A Hybrid Decision Tree and Error Correcting Technique for Printed Tamil Character Recognition

M.Ramanan, A.Ramanan and E.Y.A. Charles

Introduction

- Scanned images are needed to be in ediatable format in order to perform the following:
  Searching, Copying, Editing, Grammar checking, Spell checking, Formatting, etc.
- Why Tamil OCR?
  To preserve or reproduce ancient or decrepit Tamil books.

Challenges to Tamil OCR

1. Printed Tamil documents poses challenges owing to:
   One line may have different font styles, presence of pictures, multi columns, etc.
2. Existing Tamil OCRs show moderate recognition rate.
   Not evaluated on a common or standard datasets due to the absence of a standard dataset.

Objectives

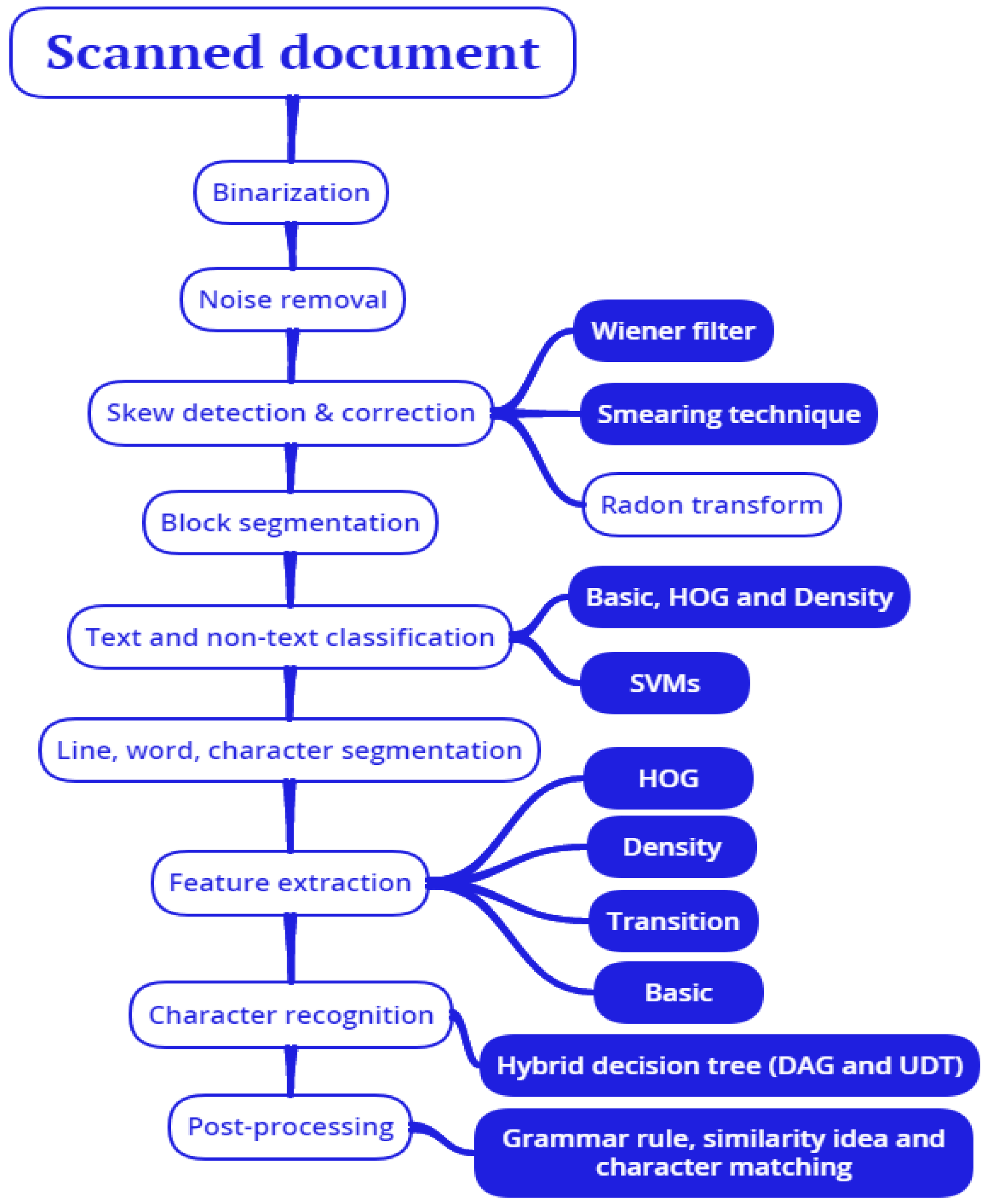
- To improve the recognition rate for printed Tamil text.
- To provide a standard and challenging dataset to the research community consisting scanned printed Tamil text.

