Int. j. inf. tecnol. (March 2020) 12(1):57–64
https://doi.org/10.1007/s41870-019-00366-y

ORIGINAL RESEARCH

Effect of supervised learning methodologies in offline handwritten Thai character recognition

Ferdin Joe John Joseph[1]

Received: 10 January 2019 / Accepted: 24 September 2019 / Published online: 1 October 2019

Abstract Offline handwritten character recognition is a conversion process of handwriting into machine-encoded text and predominantly used for digitizing handwritten texts and forensic applications. Currently, several techniques and methods are proposed to enhance accuracy of offline handwritten character recognition for many languages spoken across the globe like English, Tamil, Chinese and Arabic. In this paper, a local feature-based approach using supervised learning techniques is proposed to enhance the accuracy of handwritten offline character recognition for Thai alphabets using unsupervised learning for individual character as a class, whereas most of the existing methodologies for Thai character recognition is done with group of similarly looking characters as a class. The classification is operated by using support vector machine (SVM). The accuracy would be the percentage of correct classification for each class. For the result, the highest accuracy is 74.32% which has 144-bit shape features and uniform pattern LBP for the features.

Keywords Offline character recognition · Local binary pattern · Thai handwriting

1 Introduction

Computer vision is used for the past couple of decades to solve problems related to images and video with the combination of machine learning and statistics. It is a study that deals with computer development in digital images analysis, mimic human's visual ability. An example of this field is face detection system [1] that can be used in criminal investigation. Character recognition is also a branch of computer vision; it is a conversion of printed or handwritten texts into digital ones. Technically it is known as Offline/Online Character Recognition, where offline corresponds to the detection of handwritten text on paper and online is the characters scribed on smart devices like tablets and smart phones. Research in this field has been popularly done with major languages such as English [2], Tamil [3, 4], Chinese [5] and Arabic [6]. Each language has its uniqueness of shape and curve, so effective recognition techniques are selected differently. Methodology holding good for one language may not perform same for another language.

The purpose of this research project is to try various types Local Binary Pattern (LBP) descriptors [7], which are widely used in texture classification as a visual descriptor, with character recognition. Without grouping identical characters and loops before classification, optimal condition for features such as size of shape matrix and an addition of binarization during preprocessing were determined to obtain better performance possible.

Thai language is the mother tongue of over 69 million people [8] living in Thailand and it has a script which follows vatteluthu format [9]. The author also expected that a new framework for automated offline handwritten Thai character recognition to be proposed for a new dataset available online for free and be further developed by other researchers. Thus, an application that detects Thai characters with their unicodes, translate from Thai to other languages, such as English, from a photo could be possible with a robust method for character image recognition.

✉ Ferdin Joe John Joseph
ferdinjoe@gmail.com

[1] Faculty of Information Technology, Thai-Nichi Institute of Technology, Bangkok, Thailand

Consequently, it is an opportunity for the authors to study about the performance of OCR and evaluating the performance using the most popular image processing platform MATLAB and review state-of-the-art methodologies for character recognition, especially for Thai handwritten alphabets. The attempt of the methodologies proposed is aimed at using the minimum number of features for quicker and better classification of characters. The authors also explored the performance of Thai OCR using support vector machine (SVM). SVM is opted as a classifier and after studying various types listed in [10], radial basis kernel is used to perform classification on the images used in the experiments reported in the proposed methodology.

2 Related work

Unlike languages like English [2], Tamil [3, 4], Chinese [5] and Arabic [6], Hindi to Dogri [11] Thai OCR has relatively fewer literatures available as existing technologies. Thai alphabets in its written form has unique characteristics such as loops, curls and junctions. Some alphabets only differ by the existence of a curl, leads to an anomaly and they will be much more similar from an informal writing. It is very challenging to make robust error-free handwritten Thai character recognition. There have been many feature extraction techniques, frameworks and classifiers proposed to enhance the accuracy of handwritten Thai character recognition such as Genetic Algorithm (GA) [12], Neural Network [13], Single classifier with global features alone [14], Single classifier with global and local features [14] and Query Matching [15]. An appropriate combination of feature extraction techniques could also result a significantly high accuracy. For example, pixel distribution [14] was used to sort handwritten Thai characters into subgroups, then combine with structural features, modified for specific subgroup, to classify them.

