Engineering Science and Technology, an International Journal 22 (2019) 637–645

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Engineering Science and Technology,
an International Journal

journal homepage: www.elsevier.com/locate/jestch

Full Length Article

On developing handwritten character image database for Malayalam language script

K. Manjusha [a],*, M. Anand Kumar [b], K.P. Soman [a]

[a] *Center for Computational Engineering & Networking (CEN), Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India*
[b] *Department of Information Technology, NIT K - Surathkal, Mangalore 575025, India*

ARTICLE INFO

Article history:
Received 24 January 2018
Revised 14 October 2018
Accepted 29 October 2018
Available online 5 February 2019

Keywords:
Malayalam language
Handwritten character recognition
Handwritten character image database
Active contour minimization
Optical character recognition
Feature extraction

ABSTRACT

The objective of this paper is to build a handwritten character image database for Malayalam language script. Standard handwritten document image databases are an essential requirement for the development and objective evaluation of different handwritten text recognition systems for any language script. Considerable research efforts for handwritten Malayalam character recognition are present in literature. Still, no public domain handwritten image database is available for the Malayalam language. The present work focuses on building an open source handwritten character image database for Malayalam language script. The unique orthographic representation of the Malayalam characters forms the different character classes, and the current version of the database contains 85 character classes frequently used in writing Malayalam text. Handwritten data samples collected from 77 native Malayalam writers. For extracting the character images from the handwritten data sheets, active contour model-based image segmentation algorithm utilized. Recognition experiments conducted on the created character image database by employing different feature extraction techniques. Among the considered feature descriptors, scattering convolutional network-based feature descriptors attain the highest recognition accuracy of 91.05%.

1. Introduction

The capability of a machine to convert handwritten documents captured through imaging devices to its equivalent machine-readable and searchable document format often referred to as offline handwritten text recognition. Handwritten text recognition is one of the popular, and active research domain under pattern recognition. Offline handwritten document text recognition is an integral part of different real-world applications such as postal mail sorting, processing of bank cheques, reading aid for the blind and data entry applications [1]. Document text recognition eases automatic data entry applications by identifying the text contents in document images and thus greatly reduces the manual effort. The foremost step in implementing document recognizer for any particular language script is the selection of a suitable document image database. If such databases are not available, then the first task is creating a new document image database for that script. Benchmark databases are often necessary for implementing, evaluating and comparing different handwritten document recognition systems [2]. Several such benchmark databases for research and developments in handwritten document recognition are available for Latin [3,4], Arabic [5,6], Chinese [7–9], Japanese [9] and Korean [2] language scripts.

In the case of Indian language scripts, very limited number of benchmark document image databases are available. Indian statistical institute (ISI) Kolkata have already released offline handwritten character image databases for Bangla, Devanagari and Oriya scripts [10]. The ISI databases include isolated numeral, basic character, vowel modifier and compound character based datasets for Bangla script isolated numeral and basic character based dataset for Devanagari script and isolated numeral based dataset for Oriya script. Another major provider of character image database repository for Indian language script is Center for Microprocessor Application for Training Education and Research (CMATER), Jadavpur university, Kolkata [11]. CMATERdb includes unconstrained handwritten document images in page, line, word and character level for Bangla, Devanagari, Arabic and Telugu scripts. Indic handwriting datasets published by HP Labs India contains offline image versions of the online handwritten character data for Tamil, Telugu and Devanagari scripts [12].

Even though so many isolated research works have been carried out in Malayalam handwritten document recognition, the lack of

* Corresponding author.
E-mail addresses: k_manjusha@cb.amrita.edu (K. Manjusha), kp_soman@amrita.edu (K.P. Soman).

Peer review under responsibility of Karabuk University.

https://doi.org/10.1016/j.jestch.2018.10.011
2215-0986/© 2018 Karabuk University. Publishing services by Elsevier B.V.

standard open source Malayalam document image database is one of the problems that every researcher working in Malayalam handwriting recognition has to face. The prior Malayalam handwritten character recognition research works utilized their own created databases for experimenting and evaluating different character recognition algorithms. The non-availability of standard benchmark handwritten document image database for Malayalam language script makes the uniform comparative assessment of the research efforts reported in Malayalam character recognition, difficult. This paper is about the creation of a public domain character image database for Malayalam language script.

Section 2 describes the Malayalam language script and the research efforts reported in the literature of handwritten Malayalam character recognition. Section 3, 4 gives an overview of the document image collection process and about the statistics of the created Malayalam character image database. The performance analysis of different recognition systems on the created database is described in Section 5. Section 6 discusses the summary of the present work and the scope of future works.

2. Malayalam language script

Malayalam is one of the classical languages in India, and have more than 33 million native speakers according to the 2001 census of India. The Malayalam language is most widely used in Kerala, one of the south-western states of India and in Lakshadweep, one of the union territories in India. The Malayalam language script is one of the official Indian scripts and it belongs to the Dravidian family of languages. The Malayalam language has a close connection with Tamil and Sanskrit languages and these languages have greatly influenced the formation of Malayalam grammar and vocabulary [13]. Because of this, Malayalam script makes use of symbols from *Vatteluttu* and *Grantha* script in writing syllables. Like most of the Indian language scripts, the Malayalam script has alpha-syllabary nature. The script is written from left to right non-cursively and has no concept of lower and upper case characters. Malayalam graphemes are generally more rounded in shape compared to other Indian language scripts. The presence of straight lines are found only in 27% of cases of basic character set and the direction of concavities in most of the Malayalam characters are downwards [14].

