6.1 Optical Character Recognition Systems in Indian Languages

6.1.1

- **Name of the Technology** : ***Optical Recognition for Assamese.***
- **Nature of Technology** : Human Machine Interface System
- **Level: (Product / Technology / Sub-system)**: Software product.
- **Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product:** Optical Character Recognition (OCR) is the process of converting scanned images containing texts into a computer processable format (such as ASCII, ISCII, and UNICODE etc.). A MoU to transfer the OCR technology from the Resource Centre at the Indian Statistical Institute, Kolkata was signed in August 2002 and accordingly the same was effected in September 2002.

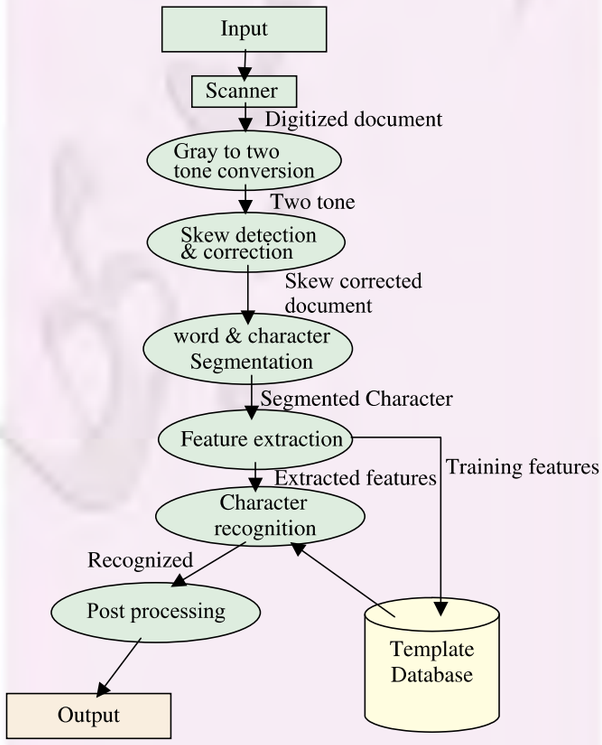
Technical Description

The system works with a gray level (8bit) TIFF (Tagged Image file format) image. The input image scanned either from books or from paper documents printed in Assamese script is then processed by OCR software and all the text from that document available inside the computer just as if it was typed in. The System produces the output in the ISCII format and output can be viewed or edited with any editor that supports ISCII. Some of the potential applications of OCR system are Newspaper (printed in Assamese) Houses, Libraries, Office automation, Linguistic Community (for creating corpus), reading aid for the Blind people, etc.

Features of the OCR System are:

Features	Assamese OCR
Scanning resolution	300-600 dpi
Input image	TIFF(8 bit gray scale)
Skew detection & correction	+5 to -5 degrees
Fonts name	Gitanjalilight, Luit & AS-TTDurga
Font size	12-28 points
Test data size	600 pages from books & printed documents
Post processing	Dictionary search & Morphological analyzer
Output file format:	ISCII format
Accuracy	97%

Basic block diagram of the system



Algorithms used in the different blocks of the system are

i. Gray to two tone conversion: The Input gray image is converted into binary image using the histogram based global threshold approach.

ii. Skew detection: A new approach proposed by Chaudhary and Pal that employs the headline of Assamese script. The method works by finding the headline of the document image. The advantage of the method is that the headline

information can be used for the detection of upper zone, character segmentation etc. To detect the headline, the large connected components of the document are found by component labeling.

iii. Word & character segmentation: Words are segmented using the connected component analysis with block absorption and block merging. For segmenting individual character from a word, we delete the headline from the word. Since characters in a word are usually connected to each other through the headline, they gets disconnected from each other when the headline is deleted.

iv. Feature extraction & character recognition: In this system two type of classifiers are used. The classifier one is designed to separate the basic characters, modifiers and compound characters. In the system, the middle and bottom portion characters are recognized as single units. The character portion in the top portions are recognized separately. Components of Individual characters are combined with respect to the positional information of their corresponding bounding box.

