# International Journal of Computer and Communication Technology

Volume 3 | Issue 1

Article 7

January 2012

## **Oriya Character Recognition using Neural Networks**

Soumya Mishra P.G.Department of CSA Utkal University, soumyalitun@gmail.com

Debashish Nanda P.G.Department of CSA Utkal University, emailtodebasish@gmail.com

Sanghamitra Mohanty Dept of Computer Sc. & Application Utkal University Bhubaneswar, Orissa, India, sangham1@rediffmail.com

Follow this and additional works at: https://www.interscience.in/ijcct

## **Recommended Citation**

Mishra, Soumya; Nanda, Debashish; and Mohanty, Sanghamitra (2012) "Oriya Character Recognition using Neural Networks," *International Journal of Computer and Communication Technology*: Vol. 3 : Iss. 1 , Article 7. DOI: 10.47893/IJCCT.2012.1111 Available at: https://www.interscience.in/ijcct/vol3/iss1/7

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer and Communication Technology by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

## **Oriya Character Recognition using Neural Networks**

Soumya Mishra<sup>1</sup>, Debashish Nanda<sup>2</sup>, Sanghamitra Mohanty<sup>3</sup>

P.G.Department of CSA

Utkal University

soumyalitun@gmail.com<sup>1</sup>, emailtodebasish@gmail.com<sup>2</sup>, sangham1@rediffmail.com<sup>3</sup>

Abstract—The biggest challenge in the field of image processing is to recognize documents both in printed and handwritten format. Optical Character Recognition (OCR) is a type of document image analysis where scanned digital image that contains either machine printed or handwritten script input into an OCR software engine and translating it into an editable machine readable digital text format. Development of OCRs for Indian script is an active area of research today. We are making an attempt to develop the OCR system for Oriya language, which is the official language of Orissa. Oriva language present great challenges to an OCR designer due to the large number of letters in the alphabet, the sophisticated ways in which they combine, and the complicated graphemes they result in. In this paper, we argue that a number of automatic and semi-automatic tools can ease the development of recognizers for new font styles and new scripts. We discuss briefly and show how they have helped build new OCRs for the purpose of recognizing Oriya script. We have used the Back propagation Neural Network for efficient recognition where the errors were corrected through back propagation and rectified neuron values were transmitted by feed-forward method in the neural network of multiple layers, i.e. the input layer, the output layer and the middle layer or hidden layers.

Keywords-Optical Character Recognition, Artificial Neural Network, Histogram, Binarization, Back Propagation.

#### I. INTRODUCTION

Optical character recognition (OCR) systems have been under research for few decades. But still it remains a highly challenging task to implement an OCR that works under all possible conditions and gives highly accurate results. Optical character recognition of Oriya characters is a challenging field. Being an Indian language Oriya is ornamented with fabulous literature and sophisticated grammar. Identification of Oriya scripts though is not an easy task; still many fruitful attempts have been made. This proposed approach is very much different from the previous approaches. Artificial Neural Network technique is totally based on feature matching. Therefore various geometric features and character features are extracted from segmented text. Features are collectively evaluated and a combined score is obtained. text blocks, sentence/line blocks, word blocks character blocks. The character blocks are then further broken down into components parts, pattern recognized and compared to the OCR engines large dictionary of characters from various fonts and languages. The approach used here is Artificial Feed Forward Neural Network, which fuses different extracted geometric such as height, width, number of pixels in columns and rows and textual features such as histogram, centroid etc to get a valid epoch or score. If the score is within the desired range then the character is recognized, else it suggests the user to back propagate and balance the input parameters by adding or subtracting weight values and bias values with overall score.

Verification of provided information and validation of data item depends on a threshold value previously imposed on the knowledge base system. Proposed technique is efficient under idle circumstances and limited corpus size.

