Digitization of Tamil Text

- Peratchi Hariharasudhan Kannan – U6366102
- Vinayak Ravi - U6561524

A Brief about Tamil script

- *Tamil script was said to have existed from 5th century BCE.*
- Tamil script consists of 31 letters in its independent form and an additional 216 combinant letters, summing to 247 combinations *("uyirmeyyeḻuttu -"soul-body-letters"*
- about 100 to 120 million people speak Tamil as of 2015 census of India.

https://en.wikipedia.org/wiki/Tamil_script

Tamil Script

ஃ		அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
		a	ā	i	ī	u	ū	e	ē	ai	o	ō	au
க்	k	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ங்	ṅ	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
ச்	c	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ்	ñ	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட்	ṭ	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண்	ṇ	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த்	t	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந்	n	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப்	p	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம்	m	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய்	y	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர்	r	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல்	l	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
வ்	v	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ்	ḻ	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள்	ḷ	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற்	ṟ	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	ṉ	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

Quick Fact:

"One of the languages which are spoken from time when humans were hunters".

https://www.nammakalvi.org/tamil.html

Literature Review for object recognition of Tamil Script

S. No & Paper Ref	Pre-Processing	Segmentation	Feature Extraction	Classification	Limitation
1 [1]	Morphological Operations were used for noise removal and morphological gradients to identify character boundaries	Histogram profile analysis on each component for line character segmentation	Using Scale Invariant Transform (SIFT) and K-means clustering to create code books for each characters	Bag-of- key points to count number of patches and coupling K-means and SIFT to transfer character image to set of local frame.	To recognize abnormal writing and similar shaped characters.
2 [2]	Thresholding method used for binarization, thinning algorithm and Hilditch algorithm was used for skeletonizing	Spatial space detection technique and histogram method used for converting image into glyph	The output from glyphs are chosen to be the features	Support Vector Machines (SVM) were used for classification, which creates a self-organizing map, to minimize error and also used neural classification algorithm and Radial Bias function (RBF) network.	Doesn't work well with increased sample set and for skew angles more than $\pm 15°$

Literature Review for object recognition of Tamil Script (contd.)

S. No & Paper Ref	Pre-Processing	Segmentation	Feature Extraction	Classification	Limitation
3 [3]	Binarization: high peak denoting the background, small peaks noting the foreground, & utilized median and Wiener filtering method used for noise removal.	Line segment, word and character segmentation	Time-domain features, which is normalizing x-y coordinates, normalizing first and second derivatives, curvature aspects and Discrete Cosine transform (DCT), acting as a sliding window.	Hidden Markov Model (HMM) for training and re-estimation using Baum-Welch algorithm.	It was considered using minimum hand-written scanned documents, and not continuous characters & sliding characters.
4 [4]	Converted input hand-written character into Unicode characters by segmentation.	Paragraph into lines and lines into words segmentation.	The segmented outputs are accounted to be features.	The extracted features are inputted into self-organizing maps and using Fuzzy neural networks into character segmentation.	Doesn't work well with increased sample set and for skew angles more than $\pm 15°$

Troubles faced in classification of Tamil Characters

- Curves in Tamil characters
- Complex letter structure
- Variation in writing styles (causing mis-interpretation when defining the strokes)
- Difficulties in viewing angles, definition of amount of skew in the characters