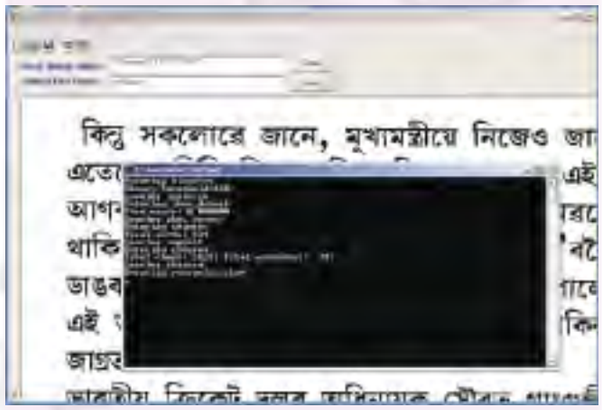
* Classifier-1: This classifier detects simple stroke features like vertical and horizontal lines, if such features are present. Characters like punctuation marks, special marks likes quotes are recognized by this classifier. The characters for which classifier-1 is not confident enough are passed on to the second classifier.

* Classifier-2: This classifier extracts some features from individual characters by counting the horizontal and vertical black runs (i.e. a black run is a set of black pixel with white pixel in either end).

In the classification phase, features vector of an unknown character are computed and matched with the stored prototypes. Matching is done by calculating distance (dissimilarity) measure between character and stored prototypes.

v. Postprocessing: In the postprocessing, error detection and correction is done at word level. There are two types of word errors, non-word and real word error. Dictionary match and Morphological analyzer are used to select the correct word from the set of alternatives that are generated during the recognition phase.

- **Operating System used**: The system is developed for both Windows and Linux.
- **Graphical User Interface**: The GUI of the System for Windows is developed using Microsoft Visual Basic that support TIFF image viewer and Scanner.
- **Representative snapshots / Screenshots of the Technology / Product**:



- **Scalability / Portability / Expandability** : The System is developed using C programming language. The Technology can be used with any platform including WINDOWS, UNIX and LINUX since the program is totally platform independent. The OCR software can be integrated with a Speech synthesis System to make a Text to Speech in Assamese.
- **Readiness of Transfer of Technology (ToT)**: Installation guidelines and User guides to use and configure the system are being prepared.
- **Availability of documentation**: Yes
- **Testing of the Product / Technology**: The results of the Alpha testing says that the performance of the OCR is up to an acceptable level. The system is able to take care of document skew in between (+5) degree to (-5) degree. Documents printed with mostly used fonts are well recognized. Variation in font size and style (except Italic) do not affect recognition accuracy.

- **IPR / Open-source** : Not yet applied.
- **Potential beneficiaries:** Libraries, DTP Operators, Offices looking for Office Automation, Linguistic Community, etc.
- **User-agency tie-up** : None.
- **Name and address of the Resource Person:**

S.V.Rao svrao@iitg.ernet.in
S.K.Deka sdeka@iitg.ernet.in
Department of Computer Science & Engineering,
Indian Institute of Technology Guwahati
Assam – 781039 (India)
Web : http://www.iitg.ernet.in/rcilts

6.1.2

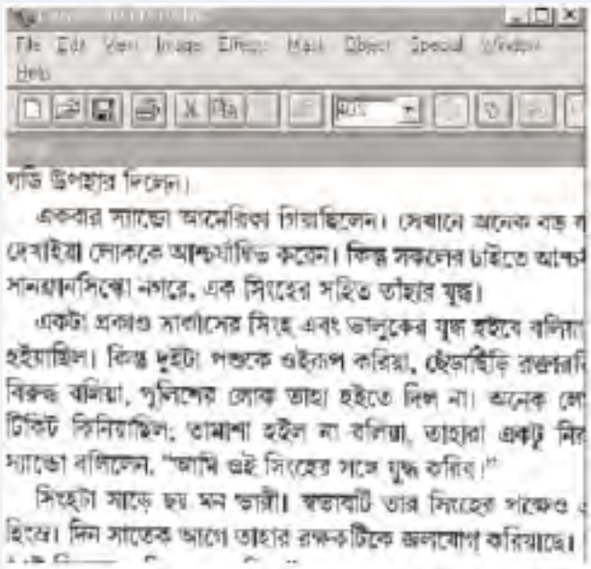
- **Name of the Technology :** ***Bangla Optical Character Recognition***
- **Nature of Technology :** Human Machine Interface, System
- **Level:** (Product / Technology / Sub-system) technology
- **Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product :** Bangla OCR System converts printed Bangla text into electronic version. It has many application potentials like reading aid for the blind (OCR and speech synthesis), automatic text entry into the computer, desktop publication, library cataloging and ledgering, automatic reading for sorting of postal mail, bank cheques and other documents. Also, document data compression is possible by converting the document image into ASCII format.

