Recognition of Off-line Hand printed English

Characters, Numerals and Special Symbols

Nisha Sharma , Bhupendra Kumar

CDAC Noida

Vandita Singh

Vishveshwarya College,Uttar Pradesh

India

India

Nishasharma062@gmail.com,

bhupendra@cdac.in

vandita.it.24@gmail.com

Abstract—

The

generic

process

of

Optical

Character

challenges for higher recognition rates. Special Applications

and systems include Hand written formula recognition, Bank

check analysis and recognition and Information retrieval,

Forms Processing, OMR Sheet Processing.

Recognition (OCR), an area of intensive research in the field of

Artificial Intelligence, Pattern Recognition and Computer Vision,

aims to recognize text from scanned document images, where

data can be in machine printed or hand written format. Optical

Character Recognition can improve the interaction between man

and machine in various applications including data entry, office

The objective is to develop an offline OCR system to

recognize Hand printed English Characters, Numerals and

Characters from the document images, which

comprises of text in hand printed format, which would be

converted into editable form. Hand printed document image

implies that input image comprises of mono-spaced characters

whereas in handwriting we have characters that connect,

whereas Hand printed input document has more or less

uniform height or width.. Hand printed document images also

imply that there must be present uniform base-line character

images (same horizontal base-line).

automation,

digital

library,

banking

applications,

health

Special

insurance and tax forms etc. Much of work has been done in the

recognition of machine printed characters in various languages

with considerably good efficiencies, however making robust

recognition engines that can be put to recognize hand written and

hand printed data with commendable recognition rates still

remains as an active area of research owing to the challenges like

diverse human handwriting style, variation in shape, angle and

style of characters. Taking into account the challenges and scope

for improvement in this domain, the work of off-line character

recognition of hand printed document images containing English

Characters-Uppercase and Lowercase, Numerals and Special

Characters

has

been

presented.

Statistical,

Geometric

and

There can be two modes of recognition, namely off-line

and on-line [16] for obtaining input document image. On-line

document image recognition implies to storing the character

image as a function of time dynamically as the character is

Directional Feature Extraction techniques have been applied

over segmented character image. Classification was done using

Multilayer

perception

neural

network

(NN)

with

back

propagation and Support vector machine (SVM) classifier. The

recognition rates achieved were up to 98% for Numerals, 96.5%

for Special Characters, 95.35% for Uppercase Characters, 92%

for Lowercase Characters. The system for combined data set-

Characters, Numerals and Special Symbols resulted out to be

92.167% accurate, using SVM as classifier.

being

drawn

electronically;

hence

the

spatio-temporal

information [16] such as order of strokes made by the writer,

information about pressure and angle of the pen is readily

available. In case of off-line document where acquisition is

done prior to recognition, the spatio-luminance [16] of the

image is analysed. Therefore, more challenges would be

present while recognizing documents in offline mode since we

have only static information about the document.

Keywords— — Hand printed character recognition, hand

printed numeral recognition, Statistical, geometric and topological

features, neural network classification, SVM classifier

An OCR system comprises of different phases as Data

Acquisition, Pre-processing, Feature extraction, classification

and post-processing [Fig. 1]. Pre-processing includes noise

cleaning, skew correction, binarization, segmentation and

normalization techniques. In feature extraction phase a set of

useful properties of a character are extracted. Statistical,

I.

INTRODUCTION

Optical

Character

Recognition

aims

at

identifying characters in images of printed or handwritten text,

in order to encode the text in a format more convenient to

edit. Recognition of cursive text has been an active area of

research, due to the challenges faced during recognition

process as the process incurs high uncertainty in the input

documents as writing styles may vary abruptly depending on

the interpersonal and intrapersonal variations.. Therefore, it

stands out to be a challenging task to devise an OCR system

for handwritten document image. Noise, broken characters,

touching characters and inappropriate scanning induces further

Directional,

Topological and Geometric features were

extracted to uniquely identify each character. Based on these

properties character is assigned to a class in the classification

phase. Classification phase involves two major steps -training

and testing. Neural Network and Support Vector Machine

were used independently for the experiment at the classifier

level. Post processing has been applied on the classified

characters to generate text output according to input character

978-1-4799-4236-7/14/$31.00c 2014 IEEE

640

sequence. Fig. 1 shows the major steps involved in the

approach.

