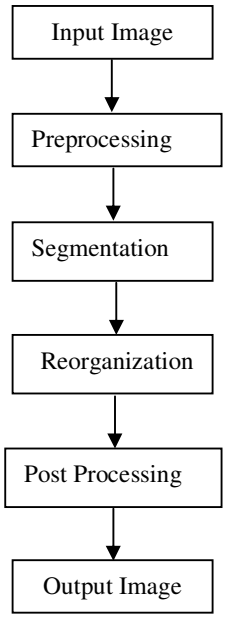


Figure 1 Block Diagram of OCR.

Antani[2] describe the classification of a subset of printed or digitized Guajarati characters using Template Matching Technique, it has low recognition rate of 67 %.

## 2. CHARESTIRESTIC OF GUJARATI SCRIPT

Gujarati is phonetic language in western India. Gujarati script is written from left to right, with each character representing a syllable. The character set of Gujarati script consist of 12 vowels, which are called *Swar* and 34 consonants, which are called *Vyanjan*. These are shown in Figure 2 and Figure 3 respectively.

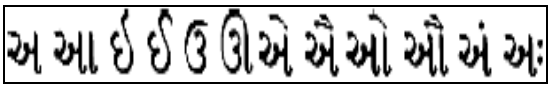

Figure 2 Vowels of Gujarati Language.

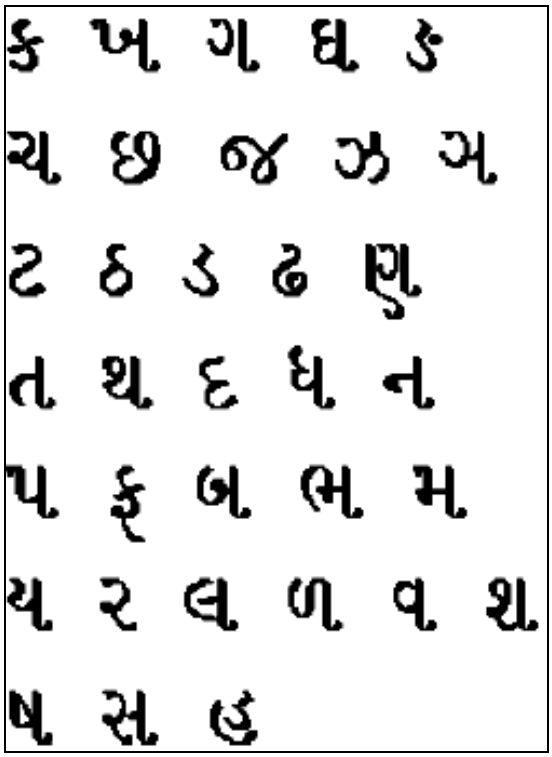

Figure 3 Constants of Guajarati Language.

Gujarati consist of set of special modifier symbols called *Maatras*, corresponding to each vowel, which are attached to consonants to change their sound. The modifiers corresponding to each vowel is shown in Figure 4.

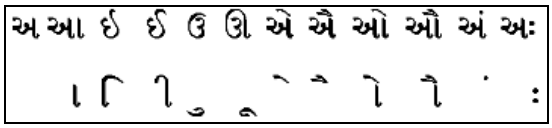

Figure 4 Special Symbols of Guajarati Script.

Gujarati consists of set of special modifier symbols called *Maatras*, corresponding to each vowel, which are attached to consonants to change their sound. The modifiers corresponding to each vowel is shown in Figure 3. First, Vowel does not have any corresponding modifier but is basic sound for the consonants. Modifiers are placed at the top, at bottom right or at bottom part of the consonant. They can be attached at different positions for different consonants. They can occur in different shapes depending on the consonant to which it is attached. All the Character (*Vyanjans and Swars*) and modifiers (*Maatras*) together roughly provide basic orthographic units, which are referred as glyphs that are combined together in different ways to represent all the frequently used syllables **Error! Reference source not found.**.

In Guajarati each consonant visualy is the combination of its original form and *Haswaksher* and *Maatra* Vowels. Each haswakshar is obtained by placing ◌ below the consonant. But when there is need of consonant without vowels haswaksher is used, shown in Figure 5.

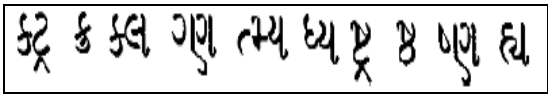

Figure 5 symbols for consonants without vowels sounds

A character is *conjunct* if two half consonants are joined (Figure 5). In *Conjunct* the shape of consonant sometimes get change. A character is said to be simple if it is a consonant alone or with a *Maatra*(Figure 2 and 3).

### 3. RECOGNIZATION TECHNIQUE

There are number of models for pattern and character recognition, but with change in character set and scripts there is drastic change in performance. S. K. Shah and A Sharma[4] has used Template matching technique for Guajarati script recognition and Consonants recognition rate was 78.34 %, and overall recognition rate was 70%.

Nearest Neighbors is one of mostly used classifier. NNC(Nearest Neighbor Classifier) takes an input character and than it searches for the best- nearest matching character from the training set. But when there is an input matches to training set and partially, its not completely matching training set than NNC will not recognize.[5]

K-nn(K nearest Neighbor) is modified from NNC, where K is weight, used for comparison from training set. For Example K=3*(odd value of K is mostly preferred to resolve the problem of equal voting)*, than system will search for three nearest neighbors and from the result of comparison the winning neighbors will be given as output from training set. K-nn works perfect in limited condition. When training set becomes too large its search space increases and that increase its recognition time. Second problem is if a unknown sample is there than there is higher chance of false recognition, instead of considering it as new pattern. K-nn is simple and effective.

