TAMIL FONT TYPE IDENTIFICATION FROM TEXT IMAGES

P.Janani*[1], E.Y.A.Charles*[2]

*Department of Computer Science, Faculty of Science, University of Jaffna

[1]jenipathmanathan3@gmail.com, [2]charles.ey@univ.jfn.ac.lk

Introduction

- Font type identification is a process of finding the font style of texts in images.
- Font type can be manually identified with experience to some extent.
- It is difficult to differentiate a font type from another due to the vast number of available fonts.
- This work proposes a method to automate the font type identification using machine learning.

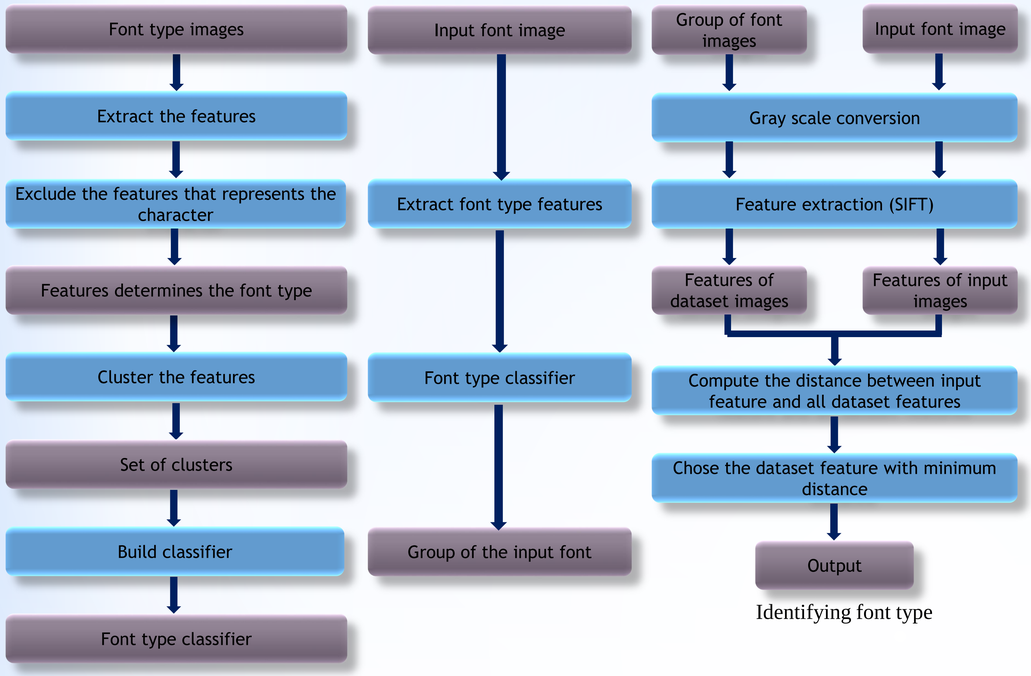
Motivation

- Graphic designers need to identify font types in images they encounter in their day to day life for later use.
- During the designing process of printed material, there is a need to identify similar fonts for a selected font.
- During OCR, the font style information are lost since the characters are identified using selected features. Font type information is needed to recreate a document from an image.

Challenges & Solution

- Font type identification can be done by finding the closest match of features of a given font among the available fonts.
- When the number of fonts increases, comparing a font with other fonts will be a time consuming and computationally intensive task.
- A better approach is to cluster fonts based on their style features and matching a given font image to a potential cluster.
- To cluster font types, style features and character type features should be distinguished from the available features.

Methodology



Identifying font type

Identifying group of font type

Dataset

To assist this research work, a Tamil font type dataset was created. This contains 1540 images covering 10 font styles.