  Performance Summary: It deals with single column text image only. It can also handle small amount of skew in the range of –5 to + 5 degree. The system has been extensively tested on desktop and old published documents (more than thousand pages) of various font-types and sizes. The average character level accuracy is about 96% and word level accuracy is about 86%.

  Supporting platform(s): Linux, UNIX, DOS, MS-Windows.

  Memory requirements: 16MB RAM.
- **Representative Snapshot / screenshot of the Technology / Product:** Screenshot of Bangla OCR:

The Image:



The Corresponding Ground Truth:

- **Scalability / Portability / Expandability: Portability:** The system is implemented using ANSI C and hence portable to any platform.
- **Readiness of Transfer of Technology (ToT) :** Yes
- **Availability of documentation :** Yes
- **Testing of the Product / Technology:** Testing done
- **IPR / Open-source :** Not open source
- **Potential Beneficiaries:** Typists, Linguists, Students, Teachers, Educationists, Office Personnel, Corpus Developers etc.
- **User–agency tie-up :** Technology transferred to C-DAC, Pune for productization.
- **Name and address of the Resource Person:**

B. B. Chaudhuri,
Professor & Head,
Computer Vision & Pattern Recognition Unit
Indian Statistical Institute
203 Barrackpore Trunk Road
Calcutta 700 108, INDIA
E-mail : bbc@isical.ac.in
Web : http://www.isical.ac.in/rc-bangla

6.1.3

- **Name of the Technology :** ***Chitraksharika Optical Character Recognition for Devanagari Script***
- **Nature of Technology :** Human Machine Interface System (Recognition Software for Devnagari Script)
- **Level:** (Product / Technology / Sub-system) Technology
- **Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product :** An efficient way to convert optically scanned images of printed materials into computer processable data files. The technical attributes are:

  (i) Based on technology by ISI Kolkata.

  (ii) Recognizes Hindi, Marathi & Nepali.

  (iii) Scanning images via TWAIN interface.

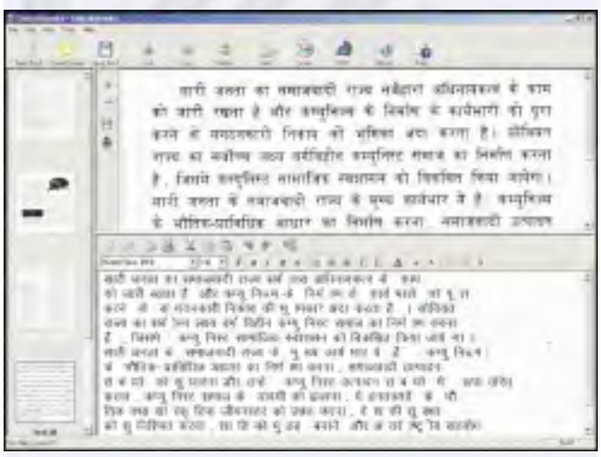
  (iv) Auto Image segmentation, de-skewing, detection of text, table & pictures

  (v) Image editing features.

  (vi) Embedded spell checker for Hindi.

  (vii)text & scanned images in different formats.