Strategy Proposed

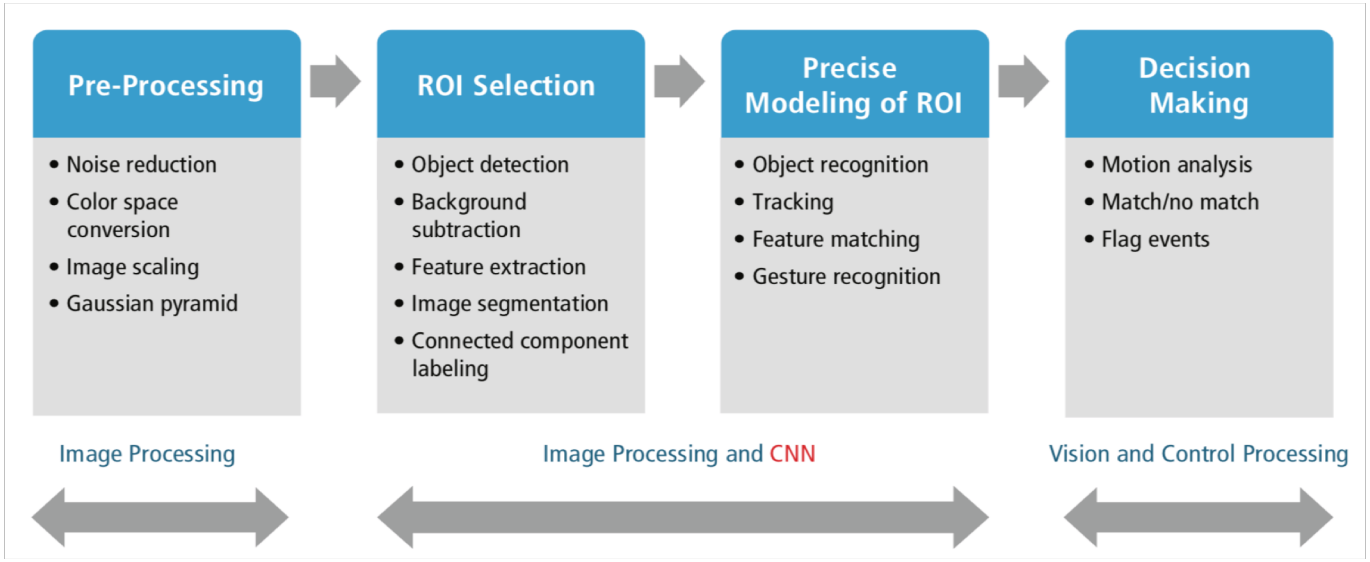
In the literature review [5], Tamil characters were classified using Convolutional Neural Networks from data obtained from HPL database. They utilized various pooling techniques (max, stochastic), probabilistic weighting, activation functions (tanh, ReLu) and classifiers (Softmax).

CNN has been around for 50 years, with certain advantages for identifying and classifying images (in our case Tamil characters), they are:

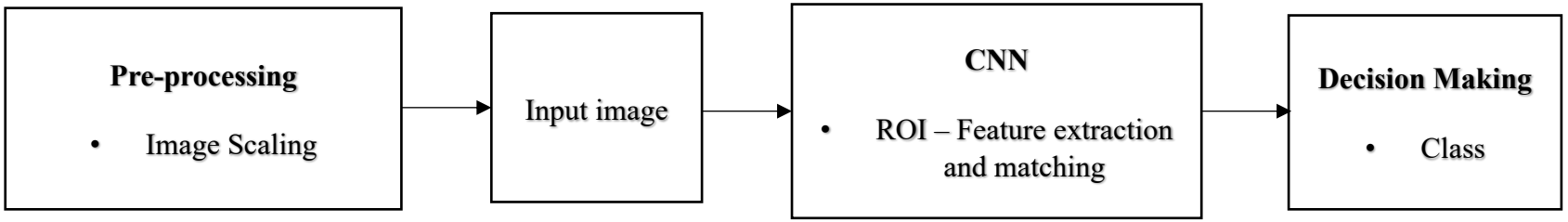
- Rugged to intrinsic distortions (changes caused due to skew in camera angle) and different lighting conditions, different poses and are shift invariant.
- Range of parameters choice is high
- Models can be generalized with ease (inducing Robustness).

https://ip.cadence.com/uploads/901/TIP_WP_cnn_FINAL-pdf

Pipeline for CNN in Image Processing



Pipeline Implemented:

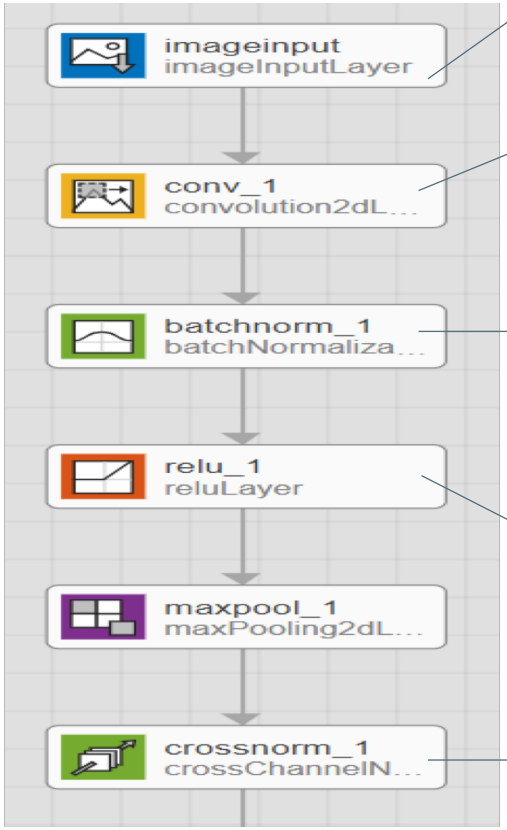


https://ip.cadence.com/applications/cnn

Dataset Pre-processing

- HPL Dataset consisting of 50683 with 155 classes [6].
- Resizing to size (32 x 32) images.
- Label definitions (no 1 to no 155)
- Next step, is to input resized images into network

Network architecture



(32 x 32 x 1) size

Feature map sizes (8, 16, 32, 64)
Filter size (5 x 5) . optimizers:
ADAM & SGD

$$output\ width = \frac{W - F_w + 2P}{S_w} + 1$$

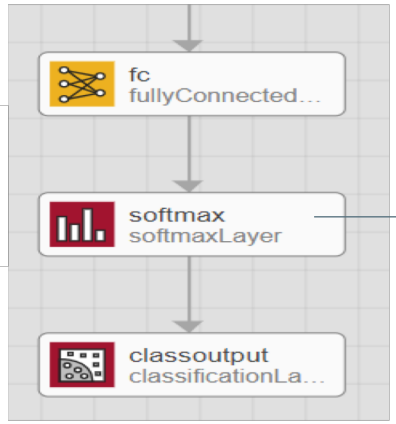
$$output\ height = \frac{W - F_h + 2P}{S_h} + 1$$

batch normalization layer would normalize the activation of a particular convolution layer with respect to mini batch mean and standard deviation.

helps in identifying the features extracted

$$f(x) = x^+ = \max(0, x)$$

cross channel normalization, after the convolution layers where it normalizes with respect to channel wise mean than the whole mini batch.



Used in identifying features w,r,t. classes.

Parameter Combinations

- Learning rate
  - 0.001, 0.01 (default), 0.1
- Minibatch size
  - 32, 64,128
- Optimizer
  - Adam/ Stochastic Gradient Descent (SGD)

Demo

Troubles faced during Learning

- Finding the right data (insufficient data/ Data augmentation)
- Finding the right activation function (tanh/ReLU)
- Parameters selection (Learning rate/ Optimizer)

Future Recommendations & conclusion

- Expansion of dataset
- Generalization of dataset, by preprocessing of image input (through thresholding, skeletonizing, segmentation)
- Local contrast normalization (Subtractive or Divisive, relates every value by standard deviations of its neighbors over space), which helps under non-uniform illumination or shading artifacts [5].

References

[1] A. Subashini, N.D. Kodikara, " A Novel SIFE-based Codebook Generation for Handwritten Tamil character Recognition" , *6th IEEE Int. Conf. on Industrial and Information Systems (ICIIS)*, Page(s): 261 – 264, 2011

[2] C. Suresh Kumar and T. Ravichandran, "Handwritten Tamil Character Recognition using RCS algorithms", Int. J. of Computer Applications, (0975 – 8887) Vol. 8– No.8, 2010

[3] R.J. Kumar, R. Prabhakar and R.M. Suresh, "Off-line Cursive Handwritten Tamil Characters Recognition", International Conference on Security Technology, pp. 159 – 164, 2008

[4] C. Sureshkumar and T. Ravichandran. "Handwritten Tamil character recognition and conversion using neural network". InIntJComputSciEng 2 (7 2010), pp. 2261–2267.

[5] P. Vijayaraghavan and M. Sra. "Handwritten Tamil Recognition using a Convolutional Neural Network".

[6] HP Labs India Pvt. Ltd, http://lipitk.sourceforge.net/datasets/tamilchardata.htm .

Github link:

https://github.com/harish-kp/Data-analytics