Contributions

- Created a dataset for the research community working on printed Tamil text.
- Proposed a skew detection and textual classification using Wiener filter, smearing and Radon transform methods.
- Proposed a hybrid decision tree using SVMs for character recognition of Tamil text.
- Proposed a post-processing error correction technique.

Methodology

The system consists of pre-processing, feature extraction, character recognition and post-processing.

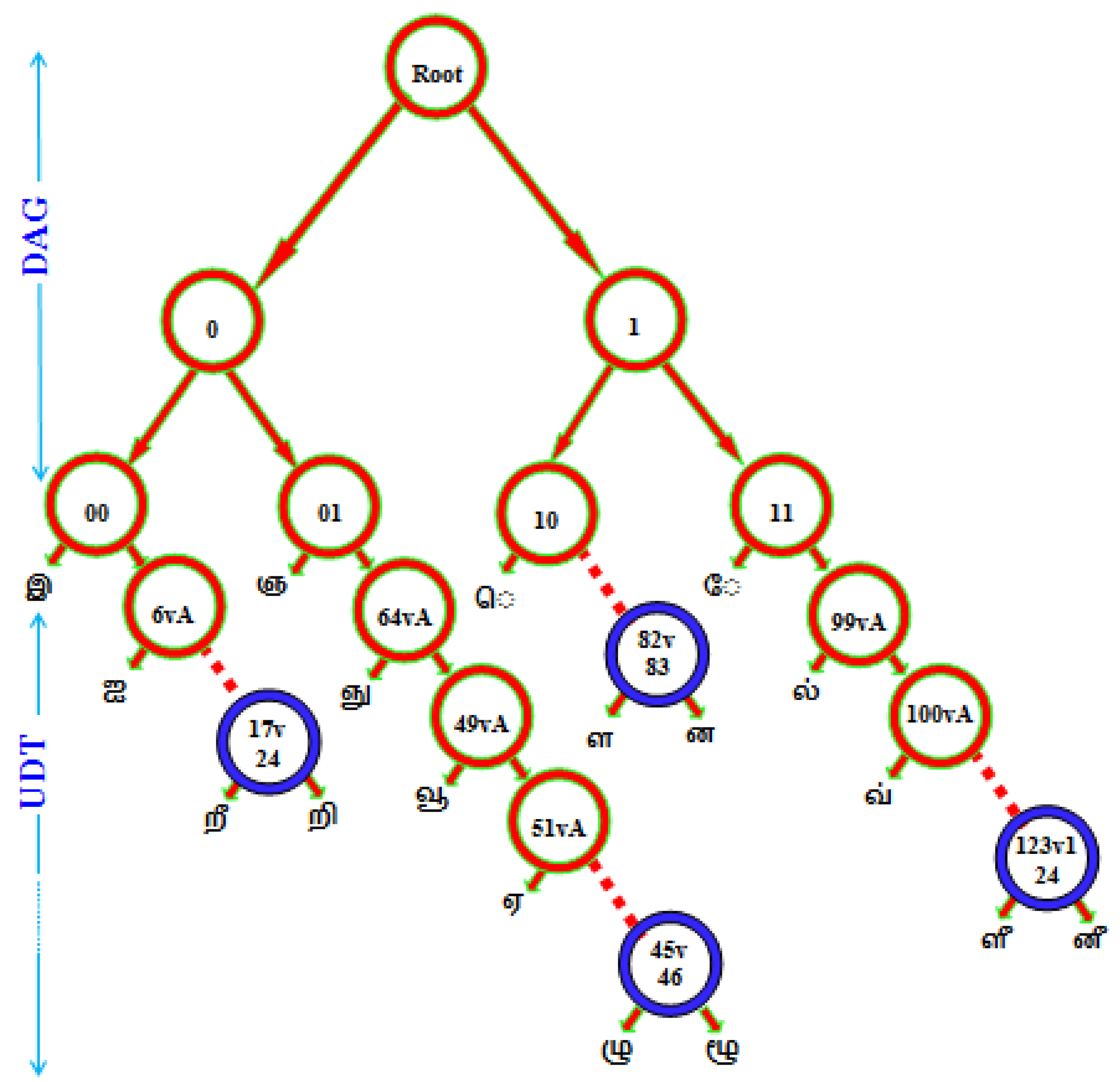


Preprocessing

1. Binarisation:
   Gray scale image ⇒ binary image.
2. Noise removal:
   Applying median filter to the binarised image.
3. Skew correction and detection:
   Using Wiener filter, smearing and Radon transform technique.
4. Block segmentation:
   Using run length smearing algorithm (RLSA) and connected component analysis.
5. Text or non text classification:
   Feature vectors (Basic, HOG and density) are analysed using SVMs to classify text or non-text block.
6. Character segmentation: Using projection technique and connected component analysis

Character recognition

Features used in this work: Basic (6 dimension), density (89 dimension), HOG (1154 dimension) and transition features (372 dimension).
The extracted feature vectors are analysed using the novel hybrid decision tree of DAG [2] and UDT-SVMs [1].



Hybrid decision tree - DAG

- Characters were grouped by applying K-means algorithm to find the root node of the hybrid decision tree.
- The same technique was applied to the next level of the hybrid decision tree.