Words in the Malayalam language are usually written as a sequence of syllables and the unique orthographic unit in the script is usually termed as *akshara*. The basic *akshara set* (*aksharamala*) of the Malayalam script contains 15 vowels (*swarangal*) and 36 consonants (*vyanganagal*) symbols. Vowels in the Malayalam script are represented in independent and dependent forms. Independent vowel forms are used when the Malayalam word starts with a vowel. Dependent vowel forms (vowel modifiers) are represented as diacritic and are attached with consonants to indicate that the consonant is followed by a vowel. For all vowels other than 'a', there is diacritic symbols. Vowel 'a' follows basic consonants by default and is termed as inherent vowel. Special diacritic 'virama' symbol is used to represent pure consonant (without the inherent vowel). The Malayalam script also contains symbols for half-consonants (*chillu*), conjunct characters (combination of two or more consonants), consonant modifiers, numerals and some special characters. The traditional Malayalam script (old lipi) had huge number of character glyphs due to the presence of large number of conjunct characters. Conjunct characters have complex orthographic structures and they produces large number of similarly shaped characters with small change in their orthographic structure. In 1971, reformed version of Malayalam script (new lipi) was introduced by Government of Kerala, in which complex conjunct characters are represented through sequence of consonant characters and diacritics. But the reformed version of the script only partially changed the traditional old script, and with the arrival of modern word processors, most of the old or traditional lipi characters re-emerged in print. Due to this, the present Malayalam script is often a mixture of old and new lipi characters. The total number of unique character glyphs in Malayalam language script is more than 250 because of the presence of both old and new lipi characters in documents. Among these, so many conjunct character glyphs are nowadays used only in printed documents which make use of some specific font styles. When comes to writing, nowadays the new lipi characters along with some selected conjunct characters are only commonly used.

The research efforts on Malayalam handwritten character recognition started a little late compared to Bangla and Devanagari scripts, the two most popular languages in India [15]. In character recognition scenario, the large number of character classes and the structural resemblance among character shapes make Malayalam character recognition, a challenging large multi-class classification problem. Most of the works reported in literature of Malayalam handwritten character recognition is based on structural [16,17], statistical [18–20], gradient [21,22] and transform domain [23–28] features. Support vector machines (SVM) and neural network (NN) based classifiers are found effective in handling Malayalam character classification problems [29]. In this paper, different feature extraction techniques are evaluated on the created Malayalam handwritten character image database.

3. Malayalam handwritten document image data collection

This section describes about the Malayalam handwritten image data collection process conducted for creating character level image database for Malayalam language script.

3.1. Malayalam character classes

Before starting handwritten data collection, the Malayalam character classes are decided based on the unique orthographic structures in Malayalam language script. 85 Malayalam character classes representing vowels, consonants, half-consonants, vowel modifiers, consonant modifiers and conjunct characters that are frequently used while writing are considered for database creation. Table 1 lists the orthographic symbols representing the considered 85 Malayalam character classes along with their unicode representations. The conjunct characters are combination of consonants and are represented with combination of unicode representation of constituting consonants and diacrictics.

3.2. Document image data collection process

For pattern recognition related applications, data patterns are one of the most necessary requirements. If the data patterns for the particular recognition application is not available, then the first and foremost task in implementing the recognition system is to collect the data patterns. Data collection is one of the tedious task in most of the pattern recognition applications. The handwritten documents are collected from different native Malayalam writers under supervision. For collecting character images, the writers are instructed to write the considered Malayalam character classes on pages five times using ball point pens by keeping attention on space between each written characters. No restriction is kept on the type or quality of the paper and the ball point pen used for writing. Fig. 1 shows the sample handwritten data sheets collected as part of data collection process.

The handwritten data collected from 77 (60 Female and 17 Male) native Malayalam writers between 20 to 60 age groups

Table 1
85 Malayalam character classes considered for the current database creation.

Independent Vowels									
അ	ആ	ഇ	ഉ	ഋ	എ	ഏ	ഒ		
0D05	0D06	0D07	0D09	0D0B	0D0E	0D0F	0D12		
Consonants									
ക	ഖ	ഗ	ഘ	ങ	ച	ഛ	ജ	ഝ	ഞ
0D15	0D16	0D17	0D18	0D19	0D1A	0D1B	0D1C	0D1D	0D1E
ട	ഠ	ഡ	ഢ	ണ	ത	ഥ	ദ	ധ	ന
0D1F	0D20	0D21	0D22	0D23	0D24	0D25	0D26	0D27	0D28
പ	ഫ	ബ	ഭ	മ	യ	ര	റ	ല	ള
0D2A	0D2B	0D2C	0D2D	0D2E	0D2F	0D30	0D31	0D32	0D33
ഴ	വ	ശ	ഷ	സ	ഹ				
0D34	0D35	0D36	0D37	0D38	0D39				
Half Consonants									
ൺ	ൻ	ർ	ൽ	ൾ					
0D7A	0D7B	0D7C	0D7D	0D7E					
Vowel and Consonant Modifiers									
◌ാ	◌ി	◌ീ	◌ു	◌ൂ	◌ൄ	െ◌	േ◌	◌ൌ	◌ം
0D3E	0D3F	0D40	0D41	0D42	0D44	0D46	0D47	0D4C	0D02
◌്	്യ	്ര	്വ						
0D4D	0D4D; 0D2F	0D4D; 0D30	0D4D; 0D35						
Conjunct Characters									
ക്ക	ക്ഷ	ങ്ക	ങ്ങ	ച്ച	ഞ്ച	ഞ്ഞ	ട്ട	ണ്ട	ണ്ണ

0D15;
0D4D;
0D15
	0D15;
	0D4D;
	0D37
		0D19;
		0D4D;
		0D15
			0D19;
			0D4D;
			0D19
				0D1A;
				0D4D;
				0D1A
					0D1E;
					0D4D;
					0D1A
						0D1E;
						0D4D;
						0D1E
							0D1F;
							0D4D;
							0D1F
								0D23;
								0D4D;
								0D1F
									0D23;
									0D4D;
									0D23

ത്ത	ദ്ധ	ന്ത	ന്ദ	ന്ന	പ്പ	മ്പ	മ്മ	യ്യ	ല്ല

0D24;
0D4D;
0D24
	0D26;
	0D4D;
	0D27
		0D28;
		0D4D;
		0D24
			0D28;
			0D4D;
			0D26
				0D28;
				0D4D;
				0D28
					0D2A;
					0D4D;
					0D2A
						0D2E;
						0D4D;
						0D2A
							0D2E;
							0D4D;
							0D2E
								0D2F;
								0D4D;
								0D2F
									0D32;
									0D4D;
									0D32