#### II. MOTIVATION

Orissa has a rich heritage of manuscripts and novels, which are need to be preserved in Oriya language and Oriya scripts. The invaluable ancient works are in the process of being lost to termites in libraries in India, due to the lack of proper OCR systems. Automatic recognition of alphabetic characters through computers is a basic need for text recognition in different languages .OCR system for different foreign languages like Japanese, Chinese and Korean are already been developed. For Indian Languages attempts are made for Devanagari, Gurmukhi, Marathi, Bengali, Telugu and Tamil. We are making an attempt to develop the OCR system for Oriya language. In order to read those scripting we have made a sincere attempt to develop Oriva OCR. Therefore, our attempt here is a humble effort towards this noble goal of developing an efficient OCR system, in public domain, for Oriya characters.

#### III. METHODOLOGY

There are two basic methods used for OCR: Matrix matching and feature extraction. Of the two ways to recognize

OCR works by first pre-processing the digital page image into its smallest component parts with layout analysis to find

International Journal of Computer and Communication Technology (IJCCT), ISSN: 2231-0371, Vol-3, Iss-1

characters, matrix matching is the simpler and more common. But still we have used Feature Extraction to make the product more robust and accurate. Feature Extraction is OCR without strict matching to prescribed templates. This method is much more versatile than matrix matching. Matrix matching works best when the OCR encounters a limited repertoire of type styles, with little or no variation within each style. Where the characters are less predictable, feature, or topographical analysis is superior.

The Process of Optical Character Recognition of the document image mainly involves six phases:

- 1. Acquisition of Grayscale Image
- 2. Digitization/Binarization
- 3. Thinning and Edge Detection
- 4. Feature Extraction

Such as: Dimension Measurements

Storage Details

Centroid and Histogram

Pixel Counts in Total, Rows, and Columns.

5. Feed Forward Artificial Neural Network based Matching.

6. Recognition of Character based on matching score.

The acquisition phase uses a scanner or digital camera that catches photocopy of the text document as an image. The scanned image must be a grayscale image or binary image, where

binary image is a contrast stretched grayscale image. That grayscale image is then undergoes digitization. In digitization a rectangular matrix of 0s and 1s are formed from the image.

Where 0-black and 1-white and all RGB values are converged into 0s and 1s.The matrix of dots represents twodimensional array of bits. Digitization is also called binarization as it converts grayscale image into binary image using adaptive threshold.

After that thinning is done to narrow edges in the image and along with eliminate the unwanted noise pixels. Thinning is a morphological operation that is used to remove selected foreground pixels from binary images. Edge detection is the process of identifying points in a digital image at which the image brightness changes sharply or more formally has discontinuities.

Feature Extraction such as dimension measurement like height and width of the image, storage details like file size and vector size. Centroid of the character image in horizontal and vertical coordinates, Histogram of luminance for each gray level value. Lastly pixel counts in the image in total and in each row and columns separately are calculated. After all the relevant features are extracted the corresponding numeric values are stored in the database.

Feed Forward Neural Network approach is used to combine all the unique features, which are taken as inputs, one hidden layer is used to integrate and collaborate similar features and if required adjust the inputs by adding or subtracting weight values, finally one output layer is used to find the overall matching score of the network. If the score is within the predefined range then the character is recognized else the system is trained again.

This OCR system is developed to recognize individual Oriya alphabets and provides a base to build complete OCR system for composite characters and entire page readability. Along with it aims to embed text-to-speech system and spellchecker to give better understanding and error correction. The Corpus size taken here is standardized to make the concerned system work smooth but intend to be enhanced in future.

Features to be incorporated in future:

There are ten structural features can be extracted from the 16x16 pixel matrix which are given below:

- 1. Upper Part Circular
- 2. A vertical line on the right most part
- 3. Holes
- 4. Horizontal Run code
- 5. Vertical Run code
- 6. Number of holes
- 7. Position of hole

## IV. ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN), usually called "neural network" (NN), is a mathematical model or

International Journal of Computer and Communication Technology (IJCCT), ISSN: 2231-0371, Vol-3, Iss-1

computational model that tries to simulate the structure and/or functional aspects of biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. An Artificial Neural Network usually consists of one input layer,more than one output layers,except the last output layer all other intermidiate layers are called hidden layer.Outcome of input layer is fed as input to the hidden layers and outcome of hiden layers are fed as input to the final output layer.