- **Representative Snapshot / screenshot of the Technology / Product:**



- **Scalability / Portability / Expandability: Portability:** The system is implemented using ANSI C and hence portable to any platform.
- **Readiness of Transfer of Technology (ToT) :** Yes
- **Availability of documentation :** Yes
- **Testing of the Product / Technology :** testing done by STQC ; Result – Character Recognition accuracy 96.8%
- **IPR / Open-source :** Technology obtained from ISI-Kolkata
- **Potential Beneficiaries :** Typists, Linguists, Students, Teachers, Educationists, Office Personnel, Corpus Developers etc.
- **User–agency tie-up:** ——
- **Name and address of the Resource Person:**
  V N Shukla
  Dir(Spl App)
  C-DAC, C – 56/1, Sector 62
  Noida – 201 307,
  Phone : 0120 – 2402551...558
  E-mail : vnshukla@cdacnoida.com
  Web : http://www.cdacnoida.com

6.1.4

- **Name of the Technology :** ***Optical Character Recognition System for Malayalam. (NAYANA™)***
- **Nature of Technology :** Human Machine Interface System.
- **Level :** Product
- **Technical Description:**

The Malayalam OCR system converts scanned images of printed Malayalam documents to editable text. It is a multi font system that works across a range of font sizes. The system has a recognition speed of fifty characters per second.

The System consists of a preprocessing module, the OCR engine and a post processing module.

The block diagram of the system is given in figure 1.1

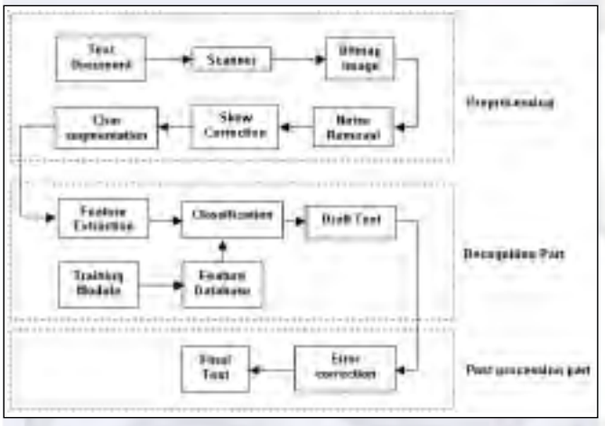


Figure 1.1 Block diagram of Malayalam OCR System

The preprocessing tasks performed by the first module include noise removal, conversion of grey scale image to binary, skew detection and correction and line, word, and character segmentation. The scanned images in grey tone are converted into two-tone (binary) images using a histogram based thresholding approach (Otsu's algorithm). Skew detection is done using the Projection profile based technique. After estimating the skew angle the skew is corrected by rotating the image against the estimated skew angle.

The OCR engine (Character Recognition Module) is based on the **Feature Extraction** method of character recognition. Feature extraction can be considered as finding a set of vectors, which effectively represent the information content of a character. The features are selected in such a way that they help in discriminating between characters. A multistage classification procedure is used, which reduces the processing time while maintaining the accuracy.

After passing through different stages of the classifier, the character is identified and corresponding character code is assigned. A training module is incorporated in the OCR engine to recognize characters, which are different from normal characters in their shape and style(example- decorative fonts).

In the post processing module, Linguistic rules are applied to the recognised text to correct classification errors. For example, certain characters never occur at the beginning of a word and if found so, they are remapped appropriately. Similarly, dependent vowel signs can occur only with consonants or consonant conjuncts; if found along with vowels or soft consonants, they are remapped into consonants/conjuncts similar in shape to the vowel sign. Independent vowels occur only at the beginning of a word and if found anywhere else, they will be mapped into a consonant or ligature having similar shape.

Performance of the OCR

Developed on VC++ platform, our Malayalam OCR runs on Windows 98/2000. It recognises 50 characters per second and gives an accuracy of 97% for good quality printed documents. The specifications and performance of the system is given below.

Skew detection and correction : - 5 to +5 degree

Supported image formats : BMP, TIFF

Image scan resolution : 300dpi and above.

Document Type : Multiple font, Multiple size.