ള്ള	വ്വ								

0D33;
0D4D;
0D33
	0D35;
	0D4D;
	0D35								

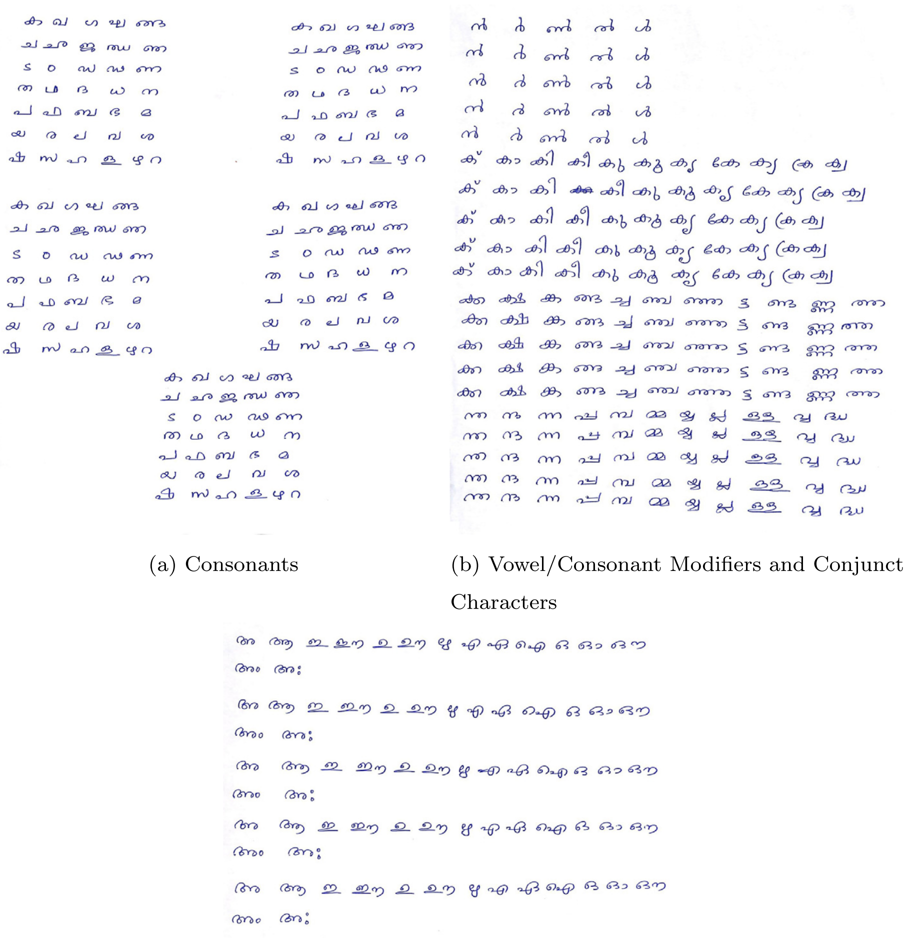
and all the writers have minimum graduation as the education qualification. The learning and testing datasets are divided based on the writers rather than collected images. Among 77 writers, the handwritten data collected from 59 persons considered for creating learning dataset while handwritten data from remaining 18 persons considered for creating testing dataset. Table 2 shows the gender and age information of writers whose handwritten data are included in learning and testing dataset. The database has three times more female writers than male writers. The age of writers has an unbalanced distribution among the divided age groups in learning and testing dataset. The testing dataset can be more challenging as the writers in 45–55 age group is not included in learning dataset but are present in the tesing dataset. The collected handwritten data sheets are optically scanned using FUJITSU Image Scanner *ScanSnap SV600* with automatic detection scanning mode.

4. Creation of Malayalam handwritten character image database

From the scanned handwritten data sheets, isolated handwritten characters are extracted by applying character segmentation algorithm. The segmented character images are manually tagged and grouped to classes for creating character image database. This section describes the character image extraction process from collected handwritten data sheets and the structure of the character image database created from those segmented images.

4.1. Character segmentation

For extracting isolated handwritten characters, active contour model based image segmentation algorithm is utilized. Geometric active contour models are variational and partial differential



(a) Consonants

(b) Vowel/Consonant Modifiers and Conjunct Characters

(c) Vowels

Fig. 1. Samples of Malayalam Handwritten data sheets collected from a single writer.

Table 2
Statistics of writers in datasets.

Dataset	Number of Writers	Gender		Age (Yrs)				Education Level
		Female	Male	20–29	30–39	40–49	50–59	
Learning	59	45	14	44	13	02	00	⩾Graduation
Testing	18	15	03	11	01	04	02	⩾Graduation

equation based methods that can extract objects of interest from input images. These models utilizes the idea of level set theory and evolve contour in images so as to fit on the object boundaries.

For representing the curve C, the level set function u can be represented as shown in Eq. (1) where the curve C is at u = 0 and $d((x,y),C)$ is the closest distance from point (x,y) to curve C.