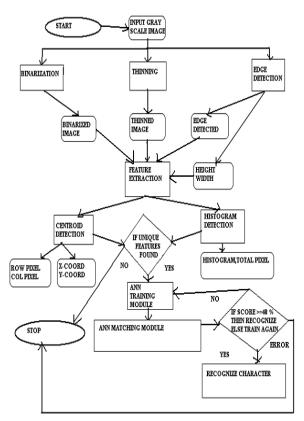
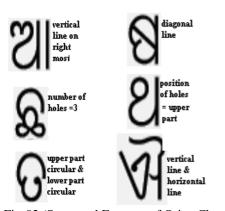
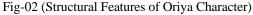


Fig-01 (Process Flow Chart)





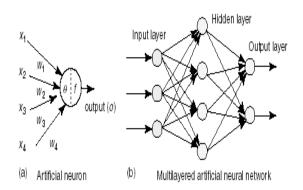


Fig-03 (a)Artificial Neural Network (b)Multilayer ANN

Algorithm of Feed Forward Neural Network:

1. Form network according to the specified topology parameters

2. Initialize weights with random values within the specified  $\pm$  weight bias value.

3. Load trainer set files (both input image and desired output text)

4. Analyze input image and map all detected symbols into linear arrays

5. Read desired output text from file and converts each character to a binary Unicode value to store separately

6. For each character:

a. Calculate the output of the feed forward network

b. Compare with the desired output corresponding to the symbol and compute error

c. Back propagate error across each link to adjust the weights

7. Move to the next character and repeat step 6 until all characters are visited

- 8. Compute the average error of all characters
- 9. Repeat steps 6 and 8 until the specified number of epochs a. Is error threshold reached? If so abort iteration
  - b. If not continue iteration

#### V. CONCLUSION

Due to no availability of 100% noise freed digitized image and presence of similar shaped characters the accuracy rate is affected. To tackle this type of problem the system has been integrated to spell checker with the help of dictionary and a huge corpus. During scanning the document may be slanted leading to problem at the time of line extraction. Noise arises due to so many reasons namely old document, low paper quality, and dust particles on the surface of scanner, low quality ink and low quality printing machines.

International Journal of Computer and Communication Technology (IJCCT), ISSN: 2231-0371, Vol-3, Iss-1

