

MULTI-LINGUAL SCRIPT RECOGNISER

COLLEGE : KVG COLLEGE OF ENGINEERING, SULLIA
GUIDE : PROF. B. S. KEDILAYA
STUDENTS : APARNA N. S
ASHRITHA S
CHITHRAKALA G
PREETHI RAI P

Introduction

In multi-lingual documents the amount of multimedia data captured and stored is increasing rapidly with the advances in computer technology.. For example, museums store images of all old fragile documents having scientific or historical or artistic value and written in different scripts which are stored in typically large database.

Script identification is key step that arises in document image analysis especially when the environment is multi-script and multi-lingual.

The salient features of this project is to design a software that can read script written in most of the INDIAN languages like Hindi, Kannada, English, etc.. Usually such type of projects undergoes the OCR problem. OCR is a type of software designed to translate images of text into machine editable text. Each OCR translates text for particular language only .So for using multi-lingual text identify different scripts and extract parts of same script so as to feed into particular OCR designed for that script.

In this project first the images are scanned. The image will be normalized, segmented and then feature extraction is done and those features will be fed into the neural network using a back propagation algorithm the scripts will be recognized. There are other methods in neural network like feed forward, support vector machines but back propagation is well suited. This will be the Multi-Lingual Script Recognition System.

Objectives

To design a software that can read script many of Indian languages.

Methodology

The script recognition system is composed of five phases.

1. Digitization
2. Segmentation
3. Normalization
4. Feature extraction
5. Script recognition

Materials used for multi lingual script recognizer are:

1. Scanned multi-lingual image
2. MATLAB software
3. Neural network tool

Details of work carried out: First a multi-lingual document is converted into a digitized image and later it is normalized. Since the line by line identification of the script is needed its further line segmented. After line segmentation each line is again normalized. After this the important features of each line are extracted and then it is fed into the neural network for the recognition.

Typical final output of the system is as follows:

THE FIRST LINE OF THE SCRIPT IS HINDI

THE SECOND LINE OF THE SCRIPT IS ENGLISH

THE THIRD LINE OF THE SCRIPT IS KANNADA

Conclusion

Using the concepts of image processing and MAT LAB it's possible to design a system which could identify the different scripts used in a document which contain different scripts.

When a multi-lingual script document is to be processed, the respective language OCR's are to be used. But this MLSR allows to feed the multi-lingual documents and helps in identifying the different scripts.

Scope of future work

- This project is dealt with noise free images. Future system can be enhanced for noisy images.
- In this project it deals with only three languages. Future enhancement can be made for other scripts also.
- This project uses monochromatic bit map images. Colored images can also be used for recognition.