6. OCR Evaluation & Benchmarking

OCR Workshop on September 5-7, 2002 at ER&DCI, Noida

The OCR workshop was held at ER&DCI, Noida to review the status of development and our readiness to launch these OCRs. The following participated:

bbc@isical.ac.in

rmk@iitk.ac.in

veena@iitk.ac.in

yamini@iitk.ac.in

rchaurasiya@cse.iitk.ac.in

nitisha@cse.iitk.ac.in

jitesh@erdcitvm.org

ravi@erdictvm.org

gslehal@mailcity.com

svrao@iitg.ernet.in

aditya@cdacindia.com

knmcs@uohyd.ernet.in

chakcs@uohyd.ernet.in

atulcs@uohyd.ernet.in

pb@cse.iitb.ac.in

rkj@cse.iitb.ac.in

pallavi_satyam@hotmail.com

sangham1@rediffmail.com,

sangham1@sancharnet.in

akp@ocac.ernet.in, a.pujari@lycos.com

utpal@isical.ac.in

srmmsu@yahoo.com

jigneshmsu@yahoo.com

profvk@softhome.net

omvikas@mit.gov.in

pkc@mit.gov.in

slata@mit.gov.in

vkumar@mit.gov.in

mjain@mit.gov.in

vnshukla@erdcinoida.com

dkjain@erdcinoida.com

schandra@mit.gov.in

bksahu@erdcinoida.com

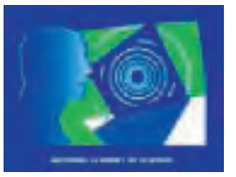
mncooper@modular-infotech.com

shyam_agarwal@indiatimes.com

Objectives

- To review the technical progress of OCR projects being implemented at various Resource Centres:

1.	Devanagari Script - (Hindi)	=	IIT Kanpur, ISI Kolkata, C-DAC, Pune
2.	Devanagari Script - (Marathi)	=	C-DAC, Pune
3.	Tamil	=	IISc Bangalore; Learn Fun Systems, Chennai* (Private developer)
4.	Gujarati	=	MS University, Baroda
5.	Bangla	=	ISI, Kolkata
6.	Assamese	=	IIT, Guwahati
7.	Oriya	=	OCAC, Bhubaneshwar
8.	Gurmukhi Script – (Punjabi)	=	TIET, Patiala
9.	Kannada	=	IISc, Bangalore
10.	Malayalam	=	ER&DCI, Trivendarum
11.	Telugu	=	University of Hyderabad



- To estimate efforts required to develop additional modules for supplementing the core OCR engine developed so far to provide total solution to the end user.
- Evolving bench marking parameters for testing of OCRs.
- Providing platform to OCR researchers to interact and share their teething problems in enhancing the accuracy of OCRs and exploring possibility of sharing algorithms to exploit commonality of methodology.
- To have a repository of OCR Software at ER&DCI, Noida for facilitating the implementation of parallel corpora project.

Methodology

- Presentation by OCR researchers on the status of technology development.
- Evolve a questionnaire to capture details of implementation of OCR technology at each centre.
- Debate on bench marking parameters evolved by ISI Kolkata for arriving at consensus.
- Demonstration of OCRs
- On the spot testing of OCRs against 10 page test data.
- Assess the level of OCR technology achieved at each centre.
- Identify languages for which OCRs have been developed with accuracy more than 95%.
- To identify the focus of RCs to refine the OCRs for delivery.

Proceedings

1. Long term objective

The ultimate objective for development of OCR technology aims at faster content creation in Indian languages and following type of documents need to be handled so that archived data can also be processed.

- Hand written multilingual content on papers, leaves, cloth etc.
- Old content printed during last 20 years using earlier technologies.
- Recently developed content using electronic publishing tools.

This is an ambitious target and would require strengthening of the basic algorithms for optical character recognition. In this process, the developers may have to adopt alternative strategies and devise new algorithms for handling archived data and hand written content based on their experiences for handling the electronically published data.

2. Current Focus

- The OCR software can be broadly categorized into three modules:
  - Document analysis
  - Character recognition engine
  - Document synthesis
- There was a general agreement to currently focus on font sizes 12-36. Auto font detection is also a desirable feature which needs to be implemented.
- Most of the developers have implemented skew detection and correction modules.
- Background noise removal algorithms have also been implemented.
- Some of them have also developed algorithms for auto inverted document recognition.
- The spell checker is also being integrated for post processing.

- The scripts in which OCR has been developed with more than 95% accuracy - Bangla, Devanagari, Gurmukhi, Telugu and Tamil.
- Scripts in which OCR with accuracy between 90-95% - Kannada, Oriya, Malayalam.

- Benchmarking

  Based on the presentation made by ISI Kolkata and University of Hyderabad, the following was discussed

  - In addition to the image file, the corresponding attribute file of the document being OCRed and the ground truth file also need to be generated. The format for attribute file and ground truth file was discussed and RCs were requested to give their feedback.
  - STQC briefed the RCs about the software test facilities available at their labs in Delhi and Chennai/Bangalore.
  - The consolidated scanned results of 10 page test data conducted during the workshop are given in the table–Gist of Consolidation of OCR Systems.

STQC has been assigned the responsibility of evaluation of OCRs, MAT systems, Spell Checkers etc. for which RCs have been requested to provide inputs and cooperate in the process of evaluation.