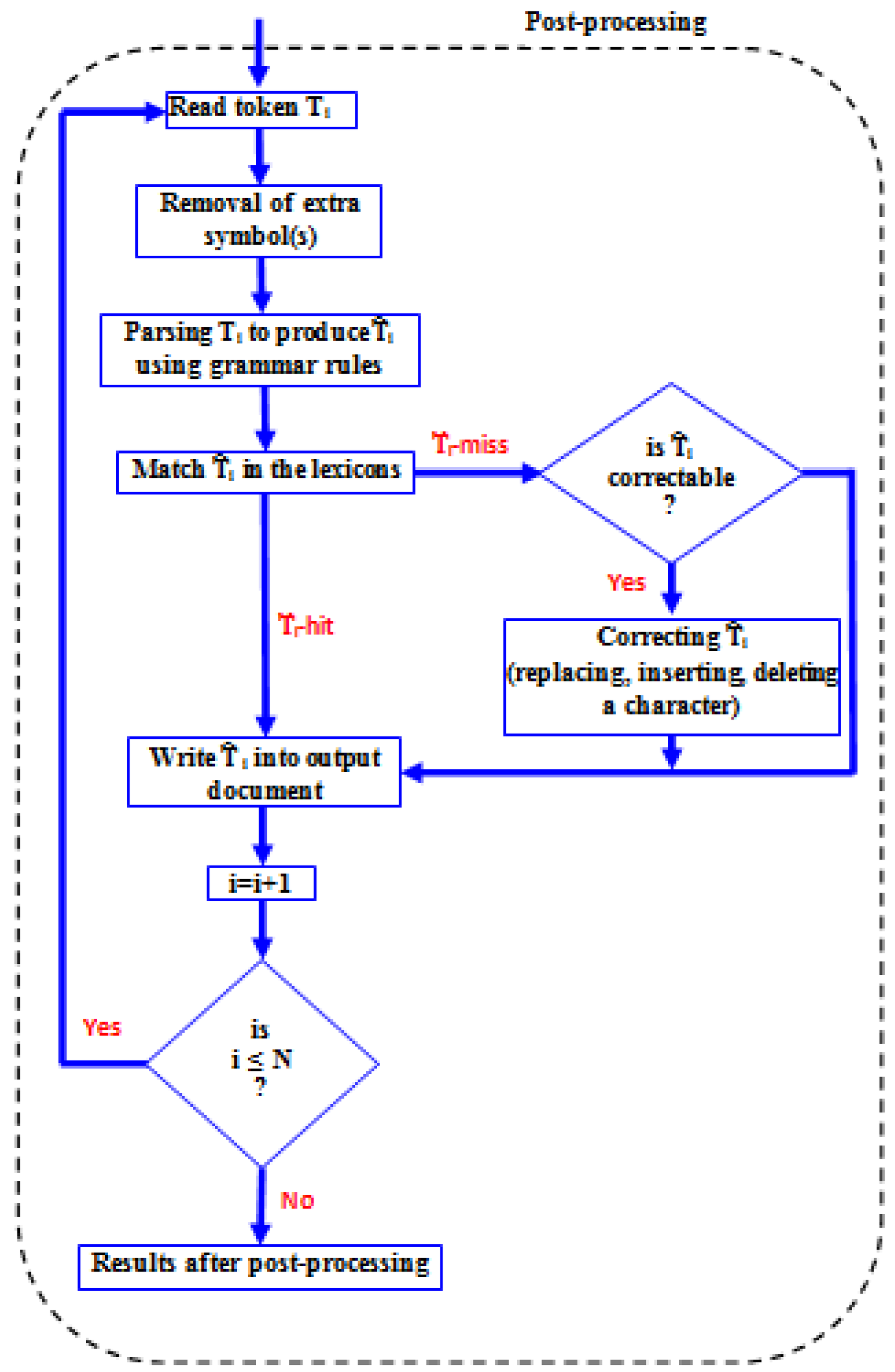
Code	Characters
0	அ ஆ இ ஒ ஓ ஐ த் ந் ழ் ற் ஞ் டீ ஞீ தீ நீ ழீ றீ ரீ டி ஞி தி நி ழி றி ரி ஞு ஙு சு ணு ணூ தூ நூ லு லூ ளு ளூ றூ னு னூ ஞூ மு மூ ரு ரூ ழு ழூ பூ யூ வூ கு ஏ த ந ழ ற டு து நு பு யு வு று ஙு ஞு ர
1	ஈ ெ ா க ச ப ம சு உ ஊ எ ை ங ய ல வ ள ன ண ட கூ டீ ே ஃ க் ங் ச் ட் ண் ப் ம் ய் ல் வ் ள் ன் கி ஙி சி ணி பி மி யி லி வி னி ளி கீ ஙீ சீ ணீ பீ மீ யீ லீ வீ னீ ளீ
00	அ ஆ இ ஒ ஓ ஐ த் ந் ழ் ற் ஞ் டீ ஞீ தீ நீ ழீ றீ ரீ டி ஞி தி நி ழி றி ரி
01	ஞு ஙு சு ணு ணூ தூ நூ லு லூ ளு ளூ றூ னு னூ ஞூ மு மூ ரு ரூ ழு ழூ பூ யூ வூ கு ஏ த ந ழ ற டு து நு பு யு வு று ஙு ஞு ர
10	ஈ ெ ா க ச ப ம சு உ ஊ எ ை ங ய ல வ ள ன ண ட கூ டீ
11	ே ஃ ர் க் ங் ச் ட் ண் ப் ம் ய் ல் வ் ள் ன் கி ஙி சி ணி பி மி யி லி வி னி ளி கீ ஙீ சீ ணீ பீ மீ யீ லீ வீ னீ ளீ

Hybrid decision tree - UDT

- Each decision node of the UDT is found based on the features that best separates the characters in recognition and the order of nodes is fixed in the decision tree.
- This process is followed for every decision node of UDT.

Error correction technique

The flowchart of the proposed post-processing technique consists of various steps.



Dataset

Four diverse types of printed Tamil documents: books, magazines, newspapers and pamphlets. Five different examples/document; ⇒ four pages / examples ⇒ Total 80 examples [3].

Results

- The recognition rate for skew correction is 95.83%
- The character recognition Proposed for hybrid decision tree is 98.80%.
- The error correction technique reduces the overall average error rate by nearly 5% in the output produced by the Tamil OCR.

References

1. Ramanan *et al.*, "Unbalanced Decision Trees for Multi-class Classification", 2007.
2. Platt *et al.*, "Large Margin DAGs for Multiclass Classification", 2000.
3. Tamil Digitising Project, Department of Computer Science, University of Jaffna, Sri Lanka, 2014. http://www.csc.jfn.ac.lk/tdp