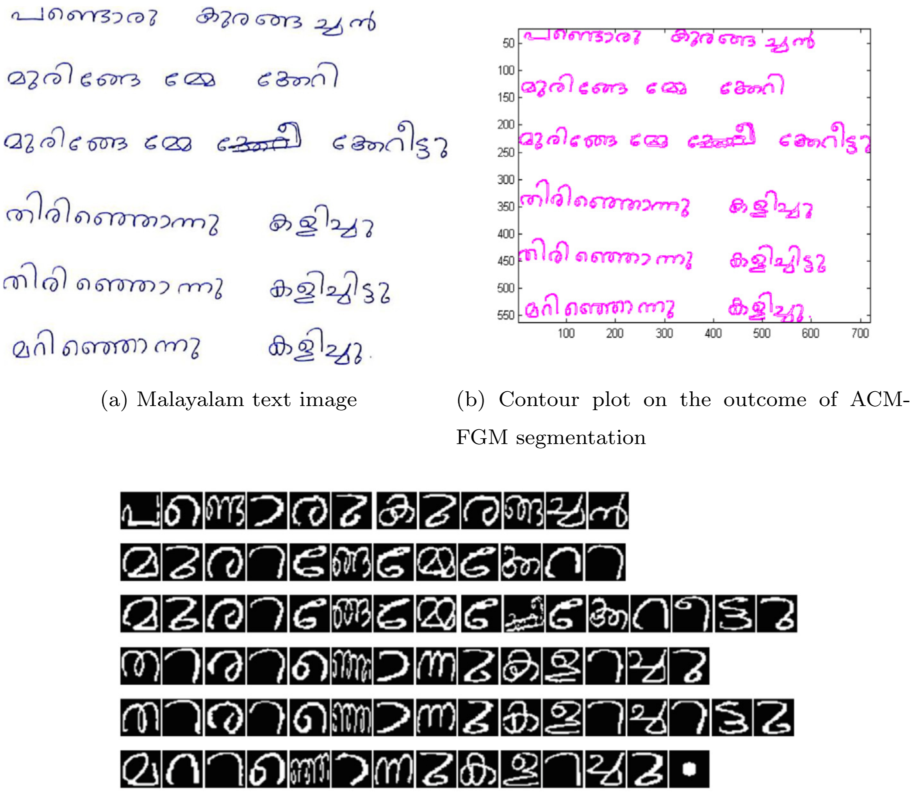
$$u(x,y) = \begin{cases} d((x,y),C), \text{if} \quad (\text{x},\text{y}) \quad \text{is} \quad \text{inside} \quad C \\ -d((x,y),C), \text{if} \quad (\text{x},\text{y}) \quad \text{is} \quad \text{outside} \quad C \end{cases} \tag{1}$$

For evolving the curve over time in order to fit on the object boundaries inside the image, partial differential equation is applied on the level set function u through mean curvature motion [30]. In order to stop the curve evolution when the curve reaches the object boundary, edge and region based information calculated from image are utilized. The mean intensity inside the curve and outside the curve are utilized in region based active contour models for fitting the evolving curve on the object boundaries inside the image. Minimization of the energy functional based on region based active contour models are utilized for effective segmentation of the images [31].

The final contour which fits on the character boundaries are processed to extract the isolated character images [30,32]. Fig. 2 shows the outcome of contour evolution and the character segmentation based on those contours fitted on the character boundaries of the input image. Fig. 2(a) shows the sample image containing Malayalam text and Fig. 2(b) shows the contour plot on the image. For converting the resultant image to binary representation, Otsu's global image threshold algorithm [33] is used and Fig. 2(c) shows the binary representation of the segmented characters that are resized to 32 × 32 dimension.

4.2. Structure of the character image database, Amrita_MalCharDb

The present version of the Malayalam character image database consists of binary isolated character images extracted from the



(a) Malayalam text image

(b) Contour plot on the outcome of ACM-FGM segmentation

(c) Character Segmentation

Fig. 2. Character segmentation based on ACM-FGM algorithm.

collected handwritten data sheets using character segmentation algorithm. As there was no constraints forced on writers in writing the isolated Malayalam characters, text portions inside the scanned images of handwritten data sheets are cropped in order to make the segmentation easier. The segmented character images are manually grouped into the considered 85 Malayalam character classes. Broken or heavily distorted character images are excluded while grouping and the segmented character images are resized to 32×32 dimension. The created character image database is available on request[1], and is released in the form of comma separated values (CSV) files. Three CSV files representing training, validation and testing images are available. Each row in the CSV files, represents a character image. The first column represents the class label and rest of the columns represent the pixel intensity values for the 32 × 32 image, vectorized in column wise. The isolated character images from 59 persons forms the learning dataset. 75% of character images in each character class from the learning dataset taken for creating training dataset while remaining 25% considered for validation. The images from rest of the 18 writers forms the test dataset. The total number of character images in the created database is 29,302. Number of character images per class is not same, as some images are discarded due to distortion or breakage after character segmentation. The average number of character images per class in the entire database is 344 and the standard deviation for the count distribution among classes is 37.96. Fig. 3 shows random samples taken from 6 different Malayalam character classes.

The total number of images in training dataset is 17,236. For validation dataset the total images are 5706 and for testing dataset it is 6360. The class-wise count of character images in training, validation and testing dataset are calculated. For each character class in training dataset, the mean of the area covered by character object in the images calculated. Table 3 shows class count information and the mean, standard deviation of the area of character object for the randomly selected 20 Malayalam character classes. The number of ON pixels in the character image depends on the writing style of the writer and the resources utilized for writing. The minimum and maximum number of ON pixels in the training dataset for the selected classes are listed in last two columns of Table 3. The statistics information for all the 85 character classes in the current version of the database is available along with the database.

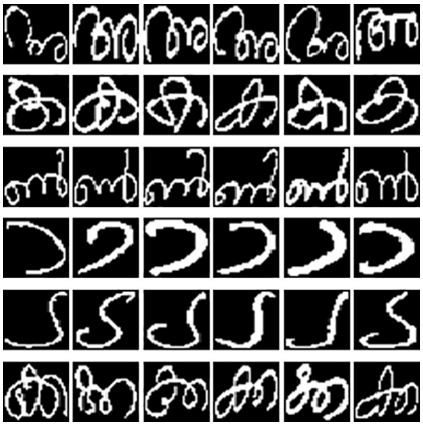


Fig. 3. Randomly chosen character images in six different Malayalam character classes.

The present database is named as Amrita_MalCharDb. The database is created as part of the research work conducted at Center for Computational Engineering and Networking in Amrita Vishwa Vidyapeetham university, and is indicated through 'Amrita' tag in

[1] Send an email to manjushagecpkd@gmail.com with subject "Amrita_MalCharDb".

Table 3
Class wise statistics of random selected 20 Malayalam character classes.