A handful of research has been done so far in the case of Thai Character Recognition in offline mode. Most of the existing methodologies like [13] proposed ant-miner based classifier which divided the entire Thai alphabet into three sets namely upper, middle and lower. These sets are taken as class and thereby having three classes. The individual character in each class will be taken as a sub class. The recognition rate reported in this methodology may not be justifiable since, there are some characters which look similar but not the same. These characters when examined using a class for each character, may go down in performance.

Pornpanomchai et al. [12] used a genetic algorithm-based methodology. This methodology has no qualitative performance reported against classification accuracy in the literature. Moreover, genetic algorithm is time consuming as it predicts over infinite loops of generations. The authors mention that their proposed methodology has a challenge in segmentation. They separate the loops and curves from each character based on a template of spaces. This is not good enough when the same character is written in different styles by same or different person. Moreover, the dataset used by them is not available for the authors of this paper to validate. Since the methodologies give different performance measures for different dataset. This is due to the fact that the contributors of the datasets may not be the same and may not have the sampling representation of different age groups. Moreover, the outcome of genetic algorithm is not the same for all the trials. This uncertainty is not advisable for character recognition of massive handwritten text.

Methasate et al. [14] proposed a framework which used various parameters. It proposed a methodology having global features, local features combined with the former, single classifier and multiple classifier. For OCR software, the methodology needs to be simple and optimal enough for quicker recognition. Multi class classifier is time consuming though the performance figures are convincing enough.

John Joseph and Anantaprayoon [15] proposed a framework with local features and query matching as classifier. The experiments reported were for 64 and 59 bin feature vectors and their combination with query matching. Euclidean distance was used to maximize the distance between the classes. The performance evaluated was quite convincing as it uses separate classes for each character and the performance was below 70%. This validation was done with tenfold cross validation.

Because of all the existing methodologies taken here, we propose a new framework which has a single tier of classifier and uses only local feature. The dataset used for validation of the proposed methodology uses an independent dataset which is freely available online.

In this paper, we proposed to use rotation invariant Local Binary Pattern (LBP) as a main visual descriptor while using shape matrix [15] as an additional feature for handwritten Thai individual alphabets. Many types of LBP were varied, and their average accuracies were compared to determine the one that yield the maximum. Query matching which is similar to [16] is used to classify in this system.

3 Proposed methodology

The framework is proposed for classifying Offline Thai characters as given Fig. 1 below.

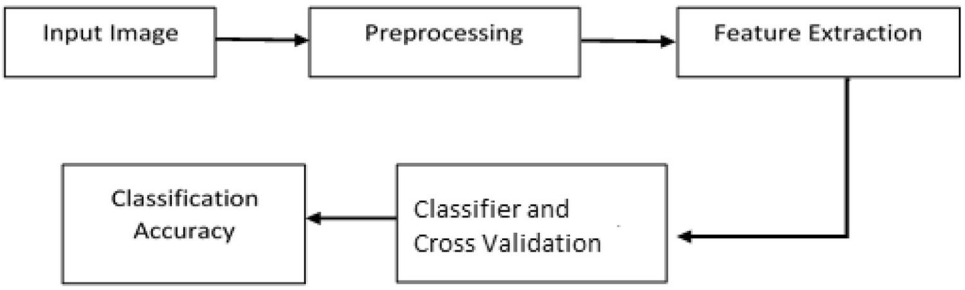

Fig. 1 Proposed framework for offline handwritten character recognition

3.1 Preprocessing

The given input image is converted to grayscale using the gradient based function available in Matlab library and then subjected to 2D median filtering. This filtering is done to remove noise from the input image. Noise reduction reduces the unwanted fragments of ink and external imprints like fingerprints and dust from the document while scanning. The resulting image is subjected to segmentation. It is done to remove whitespaces around the character and the region corresponding to the character's coverage is cropped as Region of Interest (ROI). The 2D filtered image is subjected to edge detection using canny detector. From the edges detected, the rows and columns in the image corresponding to zero summation are removed and those above zero are retained. This removes the faulty ink imprints created by the user. The resultant cropped image is now resized to the nearest integer divisible by 11 for both height and width. This is done for easier management of certain features to extract and is explained in the next phase of framework (Fig. 2).