Supporting Fonts:

Fonts Names : CDAC Fonts (ML-TTKarthika, MLW TTKarthika), Mathrubhumi Font, Manorama Font, Fonts used by DC Books

Font Size : 12-20

Font Styles : Normal, BOLD

Supported Code format : ISCII/ISFOC/ UNICODE

Supported output format : RTF/HTML/ACI/ TXT

Character recognition accuracy (%)

Document Type	Good quality Paper	Bad quality Paper
Computer Printed Document	97%	94%
Magazine	92%	90%
Newspaper	85%	82%
Books	95%	93%

Table1

Extensive testing has been done on approximately 500 pages of different quality printed documents. Table1 consolidates the results of testing. The system has undergone certification testing at ETDC Chennai.

- **Representative Screen shot**

- **Portability/Expandability/Scalability:** Porting of the software to Linux platform is in progress.

  The present system recognises around 250 glyphs, which includes all the symbols in the new Malayalam script (around 90) and around 160 characters from the old script. The system can be expanded to cater to the full set of old script characters.

- **Readiness of Transfer of Technology (ToT):** The Product is ready for Transfer of Technology.
- **Availability of documentation:** Design /TOT documents available.
- **Testing of the Product/Technology**: Product tested at ETDC Chennai.
- **Potential beneficiaries**: Visually challenged population, Publishing Sector, Content Creators, Digital Library, Corpus Development etc.
- **User-agency tie-up:** Solicited
- **Name and address of the Resource Person:**

R. Ravindra Kumar
Additional Director & Coordinator
RCILTS-Malayalam,
C-DAC, Thiruvananthapuram.
Email: ravi@erdcitvm.org.
Web : http://www.malayalamresourcecentre.org

6.1.5

- **Name of the Technology : OCR-ORIYA (Optical Character Recognition for Oriya)**
- **Nature of the Technology:** Human Machine Interface System (Optical Character Recognition and NLP)
- **Level:** (Product / Technology / Sub-system): Technology and Product
- **Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product :** The product is an intelligent document processor and recognises the Oriya characters in Oriya document image. The system also provides an Oriya text editor for viewing the text output of the OCR. This Editor also provides post-processing facilities like word check and spell check. The technologies/ algorithms involved are:

  (i) Binarization using dynamic threshold

  (ii) Skew detection using repeated angular projection profile

  (iii) Line extraction using stripwise vertical histogram analysis

  (iv) Character extraction using region growing and labelling and *matra* extraction using region analysis

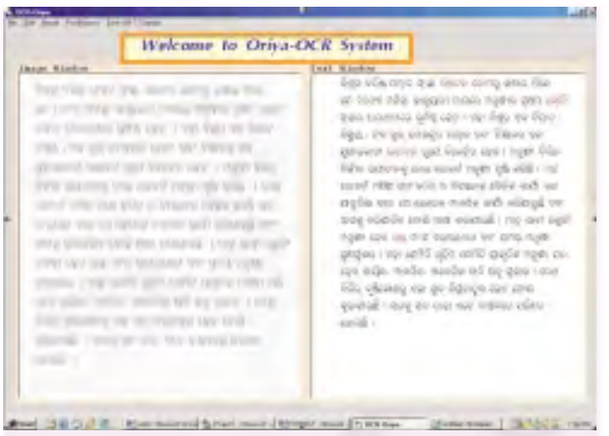
  (v) Connected component analysis by forward and backward chaining.

  (vi) Recognition by feature extraction and matching

  (vii) Post processing using a huge Oriya corpus.

  The language used for recognition engine is C/C++ and the front-end & the post processor uses VB 6.0.