Malayalam Character	Number of Character Images			Statistics of Training dataset			
	Training	Validation	Testing	Mean	Standard Deviation	Minimum Pixels	Maximum Pixels
ഇ	225	75	109	249.78	50.33	125	391
ഉ	234	77	102	231.72	50.37	107	396
എ	220	73	101	202.06	38.64	123	303
ഏ	198	65	87	204.35	42.27	118	348
ഞ	200	66	69	262.45	48.18	159	371
ഠ	190	63	44	269.78	45.16	154	384
ന	198	66	73	226.97	42.53	125	354
ള	204	67	79	221.22	43.18	108	339
ൻ	219	73	64	175.16	36.83	89	294
ർ	217	72	65	189.32	36.12	100	279
ൽ	221	73	63	184.91	37.84	90	293
ൾ	223	74	65	177.89	36.34	92	269
◌ം	171	57	67	338.32	69.10	194	512
◌ി	192	63	58	140.94	33.84	70	270
◌ീ	190	63	69	171.88	37.59	85	276
◌ൗ	230	76	105	214.91	44.30	99	394
ണ	207	68	88	244.40	44.78	148	393
ത്ത	210	70	82	241.98	47.06	125	400
ന്ന	202	67	90	212.95	37.54	134	326
ഞ്ച	216	72	86	231.11	41.74	131	425

database name. ‘MalCharDb’ stands for Malayalam character database. For the present work, only 85 frequently used Malayalam characters are included. The present database can be extended by collecting handwritten character samples for all the valid Malayalam character glyphs in use.

5. Performance analysis on Amrita_MalCharDb

The character based recognition performance is evaluated for Amrita_MalCharDb, using different feature extraction and classification algorithms. The feature extraction process extracts informative descriptors from the images and helps the classifiers in estimating decision boundaries among participating classes. Usually the feature descriptors utilized for classification greatly affects the recognition performance of the underlying system [34]. In this paper, image pixel (IMG), histogram of oriented gradients (HOG), singular value decomposition (SVD), curvelet tranform (CT), gabor filters (GF), run length count (RLC), scattering convolutional network (ScatCN) and convolutional neural network (CNN) based feature descriptors are experimented on the created Malayalam character image database, Amrita_MalCharDb. IMG features are the vectorized representation of pixel intensity values contained in the character images. HOG descriptors are localized feature descriptors obtained from normalized histograms of image gradient directions [35]. SVD features utilized in this paper uses the dimension reduction technique through the matrix factorization technique in [20]. The feature representation based on curvelet transform (CT), is obtained by computing vertical and horizontal projection profile of the coarse curve co-efficient calculated through curvelet transform on the character images [27]. For GF features, 2D gabor filter bank are applied on input images and the mean and standard deviation of resultant images are computed [36]. RLC features are based on the successive runs of white pixels on the localized blocks of the character image [22]. Scattering representations are based on scattering transform which generates invariant feature descriptors using wavelet decomposition, modulus and averaging function [37]. Scattering convolutional network generates scatting co-efficient over each node in the network with the support of cascaded wavelet decomposition and are used as features ScatCN and ReducedScatCN [28]. CNN are the state of art technique in most of the image related recognition applications [38]. Lenet-5 [38], one of the popular CNN architecture for

Table 4
Recognition on Amrita_MalCharDb.

Recognizer	Recognition Accuracy (%)		
	Training	Validation	Testing
IMG-LinearSVM	99.80	88.63	74.61
IMG-RBFSVM	100.00	90.52	77.22
HOG-LinearSVM	83.64	82.68	73.38
HOG-RBFVM	99.99	95.32	86.71
SVD-LinearSVM	97.38	89.20	75.30
SVD-RBFSVM	99.98	93.39	80.09
GF-LinearSVM	89.82	77.66	61.09
GF-RBFSVM	97.65	81.51	62.69
CT-LinearSVM	82.10	75.82	61.45
CT-RBFSVM	100.00	83.70	65.76
RLC-LinearSVM	98.94	90.89	79.42
RLC-RBFSVM	99.94	93.76	82.80
ScatCN(Layer0,1)-LinearSVM	99.83	97.69	90.17
ScatCN(Layer0,1)-RBFSVM	99.93	98.00	90.52
ReducedScatCN-LinearSVM	99.88	97.93	91.05
ReducedScatCN-RBFSVM	99.68	98.23	90.96
Lenet-5	99.99	95.97	87.30

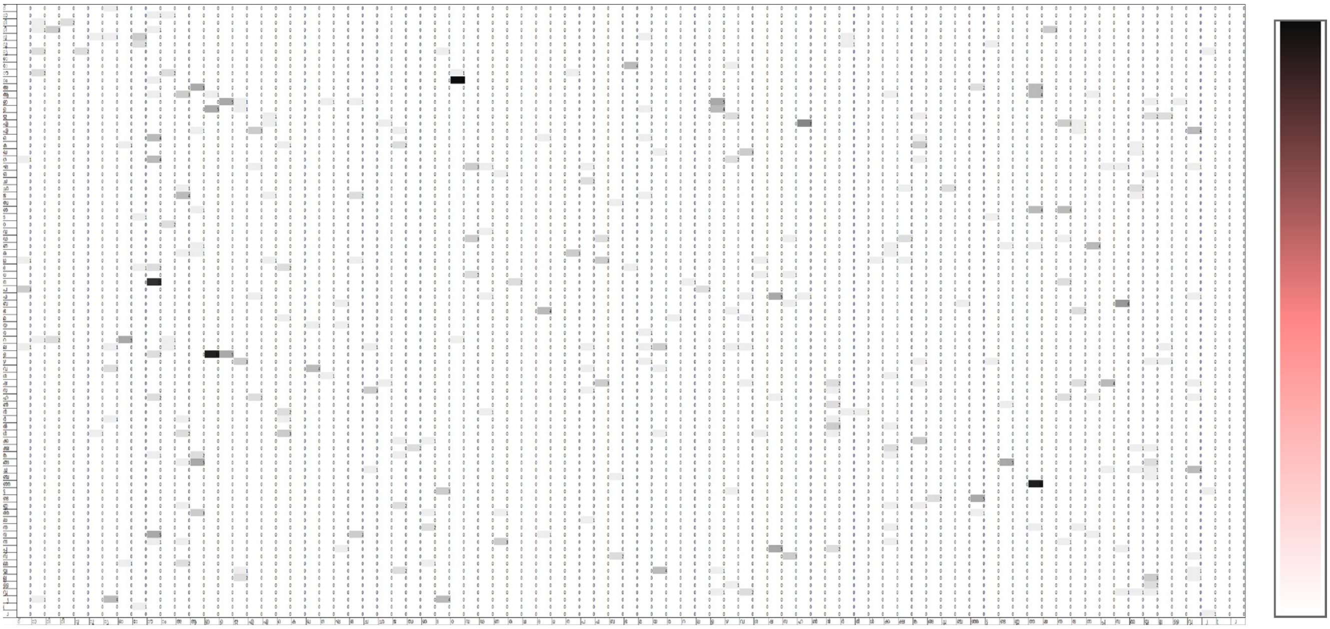


Fig. 4. Misclassifications in ReducedScatCN visualized using confusion matrix.[2].