Fig. 2. Sample Images of Hand printed Scanned characters

Fig. 3. Sample Images of Hand printed Scanned Lowercase Characters

Fig. 4. Sample Images of Hand Printed Scanned Numerals

Fig. 5. Sample Images of Hand printed Scanned Special Symbols

B. Pre-processing

Pre-Processing enhances a document image preparing it

for the feature extraction phase in the OCR system. In order to

achieve high recognition rate, prior to character recognition, it

is essential to eliminate the noise and imperfections introduced

in the image. The three techniques adopted for pre processing

in this approach were Binarization, Thinning and Size

Normalization. In Binarization, grey-scale image is converted

into binary image with the help of thresholding. The Otsu

method was applied to convert the image into binary image.

Otsu’s method selects the threshold by minimizing the within-

class variance of the two groups of pixels separated by the

thresholding operator.

Fig. 1. Major Steps involved in the process of OCR

Section II of this paper discusses the Data Collection,

various Pre-processing steps performed before the data is

forwarded to feature extraction. Section III discusses the

feature extraction techniques used to extract features which

can uniquely identify each character. Section IV discusses the

neural network and SVM classification of the characters along

with post processing. Section V and Section VI discuss the

Testing Results and the conclusion of the approach used

respectively.

The technique of Thinning converts the character images

to single pixel width. The binary image is represented as ones

and zeros. “1” is used to represent object pixel and “0” is used

to represent background. Zhang-Suen Algorithm [5] was used

for this purpose. Segmentation is a process that determines the

constituents of an image. It is necessary to locate the regions

of the document where data have been printed and distinguish

them from figures and graphics. When applied to text,

segmentation is the isolation of characters or words. In this

approach, Contour tracing was used to segment each character

image. Features extracted from the processed image should

not be affected by the size of the image. Each Segmented

Image was normalized to 20 x 20 pixels dimensions.

II.

DATA COLLECTION AND PRE-PROCESSING

A. Data Set Preparation

The data set was obtained by taking written samples of 40

people, where each writer had written a document containing

instances of each of the 26 characters, 10 numeral digits or 8

special characters in the range-[10-30].These documents were

scanned at resolution-200 and 300 dpi. There were nearly

3517, 2340 and 1804 instances of Upper case characters,

Lowercase

characters and Numerals respectively. The

collected data has been divided into the ratio of 60:40 for the

training and testing purpose. However, these partitions were

changed iteratively to test the system for best accuracy and for

obtaining the optimal data set. The digitized images were

saved in BMP format. Fig. 2, Fig. 3 and Fig. 4 show samples

of acquired images for hand printed characters (Uppercase and

Lowercase) and Numerals, respectively.

III.

FEATURE EXTRACTION

The objective of feature extraction phase is to extract the

essential and differentiable characteristics of the symbols.

Feature space is much less than input image space as we

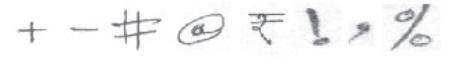
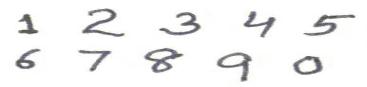
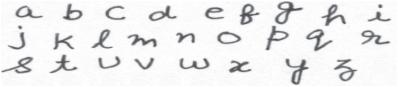
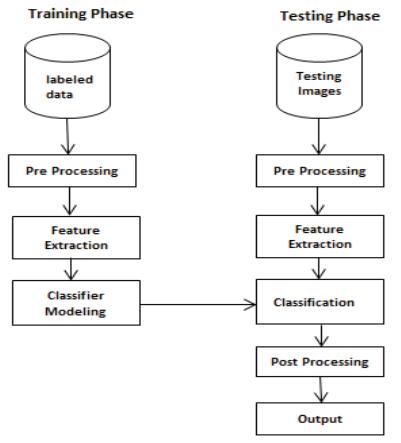
extract only essential properties for higher recognition rate.