3.2 Feature extraction

Texture and shape features are extracted for efficient classification of characters using query matching. The objective of feature extraction in the proposed framework is to increase interclass distance and minimize intra-class distance. Texture feature includes Local Binary Pattern (LBP) [7] in 2 dimensional level and shape feature includes 11×11 shape matrix.

The LBP features are extracted from the input images converted to grayscale.

Shape feature consists of the shape matrix calculated [17] as similar to Fig. 3. The given input image which has edges detected is placed in a 11×11 grid and checked on which cells does the character pass through. If the character pass through a cell, then it is marked as 1, else 0. A shape matrix is obtained because of this process. The shape

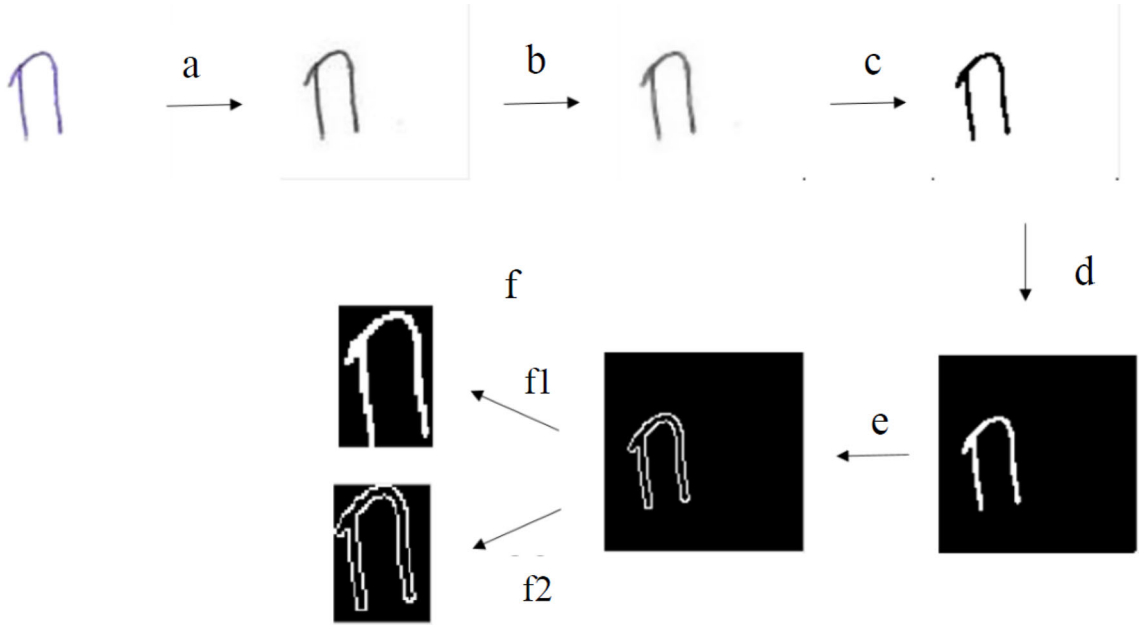

Fig. 2 Preprocessing of input image. **a** Grayscale conversion, **b** 2D median filtering, **c** binarization, **d** swapping of binary values, **e** edge detection, **f** image cropping, (f1) LBP calculation and (f2) shape matrices calculation