document recognition is employed for Malayalam character recognition. Support Vector Machine (SVM) classifiers have the ability to obtain great generalization performance in large scale multi-class classification problems [29]. For classifying IMG, HOG, SVD, CT, GF, RLC and ScatCN feature descriptors, SVM classifier is employed. For implementing SVM classifier, LibSVM toolkit [39] is employed. For Lenet-5 CNN architecture, Keras toolkit [40] on the top of Tensorflow [41] software is utilized. Apart from CNN features, all other features are classified with linear and non-linear SVM classifier. For building non-linear SVM classifier model, radial basis function (RBF) is utilized as the underlying kernel function. The classifier models are built based on the features extracted from the training dataset. Validation dataset utilized for classifier parameter selection process and the recognition performance of the built classifier models are evaluated on the test dataset. The recognition accuracy obtained for the considered recognizers on the training, validation and test dataset are listed in Table 4.

Among the considered feature descriptors, IMG and SVD work on the vector representation of input images. GF and CT are global or image level feature descriptors which obtains global feature descriptors from transform domain representations of character images, while HOG, RLC, ScatCN and Lenet-5 performs block or local region level processing for extracting informative feature descriptors. The local region level feature descriptors, either in linear or non-linear classifier acquired better performance compared to vector level or global level feature descriptors. The localized feature descriptors are more effective in capturing invariant feature descriptors where the variability among input patterns are high.

In case of training recognition accuracy, most of the recognizers obtained very high recognition accuracy excluding very few recognizers. Testing and validation accuracy determines the generalization of built recognizer models compared to training accuracy, and rest of the discussion focus on performance on testing and validation dataset. Among all considered feature descriptors, ScatCN based feature descriptors have acquired higher recognition accuracy in testing and validation dataset. The ScatCN based features have almost comparable performance in both linear and RBF SVM classifier compared with other feature descriptors (excluding Lenet-5). ScatCN features have the capability to capture invariant features from input images as the non-linear modulus and averaging functions are employed in the scattering transform [37]. ScatCN features are stable with small deformations and the non-linear nature of those features helps the linear SVM classifier to achieve comparable or high recognition performance as that of RBF SVM classifier model.

Lenet-5 CNN architecture obtained the second highest testing and validation recognition accuracy after ScatCN based features. CNN architectures are very powerful in learning the invariant feature descriptor from images with the support of train-able filter banks. But in this paper, as we have utilized the Lenet-5 architecture with same number of hidden layers and the same number of convolutional feature maps as that of [38] for Malayalam character recognition, it could obtain only slight performance improvement compared to HOG feature descriptors. CNN architectures are utilizing self learned features, and well optimized and well tuned CNN architectures are capable of attaining the comparable or even higher recognition performance as that of hand-crafted feature descriptors. Works on designing appropriate CNN architecture and parameters for Amrita_MalCharDb is on progress.

The misclassifications happened for ReducedScatCN recognizer on testing dataset are analyzed with the help of confusion matrix. Fig. 4 shows the visualization of the misclassified instances, and the number of misclassifications are highlighted based on the color formatting (minimum misclassification with No-color and maximum misclassification with Black). In Fig. 4, four entries are very dark which denotes the presence of greater number of misclassification. Table 5 lists the classes representing those four entries in confusion matrix, and it can be seen that the misclassifications are happened between character classes having very strong structural similarity.

Fig. 5 shows the top-N recognition accuracy measures for all the recognizers with N = 1,2,3 and 5. For all recognizers, the increase in recognition accuracy when N = 2 is marginally greater compared to 3, and 5. The average improvement in recognition accuracy among the recognizers for top-2 is 9.73% while for top-3 is 3.71%. This is due to the presence of similarly shaped classes in the database. The recognition accuracy of ReducedScatCN increases from 91.05%

[2] The character class labels are the indices for the confusion matrix. Zoom the figure for more clear display of the class labels and misclassification values.

Table 5
Misclassified classes for ReducedScatCN.

True Class	Predicted Class
◌ം	ഠ
ള	ഉ
ണ	ന്ന
ന	◌ൗ

to 96.34% when top-2 classes are considered. When top-5 classes are considered, the recognition accuracy reaches 98.98%. Most of the misclassifications for the classes listed on Table 5 are solved when taking top-5 candidates. The classes in first, second and fourth row of the Table 5 are resolved completely in top-5 candidates.

The scattering based methods have higher feature dimension compared to other feature extraction methods utilized in our experiments. When considering the time complexity of the feature extraction methods, ReducedScatCN based features higher time complexity as they make use of heirachical features and SVD technique, but those methods have higher recognition performance compared to other methods. Nowadays parallel processing architectures is becoming popular, and utilizing those architectures can speed up the training process of these methods. Fig. 6 shows the computational time taken by the different recognizers in SVM classifier for classifying the test data of Malayalam handwritten database. From the figure, it is evident that the ScatCN and ReducedScatCN require more computation time compared to other feature descriptors, but are effective in obtaining far better recognition accuracy.

The current work is an initial attempt towards building a character based document image database for Malayalam language script. In order to make a complete character based recognition system, the dataset needs to be extended by including all the possible Malayalam character shapes. More handwritten samples can be included in the database by extending data collection to more number of writers. Besides the character images, the document images containing Malayalam text in word, line and page level can be collected so that the researchers can work up pre-processing, segmentation and post-processing stages for Malayalam handwritten text recognition.