🛃 INPUT MODULE				
	INP	UT MODULE		
DAGE PREPROCESSING TabPage1 TabPage2 Tabl	heal			
INPUT BITMAP IMAGE	EDGE DETECTED			
ଷ	ଷ			
IMAGE PATH Cribecuments and Settle LOAD	EDGE DETECTION			
RPUT MODULE				
E TOOT MODULE	INP	UT MODULE		
BAGE PREPROCESSING	IN	CI NOL/ULE		
TahPage1 TahPage2 Tahl	Page3			
BI	NARIZATION		THINNING	
natching world.	21.707	INC MODIFIE		. • ×
height	width file_cize		w Juni en Juni ^	CLOSE
height F 6	width file_size 48 1118	vector_size 0 8064 B	FE.000000 457.00000	CLOSE SHOW DATABASE
height	width file_cize	vector_size 0 8064 B		CLOSE SHOW DATABASE ENTER INSTANCE
height F 6	width file_size 48 1118	vector_size n 2064 81 8169 1	FE.890000 4F3.00090 791.890000 1422.0009	CLOSE SHOW DATABASE
Notes ID F	width file_size 45 \$118 45 \$214 MATCH MAGE ID	vector_rize n 2064 Ri 8160 L NEW IMAGE ID	52,50000 453,0000 791,500000 1422,0000 ×	CLOSE SHOW DATABASE ENTRE DISTANCE DELETE DISTANCE C/Decuments and
Index F6 C IMAGE ID 6 Extracted Features IMAGE DIMENSION HEIGHT 62	visith file_stize 48 1118 45 8214 MATCH IMAGE ID STORAGE DETAILS FILE SIZE COprised C	NEW EMAGE ID	EL50000 453,0000 171,00000 142,0000 3 DMAGELOCATION	CLOSE SROW DALLBASE ENTER DISTANCE DELITE DISTANCE C/Decuments and TRAIN ANN LAYER
Inicia           66           7           7           7           7           7 </td <td>vidi Ele,sia 4 111 4 1214 MATCH DAGE ID STORAGE DETALLS FILE SEZ VECTOR SEZ Cytical C</td> <td>NEW EMAGE ID</td> <td>52.00000 457.0000 171.00000 1422.0000 3 100.000 1422.0000 3 100.000 1422.0000</td> <td>CLOSE SHOW DATABASE INTER INSTANCE DELITE INSTANCE Cribecuments and TRAIN ANN LAYER 1 2 3</td>	vidi Ele,sia 4 111 4 1214 MATCH DAGE ID STORAGE DETALLS FILE SEZ VECTOR SEZ Cytical C	NEW EMAGE ID	52.00000 457.0000 171.00000 1422.0000 3 100.000 1422.0000 3 100.000 1422.0000	CLOSE SHOW DATABASE INTER INSTANCE DELITE INSTANCE Cribecuments and TRAIN ANN LAYER 1 2 3
IMAGE ID 6 Extracted Features IMAGE DIMENSION HEIGHT 62 SPATIAL DETAILS L.1.	vidi Ele,sia 4 111 4 1214 MATCH DAGE ID STORAGE DETALLS FILE SEZ VECTOR SEZ Cytical C	Vector_size in 1964 Bi 169 I NEW IMAGE ID HISTOGRAM Astracter_Recognition	EL50000 453,0000 171,00000 142,0000 3 DMAGELOCATION	CLOSE SROW DALLBASE ENTER DISTANCE DELITE DISTANCE C/Decuments and TRAIN ANN LAYER
Midda           65           68           69           8           60           8	vidik 6Le,tie d 1118 d 214 STORAGE DETAILS FILE SZZ CONVERT VECTOR SZZ CONVERT UL2 L3 SEARDIN MATCH M	vecke, rise P Bi-1 III Bi-3 III NEW IMAGE ID HISTOGRAM Haracter, Recognition Can not be Recognition Can not be Recognition START VALUE 0.00	ELEIGEN 1751.50000 H22.5000 DKAGELOCATION MATCH N 55267 MATCH N	CLOSE SHOW DATABASE SHOW DATABASE SHOW DATABASE DITE POSTANCE CODUCT POSTANCE CODUCT POSTANCE CODUCT POSTANCE CODUCT POSTANCE CODUCT POSTANCE SHOW DATABASE VEGENT VALUE Weight Spatial 0