Features are classified into the following categories on the

basis of methods of extraction.

*2014 5th International Conference- Conﬂuence The Next Generation Information Technology Summit (Conﬂuence)*

641



In the approach combination of Statistical, Directional,

Geometrical and Topological features have been implemented

to uniquely identify each character.

A. Statistical Features

Initially, statistical feature (Multi zoning abbreviated as

MZ) extraction technique was applied to each segmented

character image by calculating the percentage of black pixels in

one zone of a character image as shown in Fig. 6. The character

image was divided into zones, taking data from each zone as a

feature for the image. The character image was divided 5 times

into zones. The dimensions for each of the five divisions are as

follows- 2x2, 3x3, 4x1, 3x5 and 5x5.These zones have been

chosen to cover most of the region in an image which contains

information, however can be optimized on the basis of

experimental results. The feature vector was, hence comprising

of 69(4+9+4+15+25) features which were obtained using

multiple dimensions for zoning[1].

Fig. 7. Geometric Feature Extraction

C. Topological Features

Topological features (End points and Transitions) were

extracted after thinning each pre-processed image and deriving

end points (E) with respect to the zone shown in Fig. 8. These

features have been used to eliminate confusions among

various characters at the stage of post-processing.

The number of transitions of black to white and white to

black pixel horizontally and vertically were calculated for

further rectifying the confusions at post-processing level. For

example Vertical transition can be used to remove confusions

among “y” and “z” shapes. Till one fourth of height of an

image, “y” is having two transitions and “z” is having one

transition as shown in Fig. 9.

B. Geometrical Features

Geometrical features (Distance, angle) were obtained

after

performing Thinning and Normalization on each

character image. Centre of the skeleton was calculated. Then

image was then divided into 3x3 zones as show calculated.

Then image was then divided into 3x3 zones as n in Fig 7.

In each zone Geometrical features were calculated as

follows [Fig. 7]:

D. Directional Features

A directional feature extraction technique-Chain Code



Calculating average of distances of each pixel present

in zone from center point as given by 1 (Total of 9

features) [3].

Histogram(CCH)

was implemented to obtain directional

information in Fig. 9, based on contour of an image. This

technique was also implemented at the zone level. There were

2 divisions (2x2 and 3x3), made for obtaining sub-images from

the character image. For each sub-image, i. e in each zone, the

contour was traced to find the external boundary pixels and

corresponding directions Fig. 10 were computed for each

boundary line segment. Since the chain codes computed for

each character was different, there would have aroused

difficulty in comparison when used as features. Thus the chain

code histogram for each segment was computed and used as

feature. Hence, for each zone there was an addition of 8

features. For each character image having 13 zonal divisions,

the total number of added features was 104(13x8).

(1)





Then average of angles of each pixel present in zone

from centre point as given by 2 was calculated (Total

of 9 features) [3]

(2)

Where x and y is the centre point co-ordinate of

m

m

image, x and y are the coordinates of considered pixel

i

i

and n is total number of data pixel present in the zone.

k

Fig. 8. Topological Feature Extraction : End Points

Fig. 9.

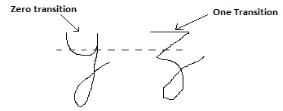
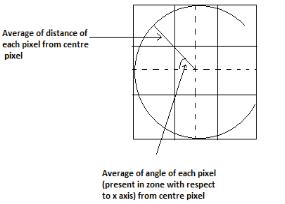
Transitions

Fig. 6. Statistical Feature Extraction:Division of character image into M X N

zones M=2,3,4,5 and N=2,3,1,5

642

*2014 5th International Conference- Conﬂuence The Next Generation Information Technology Summit (Conﬂuence)*



V.