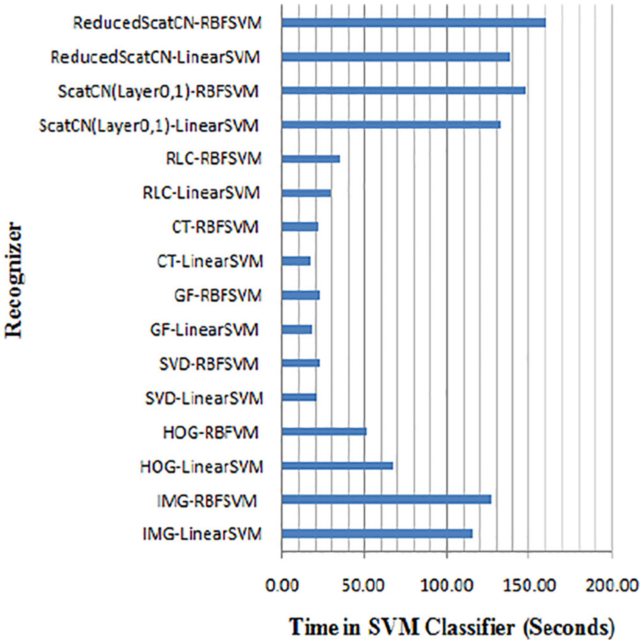


Fig. 6. Execution time on the SVM classifier for test dataset of Amrita_MalCharDb.

6. Conclusion

Handwritten character recognition is a challenging pattern recognition problem due to the high degree of intra-class and inter-class variability among handwritten character image patterns. Benchmark databases are necessary and are the foremost

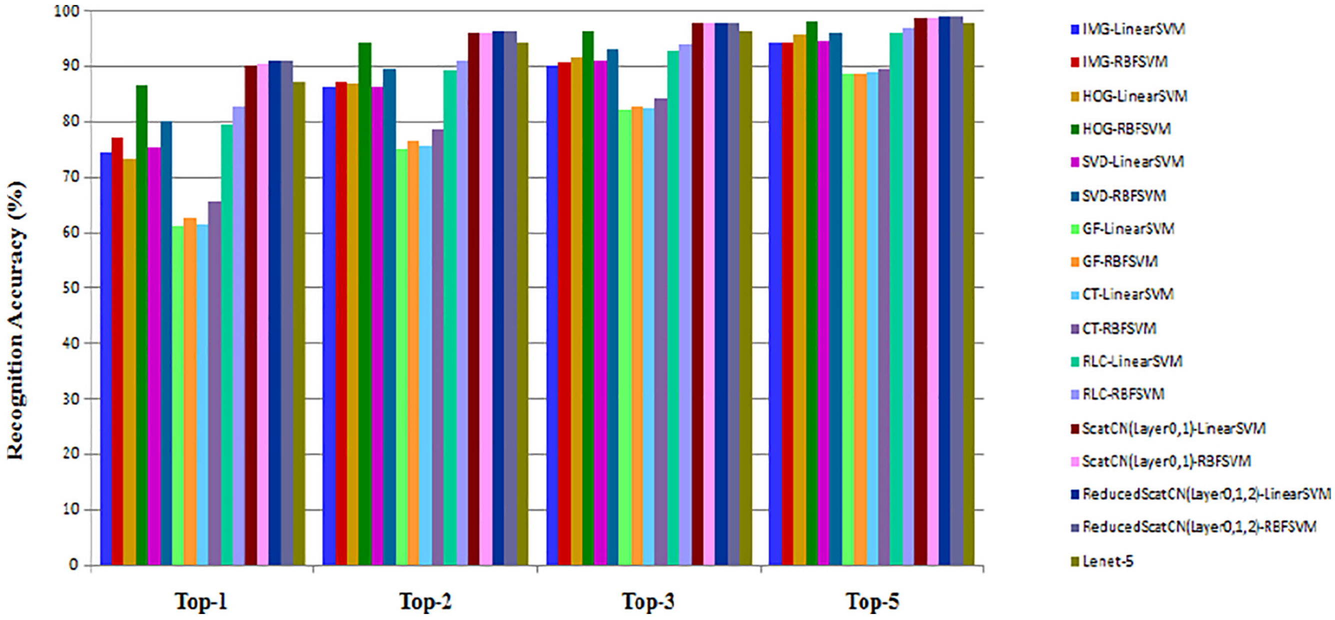


Fig. 5. Top-N recognition accuracy for the recognizers on Amrita_MalCharDb.

requirement for implementing any recognition system. Lack of standard document image resources is one of the problem in handwritten document recognition for most of the Indian language scripts. The present paper focuses on implementing a character level image database for Malayalam language script, one of the official language in India. Till date, no public domain document image database is available for Malayalam language. The character classes are formed based on the unique orthographic character shapes present in Malayalam script. 85 Malayalam character classes including Malayalam vowels, consonants, half consonants, vowel & consonant modifiers and conjunct characters are considered for database creation. Handwritten image data collected from 77 native Malayalam writers, and active contour model based minimization technique is employed for character segmentation. The present Malayalam character image database, Amrita_MalCharDb contains 29,302 Malayalam character image patterns. Recognition performance of Amrita_MalCharDb is evaluated by using different feature extraction techniques. Scattering convolutional network based features could achieve 91.05% recognition accuracy among considered techniques. Future works on the database includes extending the character class collection by including all the presently used valid orthographic shapes in Malayalam language script and creating word, line and page level collection of Malayalam document images so that the researchers can focus on other stages of document recognition system as well.

References

[1] V. Govindan, A. Shivaprasad, Character recognition – a review, Pattern Recogn. 23 (7) (1990) 671–683.

[2] D.-H. Kim, Y.-S. Hwang, S.-T. Park, E.-J. Kim, S.-H. Paek, S.-Y. BANG, Handwritten korean character image database pe92, IEICE Trans. Inf. Syst. 79 (7) (1996) 943–950.

[3] P.J. Grother, Nist special database 19 handprinted forms and characters database, National Institute of Standards and Technology.

[4] U.-V. Marti, H. Bunke, The IAM-database: an english sentence database for offline handwriting recognition, Int. J. Doc. Anal. Recogn. 5 (1) (2002) 39–46.

[5] Y. Al-Ohali, M. Cheriet, C. Suen, Databases for recognition of handwritten arabic cheques, Pattern Recogn. 36 (1) (2003) 111–121.

