Noolaham Foundation
Final Report

Project Title	Tamil Optical Character Recognition Support Project
Project Number	NF/PG/2013/0006
Project Locations	Jaffna, Colombo
Sector	Crafts, Science and Technology, Digital Library
Implementing agency and contribution	Noolaham Foundation
Grant Agency and Contribution	Noolaham Foundation
Total Budget and Expenditure	
Project Period	July 2013 – December 2013
Responsible Stakeholders	Department of Computer Science, University of Jaffna, Library, University of Jaffna (UOJ)

Summary

The **Tamil Optical Character Recognition Support Project** was aimed at providing scanned raw images of rare Tamil documents and to assist the Tamil OCR development project. Noolaham Foundation collaborated with department of Computer Science, University of Jaffna to implement this project. The Tamil digitization project is a joint venture of the Theekshana (School of Computing, University of Colombo) and the department of Computer Science, University of Jaffna and funded by ICTA. The main goal of the project is to develop the tools needed to automatically recognize the most common printed Tamil fonts from scanned images of books and documents for digitizing such content.

Noolaham Foundation provided scanned images of 51 rare documents to the department of Computer Science, University of Jaffna for training and testing the Tamil OCR system through this project. In the documents documented, 15 books (6,450 raw images) were already available at Noolaham Digital Archive. Another 36 documents were digitized specifically for this project with a view to using them for training and testing the Tamil OCR system which is being developed. All 51 documents are made available online through Noolaham Foundation's Digital Library www.noolaham.org. This project was a successful initiative and received special appreciation from the research community.

Introduction and Background

Tamil is a language spoken widely in Sri Lanka, South India and Diaspora. Tamil has the longest unbroken literary tradition amongst the Dravidian languages. The earliest available text on grammar is the Tholkaappiyam, a work describing the language of the classical period. There are several other famous works in Tamil like Kamba Ramayanam and Silapathigaram.

The information age has enabled dissemination of information sources through digital media. To do so, printed documents have to be converted to electronic versions.

Optical Character Recognition (OCR) deals with machine recognition of characters present in an input image obtained using scanning operation. The OCR software allows you to scan a printed document and then convert the electronic text in an editable format. But there was no such software available in the market which satisfactorily provided the required results in Tamil language. Tamil language consists of intricate characters. The need for OCR arises in the context of digitizing Tamil documents from the ancient and old era to the latest. The 'Tamil Optical Character Recognition Support project' tried to recognize characters of Tamil language, which helps in sharing the data through the Internet.

Therefore Noolaham Foundation initiated this project with the facilitation of the Department of Computer Science, University of Jaffna.

The implementation process

Selection criteria

Noolaham Foundation and the Department of Computer Science, UOJ spent a lot of time identifying the books and the sections and the list thus prepared has significant literary value. In addition, these documents cover a wide variety of books. Hence the pages can be used as a base for testing any of the Tamil OCR systems. The selected documents included

- Tamil books published in Sri Lanka or published by Sri Lankans elsewhere
- published before the era of desktop publishing - printed by letter press
- with significant literary value or historical value – such as books from 19th century , first prints of popular newspapers, magazines
- authored by a significant figure in the literary world

Using purposely built scanning device

During this digitization process, normal scanning apparatus and technologies such as desktop scanners, flatbed scanners, sheet fed scanners etc. were not used to digitize documents which were very fragile. Noolaham Foundation used an already built special scanning apparatus for scanning purposes. It was built with the technical guidance of the French Institute of Pondicherry.

Objectives and Achievements or Results

The objective of the project is to develop the tools needed to automatically recognize the most common printed Tamil fonts from scanned images of books and documents for digitizing such content.

51 rare and endangered books and various kinds of documents were digitized through this project. They are made available online through Noolaham Foundation's Digital Library (www.noolaham.org).

Constrains / Challenges:

- To get official permission from the library, University of Jaffna to implement this project was a big challenge. Initially most top level officials were not willing to provide the documents. Due to lack of awareness, the library officials negatively perceive this process, and fail to appreciate the actual value of the project.
- Some Tamil characters are taller than others. Some characters have closed loops.

Suggestions and Recommendations:

- More involvement on the part of the stake holders throughout the project process starting from planning will help towards the effective functioning of the project.
- Tamil OCR have been a frontline research area in the field of human-machine interface for the last few years. However, the best possible solutions need to be discussed. Noolaham Foundation should collaborate more and more with other organizations and groups in creating digital content.
- As we have recognized that these characters are based on images, this system can be used as an interactive system to support people who are physically challenged such as those who are visually impaired. . Modification is required to generate sound files of recognized characters so that a visually impaired person will be able to understand the data on the page.

  .

Relevant Attachments

News about discussion with Dr. E.Y. A. Charles, Head of Department of Computer Science, University of Jaffna on the 4th July 2013 at the Noolaham Foundation, Colombo.
http://noolahamfoundation.org/wiki/index.php?title=News/2013/2013.07.04

The details of provided existing documents

No	Description	Book title	Year of publication	No of pages provided	Selected Section	No pages Tested
1	Tamil literature and Grammar books	Tholkaappiyam Ezhuththathigaaram	1937	241	Annexed Articles	15
		Tholkaappiyam Sollathigaaram	1938	295	Annexed Articles	15
		Tholkaappiyam Porulathigaaram part 1	1948	456	Preface + Annexed Articles	30
		Tolkaappiyam Porulathigaaram part 2	1943	792		
		இலக்கண வினாவிடை Grammar Quizzes - Arumuka Navalar	1998	56	Full	56
		இலக்கணச் சுருக்கம் - Arumuka Navalar	1949	208+iv	Full	
2	Dictionaries	இலக்கண சந்திரிகை - குமாரசுவாமிப் புலவர்	2nd Ed. 1968	64	Full	64
		வினைப் பகுபத விளக்கம்- குமாரசுவாமிப் புலவர்	2nd Ed. 1967	44	Full	44
3	Religious books	மட்டுவில் திருவாதவூர் அடிகள்புராணவிருத்தி உரை-ம.க.வேற்பிள்ளை	1939	400	First 10 pages	10
4	Research books	யாழ்ப்பாணத்து ஓவியர்கள்- சோ.கிரு~;ணராஜா (Yalppanaththu Oviyarkal - S.Krishnarajah)	1997	72	First section	15
		தேடலும் படைப்பும் அ.மாற்கு (Thedalum Padaippum - A. Mark)	1987	135	First 10 pages	10
5	Newspapers and Magazines	உதயதாரகை (Uthayatharakai)			One article	
		பாதுகாவலன் (Pathukavalan)			One article	
		மறுமலர்ச்சி (Marumalarchi)			One article	
6	Novels	வெள்ளிப்பாதசரம்- இலஙிகையர்கோன் (Vellippathasaram - Illankayar kon)			First 10 pages	10

Relevant Photos

Discussion with Dr. E.Y. A. Charles