EXPERIMENTAL RESULTS

The system was tested on 40 different hand writings

having instances of each character or numeral in the range [10,

20] for 26 classes of characters and 10 different classes of

numerals and 8 special symbols. Testing was performed for

Writer Dependent and Writer Independent Systems.

For Writer Dependent System, in case of Uppercase

characters with MZ as feature and SVM as classifier, the

results came out to be 99.23%, 99.227% and 98.205% for 75%

of training data set, 50% of data set and 25% of data set

respectively, taken in training. In case of Lowercase characters

with MZ as feature and NN as classifier 90% accuracy was

achieved on 70% of training data. In case of lowercase

characters, with MZ+DATEP as a feature and NN as classifier

97.6% accuracy was achieved. In case of numerals with

MZ+DATEP as a feature and SVM as a classifier 99.4%

accuracy was achieved. With MZ+DATEP as feature and NN

as a classifier, 99% accuracy was achieved. In case of Special

Symbols MZ+DATEP as feature and SVM as classifier 91%

accuracy was achieved. With MZ+DATEP as feature and NN

as classifier, 86% accuracy was achieved.

Fig. 10. Directional Features: Direction Codes

IV.

CLASSIFICATION

The classification is the process of identifying each

character and assigning to it the correct character class. The

Classification techniques used were MLP neural network with

back propagation and Support Vector Machines.

A. MLP neural network with back propagation

a back-propagation neural network, the learning

In

algorithm has two phases. First, a training input pattern is

presented to the network input layer. The network propagates

the input pattern from layer to layer until the output pattern is

generated by the output layer. If this pattern is different from

the desired output, an error is calculated and then propagated

backwards through the network from the output layer to the

input layer. The weights are modified as the error is

propagated. A back propagation MLP NN is shown below .

For Writer Independent System, in case of Uppercase,

Lowercase, Numerals and Special Symbols, the results have

been tabulated. Using MLP back propagation neural network,

the recognition rates have been presented in Table I and using

SVM as the classifier, the recognition rates have been

presented in Table II. Here Feature Set is denoted as: MZ=

zoning; CCH=Chain Code Histogram; DATEP=

Distance, Angle, Transition, End points.

B. Support Vector Machines

Another classification technique used was support vector

machine which follows the structural risk minimization. SVM

is a kernel-based algorithm. A kernel is a function that

transforms the input data to a high-dimensional space where

the problem is solved. Kernel functions can be linear or

nonlinear. In SVM classification, weights are a biasing

mechanism for specifying the relative importance of target

values. The kernel function used in SVM was RBF (radial

basis function) denoted as [9], where gamma is 1/number of

features.

Multi

The best accuracy found was 95.35% for Uppercase on

MZ\_CCH (Table II) features, 92.31% for lowercase on

MZ\_DATEP (Table II, 96.52% for Special symbols on

MZ\_CCH\_DATEP (Table II), and 98% (Table I) and Table II)

for numeral on MZ\_DATEP, using SVM as the classifier.

The major misclassifications (Fig. 12) were due to the

similar topology of the characters, numerals. Most confusing

characters was “y” and “g” ,”l” and “t”, “g” and “z”, ”r” and

“x”, “e” and “c” ,”Q” and “G”. “U” and “O”, “Y” and “V” etc.

whereas most confusing numerals were “4” and “9”, “6”and

“5” ,”1” and “7”.

(3)

Combinations of features as explained above were passed

to SVM classifier for classifying characters.

Fig. 11 Back Propagation MLP Neural Network

Fig. 12. Example of misclassified Images

*2014 5th International Conference- Conﬂuence The Next Generation Information Technology Summit (Conﬂuence)*

643

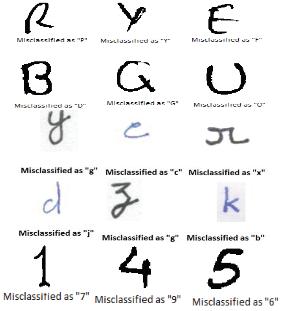
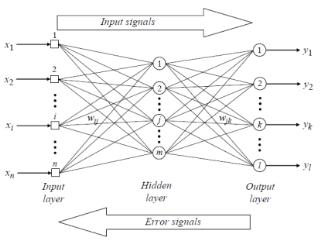
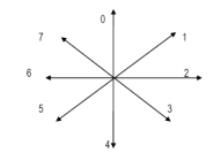


TABLE I.