Font Types	Number of Images
Anusha	148
Kamaas	155
Modern Tamil	156
Pravi	149
Rasihapriya	156
Sahaanaa	158
Siva-0002	152
Sivagami	152
Trinco Normal	157
Vairamani	157
Total	**1540**

Count of dataset images for each font styles

Examples of two character images
i)Kamaas
ii)Modern Tamil
iii)Rasihapriya
iv)TrincoNormal

Font Types	Sample Images
Anusha	
Kamaas	
Modern Tamil	
Pravi	
Rasihapriya	
Sahaanaa	
Siva-0002	
Sivagami	
Trinco Normal	
Vairamani	

Sample testing images

Experiments

- Experiments carried out on Tamil font dataset with MATLAB.
- Dataset contains 1540 images and 400 testing images covering 10 font styles.
- Each test images was compared to the dataset images and their K-Nearest neighbors were calculated.
- This test was conducted using various K values, K=1, K=5, K=10, K=15 K=20 and K=25.
- Among the testing images curved texts, rotated texts, texts with different font colors and background colors, texts with different font size and bold italic texts are included.

Confusion Matrix

Font Styles	Anusha	Kamaas	ModernTamil	Pravi	Rasihapriya	Sahaanaa	Siva-0002	Sivagami	TrincoNormal	Vairamani
Anusha	36	0	0	0	0	0	2	1	1	0
Kamaas	0	37	0	0	0	1	2	0	0	0
ModernTamil	0	3	31	1	0	1	1	3	0	0
Pravi	0	1	2	29	2	1	0	3	1	1
Rasihapriya	1	2	1	0	34	1	0	1	0	0
Sahaanaa	0	0	0	1	0	36	2	0	1	0
Siva-0002	0	1	1	1	0	2	33	1	1	0
Sivagami	1	0	2	1	0	1	2	31	2	0
TrincoNormal	0	1	3	0	0	0	1	0	34	1
Vairamani	0	0	2	0	0	0	0	2	0	36

Results

Results given in the following table shows the average recognition rate of each font face.

Fonts	K=1	K=5	K=10	K=15	K=20	K=25
Anusha	85.00%	77.50%	77.50%	77.50%	75.00%	75.00%
Kamaas	87.50%	87.50%	90.00%	90.00%	90.00%	90.00%
ModernTamil	75.00%	72.50%	77.50%	75.00%	77.50%	77.50%
Pravi	77.50%	82.50%	85.00%	85.00%	87.50%	87.50%
Rasihapriya	82.50%	82.50%	80.00%	82.50%	82.50%	87.50%
Sahaanaa	90.00%	87.50%	90.00%	87.50%	87.50%	87.50%
Siva-0002	87.50%	85.00%	87.50%	87.50%	87.50%	87.50%
Sivagami	80.00%	85.00%	87.50%	82.50%	82.50%	82.50%
TrincoNormal	87.50%	87.50%	87.50%	92.50%	92.50%	90.00%
Vairamani	82.50%	80.00%	80.00%	80.00%	80.00%	80.00%
Overall Accuracy	**83.50%**	**82.75%**	**84.25%**	**84.00%**	**84.25%**	**84.50%**

Related Works

- Deepfont : Identify Your Font from An Image
  - Authors : Zhangyang Wang, Zhangyang Wang, Jianchao Yang, Jonathan Brandt
  - Font identification from English Text images
  - Dataset : AdobeVFR (616 font styles with 4385 images)
  - Used CNN model
  - Accuracy : 80%
- Thai Font Type Recognition using SIFT
  - Authors : P. Jamjuntr and N. Dejdumrong
  - Font identification from Thai Document images
  - Dataset : 10 font styles and 10 text images in each font styles (100 images)
  - Features used: SIFT
  - Accuracy:97.37%

Conclusion & Future Work

- K-Nearest neighbor can identify the font styles of Tamil text images with an accuracy rate of 84.5%.
- This was done by comparing the input image features with features of each font type in the set of font types.
- Further research needed to be performed to distinguish the features that represent the style of a font type.
- Using the font style features we can recognize the font faces within a short period of time.