- **Representative Snapshot / screenshot of the Technology / Product:**

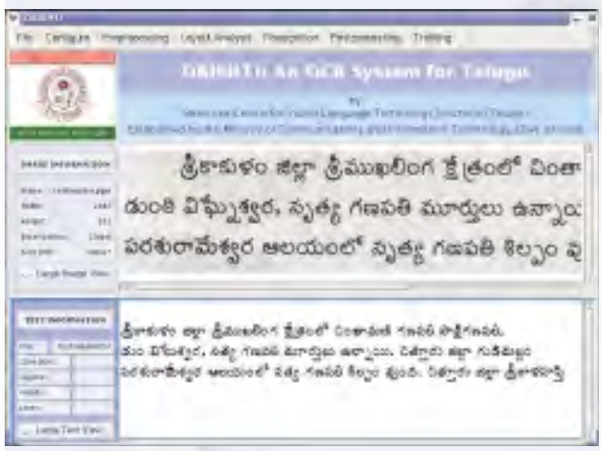


- **Scalability / Portability / Expandability :** All
- **Readiness of Transfer of Technology (ToT):** Ready
- **Availability of documentation :** Yes
- **Testing of the Product / Technology :** Under process
- IPR / Open-source : IPR: - SW-1174/2003
- Potential beneficiaries : Printing Press, Govt. Offices, Common people OCR integrated with TTS works a Reader-System for blinds and illiterates
- **User-agency tie-up:** Nil
- **Name and address of the Resource Person:**

Dr (Mrs.) Sanghamitra Mohanty
RC-ILTS-ORIYA, Deptt. of CSA,
Utkal University, Bhubaneswar, Orissa.
E-mail : sanghami@rediffmail.com
Web : http://www.ilts.utkal.org

6.1.6

- **Name of the Technology :** ***Optical Character Recognition for Oriya Telugu***
- **Nature of Technology** : Human Machine Interface System (DRISHTI is an optical character recognition system for Telugu. DRISHTI converts scanned images of texts into editable text in electronic form.)
- **Level of technology:** DRISHTI can currently handle printed texts only, not hand written manuscripts. It works across a range of fonts and font sizes. DRISHTI currently gives a character level accuracy of 95% to 97%. DRISHTI has undergone several levels of testing and evaluation both internally and through STQC unit of DIT.DRISHTI is in fact a suite of tools and modules that can be used to build OCR systems for any script. The system is designed so well that an basic OCR system for a new script can be developed in very little time.
- **Technical description:**



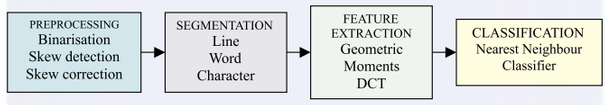
- **Representative snapshot:**
- **Scalability, Portability, expandability:** DRISHTI is written as a suite of C callable library modules. Standard C is used and so DRISHTI works on a wide variety of platforms including UNIX, Linux, and Microsoft Windows platforms. The design of DRISHTI permits it to be used very effectively as a research and exploration tool as also for developing working OCR systems for any language/script. An OCR system for the Gujarati script was developed in less than two days from scratch. DRISHTI has the core engine plus a wide variety of add-on modules. DRISHTI is exceptionally good in terms of portability, scalability and expandability.
- **Readiness for Transfer of Technology**: READY
- **Availability of documentation:** Extensive technical documentation as well as installation and user manuals have been prepared. Several research papers have been published.

  Testing:

  DRISHTI has been subjected to thorough testing and evaluation both in house and through STQC unit of Dept of Information Technology . Benchmarking standards for Indian Language OCR systems have been developed by us in collaboration with ISI (Kolkata). A Truthing tool has beendeveloped to enable large scale development of ground truth data.
- **IPR/Open-Source: To be decided**
- **Potential beneficiaries :** Corpus and content creators, archieval and preservation of treasured documents, government agencies etc.
- **User-agency tie-up:** A number of agencies have approached us.Negotiations are on.
- **Name and Address of resource persons:**
  Dr. K. Narayana Murthy, Dr. Atul Negi,
  Dr. Chakravarthy Bhagvati,
  Department of CIS, University of Hyderabad,
  Hyderabad - 500 046 AP, INDIA
  E-mail : knmcs@uohyd.ernet.in
  Web : http://www.languagetechnologies.ac.in

6.1.7

- **Name of Technology : *Tamil Gnaani (Optical Character Recognition for Tamil)***
- **Nature of Technology :** Human Machine Interface System
- **Level:** (Product/Technology/Sub-system) : Product
- **Technical Description :** A complete working model of Optical Character Recognizer for tamil script is developed. The system works in a multi-font and multi-size scenario. The input to the system is a scanned or digitized document and the output is in TAB code.



Block diagram of our OCR System

The product is developed using VC++ and takes about 24 Mb after installation.