RESULTS ON NEURAL NETWORK

Lowercase

characters+

Special

1312/

2624

MZ+CCH+DATEP(193

)

Training/

Total

8

9

34

18

51

20

15

25

94.9%

92.8%

93.3%

S.

N

o

Writer Independent Class Write

System

Accura

cy

Symbols

Database

60/40

Feature Set

es

rs

Numerals

+Special

Symbols

1051/

1751

MZ+CCH+DATEP(193

)

ratio

2314/351

7

Uppercase

Character+

Lowercase

Characters

Uppercase

characters+

Lowercase

Characters +

Numerals

1

Uppercase

26

15

MZ(69)

93.1%

3475/

5867

MZ+CCH+DATEP(193

)

10

2314/351

7

2

3

4

5

Uppercase

Lowercase

Lowercase

Numerals

26

26

26

10

15

10

10

5

MZ+DATEP(29)

MZ(69)

74.5%

79.2%

92.3%

98%

1560/

2340

1560/

2340

600/

MZ+CCH+DATEP(193

)

92.8%

81.5%

4013/

6729

11

59

67

30

40

MZ+DATEP(29)

MZ+DATEP(29)

MZ+DATEP(29)

1000

Uppercase

characters+

Lowercase

12 Characters +

Numerals

+Special

Symbols

MZ+CCH+DATEP(193

)

MZ+DATEP(29) 81.40%

92.2%

80.4%

480/

800

6

Special Symbols

8

10

4497/

7533

MZ+CCH+DATE

91.20%

P(193)

1921/

3237

2667/

4447

4013/

6729

MZ+DATEP(29)

6

7

8

Lowercase+ Numerals

Uppercase +Numeral

35

35

59

15

20

30

MZ+DATEP(29)

85.4%

77.4%

89%

MZ+DATEP(29)

Uppercase+Lowercase+

Numeral

MZ+CCH+DATE

P(193)

VI.

CONCLUSION

Geometric, statistical and topological and directional

feature were used along with MLP back propagation neural

network and SVM classifier independently. Neural networks

have been preferred to be used due to their high noise

tolerance and SVM, due to its high flexibility, scalability and

speed.

TABLE II.

RESULTS ON SUPPORT VECTOR MACHINES

Training

/Total

S.

Writer

Accuracy

SVM

Classe Writer

dataset

SVM was found to outperform MLP back propagation

neural network as evident from the results in Table I and Table

II. It can be observed that using SVM as the classifier, MZ

was found to play an important role in defining the

characteristics of the symbols in each class-Uppercase ,

Lowercase , Numerals and Special Symbols. Along with MZ,

CCH played a more significant role in defining Uppercase

characters, DATEP in case of Lowercase characters and

Numerals; and combination of CCH+DATEP lead to define

the characteristics of special symbols in the best way. It can

also be seen that symbols belonging to the class of Uppercase

characters gave best results when 173 features were used with

SVM. However, lowercase characters gave best recognition

results with 29 features with SVM. In case of Numerals, 29

features were found to be sufficient for giving best recognition

rates with both SVM and MLP back propagation neural

network. In case of Special Characters, best recognition rates

were achieved with SVM using 193 features. Thus it can be

inferred that recognition rates depend on type of features-

MZ,CCH or DATEP selected for a certain input class -

uppercase, lowercase, numerals or special symbols, number of

features-29,69,104,173 or 193 and type of classifier- SVM or

MLP back propagation Neural Network .