Inidia           6           4           4           6           5           1000 Cell (Control on Marcino)           Extracted Features           IMAGE ID (F           IMAGE ID (F           Extracted Features           IMAGE ID (F           IMAGE ID (F  <	vidita Elevitive 4 TIII 4 Elevitive MATCH INAGE ID STORAGE DE TAILS FILE SIZE VECTOR SIZE Convater L2 L3 SEARCH	VECKE, LISE 0 06-1 III 16-0 II 16-0 III 16-0 III 16-0 III 16-0 III 16-0 III 16-0 IIII 16-0 III 16-0 II	SEASONO 45.0000 102.0000 102.000 30 MAGELOCATION MAGELOCATION MAGELOCATION 52557 MAGEN N 85257 MAGEN N	CLOSE SROW DALABASE ENTRE ROSLANCE CODECEMENTS and TRAIN ANN LAYER 1 [2] 3 WEIGHT VALUE weigth_spatial
Midda           65           68           69           8           60           8	vidik 6Le,tie d 1118 d 214 STORAGE DETAILS FILE SZZ CONVERT VECTOR SZZ CONVERT UL2 L3 SEARDIN MATCH M	Vector, size a Biol and a Biol and a Biol and a NEW IMAGE ID NEW IMAGE	EXAMPLE FX.10000 INL.00000 INL.00000 INL.00000 INL.00000 INL.00000 INL.00000 I	CLOSE SROW DATABASE SROW DATABASE DUTE DISTANCE COUNT DISTANCE COU
Inida           64	viditi GLo, rice a 111 a 111 MATCH INAGE ID STOTAGE DE TALS: FILE SEX VECTOR SEX L2 L3 SEARCH MATCH N MATCH N	Vector, size of BG41 III BG64 III BG64 III BG64 III BG64 III BG64 III BG64 III BG64 IIII BG64 IIII BG64 IIII BG64 IIIII BG64 IIIII BG64 IIIIII BG64 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	EXAMPLE FX.10000 INL.00000 INL.00000 INL.00000 INL.00000 INL.00000 INL.00000 I	CLOSE SROW DATABASE SROW DATABASE SROW DATABASE DELITE DOSTANCE C/Decements and TRANN ANN LAVER TRANN ANN LAVER TRANN ANN LAVER TRANN ANN UNTER TRANN ANN UNTER TRANN ANN UNTER TRANN ANN UNTER TRANN TRAN
Inicia           #4	vidik 6La, sia e 1118 e 1118	Vector, size a Biol and a Biol and a Biol and a NEW IMAGE ID NEW IMAGE	EXAMPLE FX.10000 INL.00000 INL.00000 INL.00000 INL.00000 INL.00000 INL.00000 I	CLOSE STOW DATABASE STOW DATABASE RUTR RISTANCE Critectorist and TRAIN ANN LAYER 1 2 3 WEIGHT VALUE weight_spatial 0 weight_controld 0 BIASE VALUE 0
Midd           6           4           4           4           5           5           5           6           7           7           7           8           6           6           6           7           7           7           7           8           7           8           8           8           8           8           8           8           8           8           8           8           8           8	viela Ele,cite ele 111 ele 1121 1	WERE JOS OF THE SECOND	ELEMENT     ELEMENT     ELEMENT     ELEMENT      MARCELOCATION      MARCEN     MARCEN     N      MARCEN	CLOSE SHOW DATABARE EXTER INSTANCE Crifferenses and TRAIN ANN LAVER 1 2 3 WEIGHT VALUE weight_spatial 0 weight_spatial 0 BLOSE VALUE 0
Inida           64	vidik 6La, ite 4 TIL 4 TIL 5 TORAGE DETALLS FILL SEX Contact ID 12 L 3 SEARCH MATCH N MATCH N MATCH N MATCH N MATCH N	vector, size a 1964 III 1964 III 1964 III 1965 III HISTOGRAM IIII HISTOGRAM IIIII HISTOGRAM IIIIII HISTOGRAM IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	ELEMENT 45.0000 31.0000 122.000 DALECLOCATION DA	CLOSE STOW DATABASE STOW DATABASE WITE RISTANCE C./DECEMBER OF LAYER 1 2 3 WEIGHT VALUE WEIGH_Spatial 0 weight_centroid 0 BASE VALUE 0 CLOSE