[6] S.A. Mahmoud, I. Ahmad, M. Alshayeb, W.G. Al-Khatib, M.T. Parvez, G.A. Fink, V. Märgner, H. El Abed, Khatt: Arabic offline handwritten text database, in: International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2012, pp. 449–454.

[7] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, Casia online and offline chinese handwriting databases, in: International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2011, pp. 37–41.

[8] T. Su, T. Zhang, D. Guan, Corpus-based hit-mw database for offline recognition of general-purpose chinese handwritten text, Int. J. Document Anal. Recognit. 10 (1) (2007) 27.

[9] T. Saito, H. Yamada, K. Yamamoto, On the data base ETL9 of handprinted characters in JIS chinese characters and its analysis, IEICE Trans. 68 (4) (1985) 757–764.

[10] ISI Indian script databases,http://www.isical.ac.in/ujjwal/download/database.html (accessed: 2017-12-06).

[11] CMATERdb,https://code.google.com/archive/p/cmaterdb/ (accessed: 2017-12-06).

[12] HP Labs India Indic Handwriting Datasets,http://lipitk.sourceforge.net/hpl-datasets.htm (accessed: 2017-12-06).

[13] V. Govindaraju, S. Setlur, Guide to OCR for Indic Scripts, Springer, 2009.

[14] S.M. Obaidullah, C. Halder, K. Santosh, N. Das, K. Roy, Phdindic_11: page-level handwritten document image dataset of 11 official indic scripts for script identification, Multimedia Tools Appl. (2017) 1–36.

[15] U. Pal, B. Chaudhuri, Indian script character recognition: a survey, Pattern Recogn. 37 (9) (2004) 1887–1899.

[16] M.A. Rahiman, A. Shajan, A. Elizabeth, M. Divya, G.M. Kumar, M. Rajasree, Isolated handwritten Malayalam character recognition using HLH intensity patterns, in: Second International Conference on Machine Learning and Computing (ICMLC), IEEE, 2010, pp. 147–151.

[17] J. John, K. Pramod, K. Balakrishnan, Offline handwritten Malayalam character recognition based on chain code histogram, International Conference on Emerging Trends in Electrical and Computer Technology, ICETECT 2011, IEEE, 2011, pp. 736–741.

[18] B.S. Moni, G. Raju, Modified quadratic classifier for handwritten Malayalam character recognition using run length count, in: International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), IEEE, 2011, pp. 600–604.

[19] V. Vidya, T. Indhu, V. Bhadran, R. Ravindra Kumar, Malayalam offline handwritten recognition using probabilistic simplified fuzzy ARTMAP, Intell. Inf. (2013) 273–283.

[20] S.S. Kumar, K. Manjusha, K. Soman, Novel SVD based character recognition approach for Malayalam language script, in: Recent Adv. Intell. Inf., Springer, 2014, pp. 435–442.

[21] J. Jomy, K. Balakrishnan, K. Pramod, A system for offline recognition of handwritten characters in Malayalam script, Int. J. Image Graphics Signal Processing 5 (4) (2013) 53.

[22] G. Raju, B.S. Moni, M.S. Nair, A novel handwritten character recognition system using gradient based features and run length count, Sadhana 39 (6) (2014) 1333–1355.

[23] G. Raju, Recognition of unconstrained handwritten Malayalam characters using zero-crossing of wavelet coefficients, in: International Conference on Advanced Computing and Communications, ADCOM 2006, IEEE, 2006, pp. 217–221.

[24] R. John, G. Raju, D. Guru, 1D wavelet transform of projection profiles for isolated handwritten Malayalam character recognition, International Conference on Computational Intelligence and Multimedia Applications, vol. 2, IEEE, 2007, pp. 481–485.

[25] B.P. Chacko, V.V. Krishnan, G. Raju, P.B. Anto, Handwritten character recognition using wavelet energy and extreme learning machine, Int. J. Mach. Learn. Cybern. 3 (2) (2012) 149–161.

[26] J. John, K. Pramod, K. Balakrishnan, Unconstrained handwritten Malayalam character recognition using wavelet transform and support vector machine classifier, Procedia Eng. 30 (2012) 598–605.

[27] M. Manuel, S. Saidas, Handwritten Malayalam character recognition using curvelet transform and ANN, Int. J. Comput. Appl. 121(6).

[28] K. Manjusha, M.A. Kumar, K. Soman, Reduced scattering representation for Malayalam character recognition, Arab. J. Sci. Eng. (2017) 1–12.

[29] N.N.V, C.V. Jawahar, Empirical evaluation of character classification schemes, in: Seventh International Conference on Advances in Pattern Recognition, IEEE Computer Society, 2009, pp. 310–313.

[30] K. Soman, R. Ramanathan, Digital Signal and Image Processing-the Sparse Way, first ed., Elsevier India.

[31] T. Goldstein, X. Bresson, S. Osher, Geometric applications of the split bregman method: segmentation and surface reconstruction, J. Sci. Comput. 45 (1–3) (2010) 272–293.

[32] K. Syama, N. George, S. Sekhar, C. Neethu, M.S. Manikandan, K. Soman, Performance study of active contour model based character segmentation with nonlinear diffusion, in: International Conference on Advances in Computing and Communications (ICACC), IEEE, 2012, pp. 118–121.

[33] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.

[34] Ø. Due Trier, A.K. Jain, T. Taxt, Feature extraction methods for character recognition-a survey, Pattern Recogn. 29 (4) (1996) 641–662.

[35] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, IEEE, 2005, pp. 886–893.

[36] R. Ramanathan, A. Nair, L. Thaneshwaran, S. Ponmathavan, N. Valliappan, K. Soman, Robust feature extraction technique for optical character recognition, ACT 2009 – International Conference on Advances in Computing, Control and Telecommunication Technologies (2009) 573–575. https://doi.org/10.1109/ACT.2009.145.

[37] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886. arXiv:1203.1513.

[38] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324. arXiv:1102.0183.

[39] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.

[40] F. Chollet, et al., Keras, https://github.com/fchollet/keras, 2015.

[41] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org, 2015. https://www.tensorflow.org/.