N Independen

Feature Set

s

s

o

t System

60/40

ratio

MZ+CCH(173)

CCH(104)

95.4%

88.5%

1

2

Uppercase

Characters

Lowercase

Characters

26

26

15

10

2314/351

7

1560/234

MZ\_DATEP(29)

MZ(69)

92%

79.5%

0

1154/

1804

MZ+CCH(173)

95.3%

3

10

5

571/

MZ+CCH+DATEP(193

Numerals

96.8%

98%

951

600/1000

)

MZ\_DATEP(29)

MZ+CCH+DATEP(193

)

96.6%

Special

Characters

480/

800

4

5

8

10

20

MZ+DATEP(29)

86.9%

94.9%

3399/

5791

2667/

4447

Uppercase

Characters+

Numerals

MZ+CCH(173)

35

MZ+CCH+DATEP(193

)

95.8%

Uppercase

characters+

Special

2690/

4466

MZ+CCH+DATEP(193

)

6

8

10

15

94.9%

Symbols

Lowercase

7 Characters +

Numerals

The further research can be focused on exploring new or

hybrid feature extraction methods which can be used for

training the neural network and SVM so that the efficiency of

system can be improvised.

1921/

3237

MZ+CCH+DATEP(193

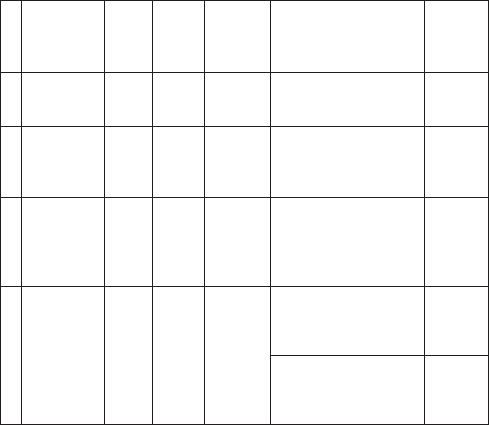
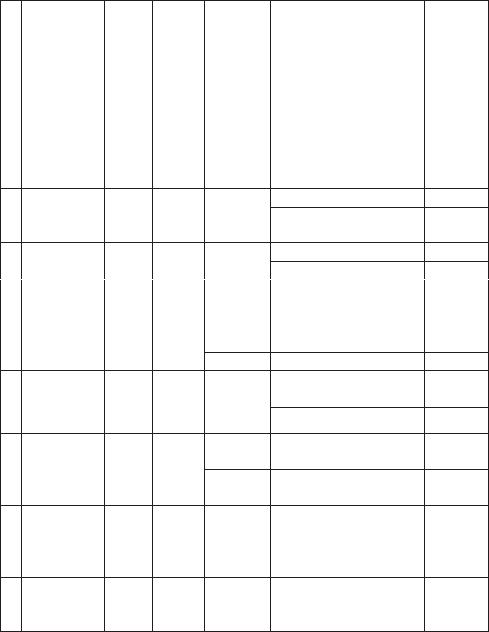
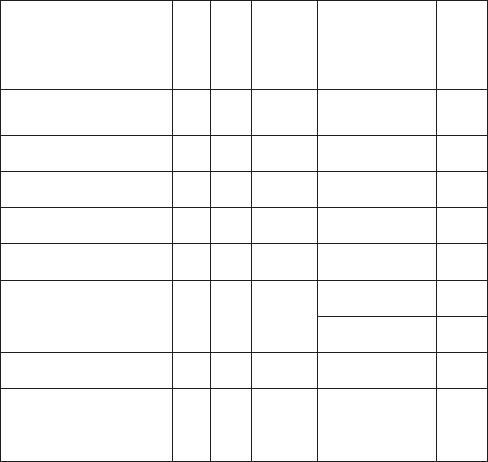
)

35

95.2%

644

*2014 5th International Conference- Conﬂuence The Next Generation Information Technology Summit (Conﬂuence)*



ACKNOWLEDGMENT

[8]

[9]

Sameer singh, Adnan Amin “Neural network recognition and analysis

of hand printed letters.” IEEE 1998.

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin “A Practical

Guide to Support Vector Classification” Department of Computer

Science National Taiwan University, Taipei 106, Taiwan.

N.Sharma, V. Singh and B.Kumar are grateful to Mr.