Inida           64	vidik 6Lo, rio 4 011 5107AGE DETALS FILE SEX OPTION VECTOR SEX 2 L 3 SEACE NUTCH N NATCH N NATCH N NATCH N NATCH N NATCH N NATCH N	Vector, size of The sector of the sector o	ELEMENT 45.0000 31.0000 122.000 DALECLOCATION DA	CLOSE SHOW DATABARE EXTER INSTANCE Crifferenses and TRAIN ANN LAVER 1 2 3 WEIGHT VALUE weight_spatial 0 weight_spatial 0 BLOSE VALUE 0
Inida           Idea	vidik Ek., ite G Eli T214 STOTAGE DE TALS FILE SIZE FILE SIZE FILE SIZE STOTAGE DE TALS FILE SIZE FILE SIZE STOTAGE DE TALS FILE SIZE STOTAGE DE TALS STOTAGE DE TA	vector, size a 1964 1 11	ELEMENT 45.000 7.Leven 122.000 BASCELOCATION MATCH N 52257 MATCH N 52257 MAT	CLOSE SHOW DATABASE STREESTACCE Criteconnes and LAYER I 2 3 VERGHT VALUE Weight_spatial 0 weight_centroid 0 BIASE VALUE 0 CLOSE SHOW DATABASE
■ Midd           ●           ●           ●           ●           ■           ●           ■           ●           ■ </td <td>vidik 6Lo, rice 4 TII 5 TORAGE DETALLS FILE SZZ OPTICAL VECTOR SZZ OPTICAL MATCH N MATCH N MATC</td> <td>vector, size a 1964 1 11</td> <td>ELEMENT 45.000 7.Leven 122.000 BASCELOCATION MATCH N 52257 MATCH N 52257 MAT</td> <td>CLOSE SHOW DALAGAR SHIR RISIANCE CONTRESSIANCE CONTRESSIANCE CONTRESSIANCE CONTRESSIANCE CONTRESSIANCE D SHOW DALAGAR D SHOW DALAGAR D DIR DISIANCE</td>	vidik 6Lo, rice 4 TII 5 TORAGE DETALLS FILE SZZ OPTICAL VECTOR SZZ OPTICAL MATCH N MATCH N MATC	vector, size a 1964 1 11	ELEMENT 45.000 7.Leven 122.000 BASCELOCATION MATCH N 52257 MATCH N 52257 MAT	CLOSE SHOW DALAGAR SHIR RISIANCE CONTRESSIANCE CONTRESSIANCE CONTRESSIANCE CONTRESSIANCE CONTRESSIANCE D SHOW DALAGAR D SHOW DALAGAR D DIR DISIANCE
Inida           Idea	vidik 6Lo, rio 4 011 5107AGE DETALLS FILLS SEX FILLS SEX ANTCH N MATCH N MATC	vector, size a 1964 1 11	ELGOBER STATES DESCELOCADON DESCELOCADON BLOELOCADON BLOELOCADON	CLOSE STOW DATABASE STOW DATABASE WITE SYSTANCE Criberansis and TRAIN ANN LAYER 1 2 3 WEIGHT AALUE weigh_centroid 0 weigh_centroid 0 BIASE VALUE 0 CLOSE SIGW DATABASE DETER DYSTANCE
Inida           Idia	vidik 6Lo, rio 4 011 5107AGE DETALS: FILE SZZ (Optical VECTOR SZZ (Optical VECTOR SZZ (Optical NATCH N NATCH N	VICE LA CONTRACTOR DE LA CONTRACTÓR DE L	ELEBORIO 45.000 71.0000 122.000 DESCELOCATION MATCH N 52267 MATCH N 52267 MA	CLOSE SHOW DATABASE EXTER ENSIANCE Criberanes and LAVER 1 2 3 WEIGHT VALUE weigh_spatial 0 weigh_centroid 0 BUSSE VALUE 0 CLOSE SHOW DATABASE CLOSE SHOW DATABASE FICE PRIAME DITER ENSIANCE Criberanes and I 2 3