- **Screenshot :** Not available
- **Scalability/Portability/Expandability :** Windows platform only
- **Readiness of Transfer of Technology :** Prototype
- **Testing of Product :** Tested by STQC with accuracy of 96 %
- **IPR / Open source :** Open source
- **Potential Beneficiaries :** Mass conversion of existing documents and literature into electronic format, reading aid for the blind as a part of text-to-speech converter, automatic sorting of mails in postal department and processing of bank documents, machine transliteration/translation of documents and literature of other scripts, etc.
- **User-agency tie-up :** NIL
- **Name and address of Resource Person :**
  Prof N.J.Rao,
  Department of Management Sciences,
  Indian Institute of Science, Banglore 560012
  E-mail : chairman@mgmt.iisc.ernet.in
  Web : http://www.iisc.ernet.in

6.1.8

- **Name of the Technology : *Optical Character Recognition for Gurumukhi***
- **Nature of Technology :** Human Machine Interface System (Printed Character Recognition Software for Gurumukhi script)
- **Level:** (Product / Technology / Sub-system) Technology – product could be readily developed form the technology
- **Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product: Gurumukhi OCR : Optical Character Recognition (OCR)** is the process whereby typed or printed pages can be scanned into computer systems, and their contents recognized and converted into machine readable code. The text, which the machine can read, has great advantages over that merely displayed as an image, since it can be edited, exported to other programs and indexed for retrieval on any word. If one needs to electronically store and manipulate large amounts of text or printed matter such as newspapers, contracts, letters, faxes, price lists or corpus development, one will find OCR programs can save lots of effort. For the first time a complete OCR package for Gurmukhi script has been developed. The OCR has a recognition accuracy of more than 97%. The OCR can automatically detect and correct skewed pages and can also detect page orientation. The OCR can recognize multiple Punjabi fonts and sizes. The main features of the Gurmukhi OCR are:

Recognition Accuracy

Recognition accuracy is around 97% for books, photocopies and medium degraded documents and around 99.5% for laser print outs and good quality documents.

Import Image

B/W Images, 24-bit color images, 256 grayscale images.

Output File Format

ISCII / text file encoded in any one of the popular Punjabi fonts.

Fonts Supported

All non-decorative Gurmukhi fonts.

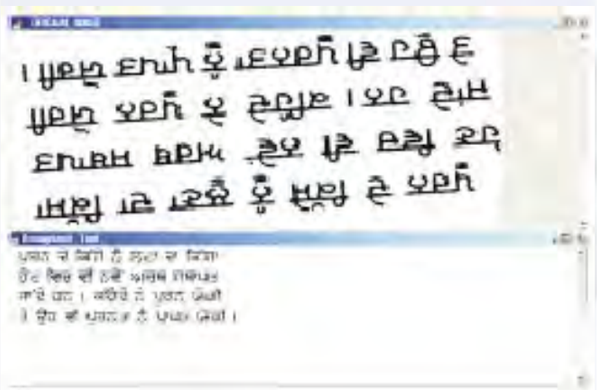
Proofing

On-screen verifier

Additional Features

Inbuilt Spell checking facility. Automatic skew detection and correction (skew range – 5 to +5 degrees). Upside down image auto detection and correction.

- **Representative Snapshot / screenshot of the Technology / Product :** Screenshot of Gurumukhi OCR:



- **Scalability / Portability / Expandability :** Presently developed for Windows operating system
- **Readiness of Transfer of Technology (ToT) :** Yes
- **Availability of documentation :** Yes
- **Testing of the Product / Technology :** Testing done
- **IPR / Open-source :** Not open source
- Potential Beneficiaries: Typists, Linguists, Students, Teachers, Educationists, Office Personnel, Corpus Developers etc.
- **User–agency tie-up :** Technology transferred to C-DAC, Pune for productization.
- **Name and address of the Resource Person:**
Dr. R.K.Shrama
Thapar Institute of Engineering Technology
Patiala
E-mail : rksharma@tiet.ac.in
Web : http://punjabirc.tiet.ac.in/rcentrei.asp