OCR Systems for Gurmukhi, Bangla, Tamil, Kannada, Marathi, Malayalam, Telugu and Devanagari have been tested by STQC.

(Courtesy : Sh. V.N. Shukla,
(Dir. Special Applications), ER&DCI,
Noida Tel. : 95120-2402551
e-mail : shuklavn@hotmail.com)

SĀDHĀNĀ - ***Special Issue on OCR***

The Indian Academy of Sciences Bangalore publishes a journal titled **SĀDHĀNĀ** (Website : http:// www.ias.ac.in/sadhana, E-mail: sadhana@ias.ernet.in, Tel: (080) 3612546), which covers Academy Proceedings in Engineering Sciences. Vol. 27 part 1 February 2002 issue of **SĀDHĀNĀ** covers papers in the area of Indian Language Document Analysis & Understanding.

There are eight papers in this special issue. The invited paper authored by Kasturi, O'Gorman and Govindaraju provides a good overview on document image analysis in an authoritative manner outlining all the issues involved. The next two papers deal with complete systems designed for processing printed text documents in a single language. The paper by Chaudhuri, Pal and Mitra, which is also an invited contribution, describes a system for recognition of printed Oriya script. The paper by Ashwin and Sastry describes a system for analysing printed Kannada documents using SVMs for pattern classification. The next paper by Bajaj, Dey and Chaudhury, which is also an invited contribution, describes a method of combining multiple neural network classifiers for robust recognition of handwritten Devanagari numerals.

The next two papers deal with script identification in multilingual documents. The paper by Dhanya, Ramakrishnan and Pati describes a system for identifying the script at each word level in bilingual documents containing Roman and Tamil scripts. The paper by Subba Reddy and Patil describes a neural network based recogniser for identifying the script in a document containing Roman. Devanagari and Kannada scripts. The paper by Lehal and Singh describes a post-processing system for Gurumukhi.

The final paper in this special issue, authored by Sen and Samudravijaya, is somewhat different in the sense that it does not deal with document image analysis; it deals with one of the possible applications of the results of document image analysis and recognition. This paper describes a text-to-speech system that can read aloud a web document in Hindi or English.

...tdilteam

(Source : SĀDHĀNĀ, IASc Bangalore, Feb., 2002)

GIST OF CONSOLIDATION OF OCR SYSTEMS (as on September 2002)

Language	Sub Modules			Test Data Size	Font Name	Font Size(in Pts)	Efficiency	Possibility of Increasing Efficiency Up To	Method for Improvement
	Document Analysis	Character Recognition Engine	Document Synthesis						
Hindi	Exist for text and images	Exist	Exist for text and images	Sample Size of 1000 pages	Major Fonts Including Digital ones used for Printing Hindi Materials DVB- TT Yogesh DVB-TT Surekh, Metal Type	12 to 36 11 to 32 11 to 32 Not Mentioned	96%-98%		Spell Checker, Algorithm Improvement, Exhaustive Testing
Marathi	Exist for text and Images	Exist	Exist for text and Images	Not Available	DVB- TT Yogesh DVB-TT Surekh, Metal Type	11 to 32 11 to 32 Not Mentioned	90% at word Level		
Punjabi provement	Not Exist	Exist	Not Exist	Sample size of 16 pages for each character fonts and 4 font sizes	Gurumukhi-IIYS, CDAC Font (PN-TT Amar), Punjabi, Primaja Anandpur Sahib	12-20, 14-24, 12-20, 12-20, 12-20, 12-20	97%		Spell Checker, Algorithm Im- Exhaustive Testing
Telugu	Not Exist	Exist	Not Exist	Tested on 80-90	Hemalatha Harshapriya SreeLipi Ann font Family	12,18,20,22,24 16,18,20 14 to 20 14	95-97%		
Tamil	Exist for multi column text.	Exist	Not Exist	Tested on 600 pages.	Any font	12 to 36	98%	99%	Spell Checker, Algorithm Improvement Exhaustive Testing

GIST OF CONSOLIDATION OF OCR SYSTEMS (as on September 2002)

Language	Sub Modules			Test Data Size	Font Name	Font Size(in Pts)	Efficiency	Possibility of Increasing Efficiency Up To	Method for Improvement
	Document Analysis	Character Recognition Engine	Document Synthesis						
Kannada	Not Exist	Exist	Not Exist	Not Mentioned	Multi Font	Multi Size	95-96 %	97%	Spell Checker, Algorithm Improvement, Exhaustive Testing
Malayalam	Not Exist	Exist	Not Exist	Not Mentioned	MLaw-TTKartika ML-TT Kartika, MalBrubhmi Manorma Fonts Current Books	12,14,16,18 12,14,16,18 12 to 16 12 to 16 12,14	90-92%		Spell Checker, Algorithm Improvement, Exhaustive Testing
Oriya	Not Exist	Exist	Not Exist	Sample size of about 100 pages	1. Modular Shree 2. C-DAC Font	14, 16	93%	97%	Spell Checker, Algorithm Improvement, Exhaustive Testing
Assamese	Not Exist	Exist	Not Exist	Sample size of about 100 pages	Ratnagiri	12 to 36	95%	97%	Spell Checker, Algorithm Improvement, Exhaustive Testing
Bangla	Partially Exist	Exist	Partially Exist	Sample size of about 100 pages	Covers all the major fonts used for publishing Bangla Materials	12 to 36	97%		