Inida           Inida </td <td>vidik 6Lo, rio 4 011 5107AGE DETALLS FILLS SEX FILLS SEX ANTCH N MATCH N MATC</td> <td>VICELL SEA OF THE SECOND SECON</td> <td>ELEMENTE AFJ. 2009 PLANON I AZJONA DALACELOCATION MACCELOCATION MACON N MACON N</td> <td>CLOSE STOW DATABASE STOW DATABASE STOW DATABASE Criberansis and CATE RISTANCE Criberansis and CLOSE STOW DATABASE CLOSE STOW DALBASE CLOSE STOW DALBASE STOW DALBASE STOW DALBASE STOW DALBASE CLOSE STOW DALBASE CLOSE STOW DALBASE STOW DALBASE</td>	vidik 6Lo, rio 4 011 5107AGE DETALLS FILLS SEX FILLS SEX ANTCH N MATCH N MATC	VICELL SEA OF THE SECOND SECON	ELEMENTE AFJ. 2009 PLANON I AZJONA DALACELOCATION MACCELOCATION MACON N MACON N	CLOSE STOW DATABASE STOW DATABASE STOW DATABASE Criberansis and CATE RISTANCE Criberansis and CLOSE STOW DATABASE CLOSE STOW DALBASE CLOSE STOW DALBASE STOW DALBASE STOW DALBASE STOW DALBASE CLOSE STOW DALBASE CLOSE STOW DALBASE STOW DALBASE
Inicia           In	with         Be,rise           9         913           9         914           9         924           1214         924           1214         924           10000000         1214           10000000         1214           10000000         1214           10000000         1214           10000000         1214           10000000000         1214           10000000000000         1214           1000000000000000000000000000000000000	VICELL SEA OF THE SECOND SECON	ELEBORIO 45.000 71.0000 122.000 DESCELOCATION MATCH N 52267 MATCH N 52267 MA	CLOSE SHOW DATABASE RYTER RISIANCE C/GUCKEN
Inicial           Inicial           Inicial           Inicial           Extracted Features           Inicial           Initial	vidik 6Lo, rio 4 011 5107AGE DETAILS FILE SZZ Cystan KATCH INAGE ID STORAGE DETAILS RES SZZ SZI STORAGE DETAILS RES SZZ SZI SZI STORAGE DETAILS RES SZZ SZI SZI SZI SZI SZI SZI SZI SZI	VICELLA LA CONTRACTOR	ELEMENT IL STAND FLANDON IL IZZAND DESCELOCATION MATCH N STATU MATCH N STATU MATCH N STATU MATCH N STATU MATCH N STATU MATCH N STATU STATU MATCH N STATU	CLOSE STOW PATABASE STOW PATABASE Criberansis and CATE PSTANCE Criberansis and LAVER 1 2 3 WEIGHT VALUE Weight_controld 0 BASE VALUE 0 CLOSE STOW PATABASE PDTDE PSTANCE CLOSE STOW PATABASE PDTDE PSTANCE CLOSE STOW PATABASE PETTE PSTANCE CLOSE STOW PATABASE PETTE PSTANCE
bidd           64           6           64           64           64           62           BAGE ID 6           Extracted Features MAGE DIALFACEL 1995           SPATIAL DETAILS L1           TOTAL POEL 1995           ROW POEL 1995           CULTPOEL 1995           CULTPOEL 1995           CULTPOEL 6474           MAGE ID 5           Extended Features WIDTH 43           SPATIAL DETAILS L1           TOTAL POEL 3038           ROW POEL 6085.           COL POEL 6474.           COL POEL 6474.           COL POEL 6474.           COL POEL 6474.	with         Be,ris           III         111           III         1121           III         1214           IIII         1214           IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	VICELLA LA CONTRACTOR	ELEMENT     ELEMENT     ELEMENT     ELEMENT      ELE	CLOSE SHOW DATABASE RYTER RISIANCE C/GUCHEN