Deepak Kumar Arya, Senior Technical Officer, C-DAC Noida

for his technical guidance. Also, N.Sharma, V. Singh and

B.Kumar would like to express our immense gratitude towards

Mr. Tushar Patnaik, Senior Project Engineer , C-DAC Noida

for his consistent guidance and support.

[10] Thresholding(Jain et al., Sections 3.2.1, 3.2.2, Petrou et al., Chapt 7)

[11] Plamondon

and

Srihari,

“On-line and

Off-line

Handwriting

Transactions on

Recognition-A

Comprehensive

Survey”

IEEE

Pattern Analysis and Machine Intelligence,Vol 22, No.1, January

2000

[12] M. Shridhar and A. Badreldin, “A High

Recognition Algorithm for handwritten Numerals”, Proceedings of

the IEEE International Conference on Systems , Man and

Cybernetics, Vol. SMC-15, No. 1, January/February 1985

Accuracy Syntactic

REFERENCES

[13] L. Heutte, T. Paquet, J. Moreau, Y. Lecourtier and C. Olivier,

“Combining Structural and Statistical Features for the Recognition of

Handwritten Characters”, ICPR, Vol. 19, pp. 629-641, 1998.

[1]

[2]

Rafael M. O. Cruz, George D. C. Cavalcanti and Tsang Ing Ren . “An

Ensemble Classifier For Offline Cursive Character Recognition Using

Multiple Feature Extraction Techniques,”Neural Network(IJCNN)

IEEE International Conference 2010.

Anshul Mehta, Manisha Srivastava, Chitralekha Mahanta “Offline

handwritten character recognition using neural network” IEEE 2011

International conference on computer applications and Industrial

Electronics.

[14] C.- L. Liu, K. Nakashima, H. Sako and H. Fujisawa,

“Handwritten

Digit

Recognition:

Benchmarking

of

State-of-the-Art”,

Pattern

Recognition, No. 36,

pp. 2271-

2285 , 2003.

[15] J.T. Favata, G. Srikantan, S.N. Srihari, “Hand printed character/digit

philosophy”.

Proceedings of the Fourth International Workshop on Frontiers of

Handwriting Recognition, Taipei, 1994, pp.57–66.

recognition

using

a

multiple

feature/resolution

[3]

[4]

[5]

Vamsi K. Madasu1, Brian C. Lovell , M. Hanmandlu “Hand printed

Character Recognition using Neural Networks”.IEEE

[16] L. Koerich, “Large Vocabulary Off-line Handwritten

Word

Superieure,

Skeletonization:

CAP400.

Zhang-Suen Thining Algorithm. Jason Rupard

Recognition”,

Ph.

D.

Thesis,

Ecole

de

Technologie

Montreal - Canada, 2004.

Christopher E. Dunn and P. S. P. Wang “Character Segmentation

techniques for handwritten text-A Survey” 11 IAPR International

th

conference on Pattern Recognition Vol II IEEE 1992.

[17] Ithipian Methasate, Sanparith Marukatat, Sutat Saetang and Thanarak

Theeramunkong, “The Feature Combination Technique for Offline

Thai Character Recognition System” IEEE, Eighth ICDAR 2005

[6]

[7]

Adnan Amin and W. H. Wilson.“Hand-Printed Character Recognition

Using Artificial Neural Networks” Proceedings of 2nd

International Conference on Document Analysis and Recognition

IEEE 1993.

[18] R.Jayadevan, Satish R.Kolhe, Pradeep M. Patil, Umapada Pal. 2011.

System

“Offline

Recognition of Devanagari Script-A Survey”, IEEE

transactions on Systems, Man and Cybernetics-Part C: Applications

and Reviews, Vol.41, No.6, November 2011

Yuk Ying Chung, Man to Wong “handwritten character recognition

by Fourier descriptors and neural network”, Speech and Image

Technologies for computing and telecommunication. Proceedings of

IEEE 1997.

*2014 5th International Conference- Conﬂuence The Next Generation Information Technology Summit (Conﬂuence)*

645