Kohonen SOM[7] is iterative training process with unsupervised learning. It assigns high dimensional input vector into a neuron(node) in low dimensional space. SOM architecture contains competitive layer where the network nodes are arrange in two dimensional grid. Unsupervised learning is a means of modifying the weights of a neural network without specifying the desired output for any input patterns. The advantage is that it allows the network to find its own solution, making it more efficient with pattern association. The advantage of using SOM is it simplifies its dimensionality and it preserves neighborhood property of input vector which is important for new unseen sample.

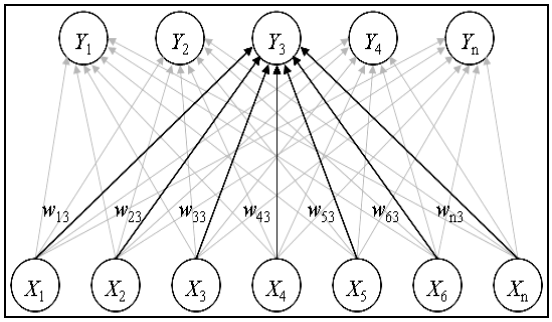


Figure 6 self-organizing network with five cluster units, $Y_i$, and seven input units, $X_i$.

The five cluster units are arranged in a linear array. Number of nodes(*Xi*) in input layer are always equal to the number of inputs. But for output cluster(*Yi*) it depends on the problem, here it varies according to script.

### 4. PROPSED TECHNIQUE

The advantage of K-NN method in comparison with other classifier is that it's simple & effective, new data samples can be added easily in training set for future classification and do not require prior training, But with this drawbacks of K-nn are, when the data is categorical its difficult to compute the distance between two samples. CPU cycle consumption and time consumption varies linearly with increasing size of training set in K-nn. Scripts like Gujarati where *consonants, haswakshers, maatras* are there it becomes quite complex and time consuming execution of K-nn.

If data is pre-processed before giving as input in K-nn. Than it can reduce time and CPU consumption for recognition. If K-nn is binded with SOM than it we can take advantage of both. SOM makes network to find out its own solution according to input. Preprocessing using SOM is helpful by-

1. Similar data samples are placed in 2 layered map to it nearest similar data set.
2. Once mapping of test samples for training is completed. Any of the suitable method for distance calculation can be used.
3. Now very similar data sets(because similarity of data set is preserved) are formed together and a cluster is made. And its mean value is calculated.

Now we need to modify K-nn, K-nn finds neighbors from the training set. It checks each training sample individually. Now K-nn will find distance with the mean of a cluster formed by SOM. Once nearest cluster is found than only member of that cluster are compared. This is because all neighborhood data

samples are stored in same cluster *(neighborhood property of training sample is preserved in SOM)*. Steps for proposed approach are as below.

The proposed method is partitioned in two phase.
1. Pre-processing using SOM.
2. Classification using K-NN.
They are detailed as follow

1. Pre-processing step using SOM
Step1: Convert all NxM character images into binary vector of size NxM.
Step2: Map all NxM binary patterns on to SOM. Output will be 2D lattice.
Step3: To find 2D coordinates Di(x,y) for each pattern in dataset.
Step4: Store 2D coordinate Di(x,y) with the class label Ci in database.
2. K-NN Classifier.
Step1: For a new real data its distance will be calculated with the cluster mean
value.
Step2: K nearest neighbors will be selected and majority votes cluster will be
declared as winner.
Step3: From that cluster data samples are used for comparison and input pattern
will be recognized.

## 5. SIMULATION AND RESULTS

Experiment was conducted for 10 data samples from 5 different fonts, fonts were selected in such a way that changes in curve of consonants of other fonts can be included in range of selected fonts. So for each character total 50 samples were tested in system. This procedure was repeated for all consonants of Guajarati script. The accuracy found in characters was 91.05% and if considered with impure consonant than it is 89.37%. When same data set is tested for k-nn than it was 80.04%. This increase in efficiency and recognition rate is due to binding k-nn with neural network.

## 6. CONCLUSION

Hence results states that biding k-nn with neural network has given good recognition rate. Merging Neural with existing methods for recognition has given optimum results and best recognition rate. This can be further extended for other scripts and other application area and can give glorious results.

## 7. REFERENCES

[1] B Krishna. 'Design and Implementation of a Telugu Script Recognition System'. Technical Report, Department of Computer and Information Sciences, University of Hyderabad.

[2] S Antani and L Agnihotri. 'Gujarati Character Recognition'. Proceedings of the International Conference on Document Analysis and Recognition, (ICDAR-99), Bangalore, India, 1999, pp 418-421.

[3] Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching, Prof S. K. Shah, A Sharma.

[4] A Nearest Neighbor Approach to Letter Recognition, Aiyuan Ji, Roy George.

[5] K-Nearest Neighbor Learning, Dipanjan Chakraborty, Presentation.

[6] Link-http://mnemstudio.org/neural-networks-kohonen-self-organizing-maps.htm

[7] Teuvo Kohonen "The Self-Organizing Map 3e", *Springer Publication,* 2000.