Fig-04 (Screenshots of Oriya OCR being developed)

Oriya alphabets are complex in nature and consist of 115 alphabets among which around 60 characters are very

difficult to recognize because they are composite characters (conjuncts). Fixed font style and font size are the biggest problem of OCR.

Of even greater concern is the problem of misreading a character (substitutions). In particular, if the system does not accurately balance dollar data, customer dissatisfaction will occur. The success of any OCR device to read accurately without substitutions is not the sole responsibility of the hardware manufacturer. Much depends on the quality of the items to be processed.

However, these limits are not objectionable to most applications, and dedicated users of OCR systems are growing each year. But the ability to read a special character is not, by itself, sufficient to create a successful system.

#### ACKNOWLEDGEMENT

The authors sincerely thank DIT,MCIT, Government of India for providing financial assistance for this work.

#### REFERENCES

- [1] Digital Image Processing, by Rafael C. Gonzalez, Richard E. Woods.
- [2] A Complete Development System for Oriya Script, Sanghamitra Mohanty, Hemanta Kumar Behera RC-ILTS-Oriya,Dept. of Comp. Sc. And Appl.,Utkal University,Bhubaneswar
- [3] S.Mohanty, K.Sahoo and H. K. Behera," A New Algorithm for the restoration of characters in old noisy document with varying level of intensities", ISC Conference, India, Jan'2003.
- [4] S. Kahan, T. Pavlidis and H. S. Bairb, "On the Recognition of Printed characters of any font and size", IEEE Trans. Pattern Analysis Machine Intelligence, Vol-9, 1987, pp. 274 –287.
- [5] LTG (Language Technologies Group), 2003. Optical Character Recognition for Printed Kannada Text Documents. SERC, IISc Bangalore. VijayaKumar, B., 2001.
- [6] Gonzalez, R.C., Woods, R.E., Eddins, S.L., 2004. Digital Image Processing Using MATLAB. PHI Pearson. Unicode, 2000. The Unicode Standard Version 3.0. Addison Wesley.
- [7] Rajavelu, M. T. Muaavi, and M. V. Shirvaikar, "A neural network approach to character recognition," *Neuml Networks*, Vol. 2, No. 5, 1989, pp. 387-389.

International Journal of Computer and Communication\_Technology (IJCCT), ISSN: 2231-0371, Vol-3, Iss-1

- [8] Chaudhuri, B.B., Pal, U., "A complete Bangla OCR system", Pattern Recognition, Vol. 31, pp. 531–549, 1998.
- [9] Complete OCR for Printed Hindi Text in Devnagari Script", Sixth International Conference on Document Analysis and Recognition, IEEE Publication, Seatle USA, 2001. Page(s):800-804.
- [10] Nallasamy Mani and Bala Srinivasan, "Application of Artificial Network Model for Optical Character Recognition", System, *Man and Cybernetics*, 1997, "Computational Cybernetics and Simulation". *International Conference* on 12-15 Oct. 1997 page(s): 2517-25203.
- [11] Veena Bansal and R.M.K. Sinha, "A Devnagari OCR and A Brief Overview of OCR for Indian Script", *PROC Symposium on Transaction support System* (STRANS 2001), Feb. 15-17, 2001, Kanpur, India.
- [12] Bansal, V., Sinha, R.M.K., "Partitioning and Searching Dictionary for Correction of Optically Read Devnagari Character Strings", *Document Analysis and Recognition*, 1999. ICDAR'99, Proceedings of the Fifth International.
- [13] A. A. Chaudhary, E.A.S. Ahmad, S. Hossain, C. M.Rahman, "OCR of Bangla Character Using Neural Network: A better Approach", 2nd International Conference on Electrical Engineering (ICEE 2002), khuln, Bangladesh.
- [14] Grain U and Choudhuri B B 1998 compound character recognition by run number based metric distance. *Proc. SPIE Annual Symposium on Electronic Imaging, San Jose, USA, pp 90-97.*
- [15] S. Kahan, T. Pavlidis and H. S. Bairb, "On the Recognition of Printed characters of any font and size", IEEE Trans. Pattern Analysis Machine Intelligence, Vol-9, 1987, pp. 274 -287

37