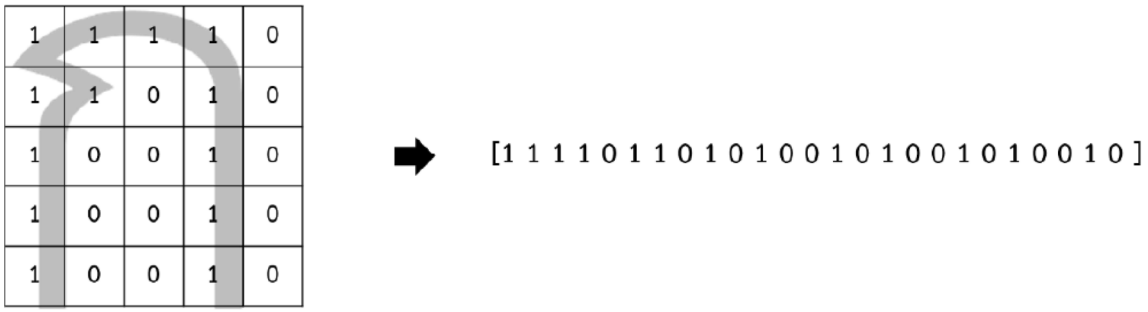

Fig. 3 Shape feature extraction

matrix is then converted to a single dimensional array and concatenated with the normalized LBP features. This results in a 180-vectored feature set corresponding to the segmented character from the input image. The resulting feature set is stored along with their corresponding labels. Since 44 consonants from Thai alphabets are taken to the experiment, the number of classes remain the same quantity. Taking 11 as the cell count for both length and breadth of the character is decided based on the performance yielded in the preliminary experimentation trials. The features are in sparse form and are calculated with subspaces using HoVer representation [18] and normalized. The normalized features are ranked on the basis of its dominance using the regression based model developed in [19]. The resultance dominance values of features are arranged and taken for training and testing.

Rotation invariant LBP is used for the texture feature in this proposed methodology and is illustrated in Fig. 5. Two LBP values are said to be the same if their shift of binary patterns are equal. For example, 100101102 is equivalent to 101101002. In this kind of LBP, equivalent LBPs are counted in the same bin in histogram by choosing the least decimal value from each equivalent set of local features. This is done to obtain same or similar histogram even if the image is tilted to a particular angle. In the case of our proposed methodology, rotation invariant LBP is used to ignore the tilt of writing characters. Every user will have a different tilt while writing a character and it is by their psychological orientation. Rotation invariant features are used to overcome this orientation.

$$f(I) = \sum t(I) U \sum s(I) \quad (1)$$

where f(I) is the feature vector if the given image, while t(I) and s(I) are the texture and shape vectors of the same image.

$$t(I) = LBP^{ri}(I) \quad (2)$$

$$s(I) = y \times y \text{ matrix of shape} \quad (3)$$

y in this proposed methodology is taken as 12 as it yielded maximum performance in the preliminary experimentation. So the shape feature takes a length of 144 vectors. This is combined with 59 vectors of $LBP^{ri}(I)$. So the total number of vectors extracted for an input image comes to 193.

4 Experiment

4.1 Kamnoetvidya Science Academy Thai OCR Dataset

Kamnoetvidya Science Academy (KVIS) Thai OCR Dataset was created by the authors of this paper. Handwritten version of all 44 Thai characters are obtained from various individuals and scanned. Each individual contributor of this dataset is a citizen of the Kingdom of Thailand and has learned Thai language at least up to their high school level. So the language proficiency of the contributors has no issue with the dataset. A questionnaire as shown in Fig. 4 was given to all the contributors and the data was collected. The obtained questionnaire is compiled and scanned using a normal scanning machine with same lighting condition and magnification. The images obtained using the scanning process are then stored with file names containing the type of letter and the serial number of sample. The resultant dataset consists of 1079 images from 44 classes (letters). This dataset consists of all Thai consonants with different writing styles of various people from different age groups. The contributors of this dataset are from ages between 16 and 75. The contributors of the dataset includes the instructors, students and the latter's parents from the first three cohorts of Kamnoetvidya Science Academy in Wangchan, Rayong. The dataset is obtained from the contributors after getting their consent and copyright transfer to use their handwritten data by anyone who use the dataset for research purposes only. Sheet of questionnaire was created to ask the contributors to write the specified Thai consonant. Each contributor was asked to write every consonant on the sheet of questionnaire. The dataset is freely available online in [15] for prospective researchers who would like to develop new methodologies and paradigms for OCR in Thai language.

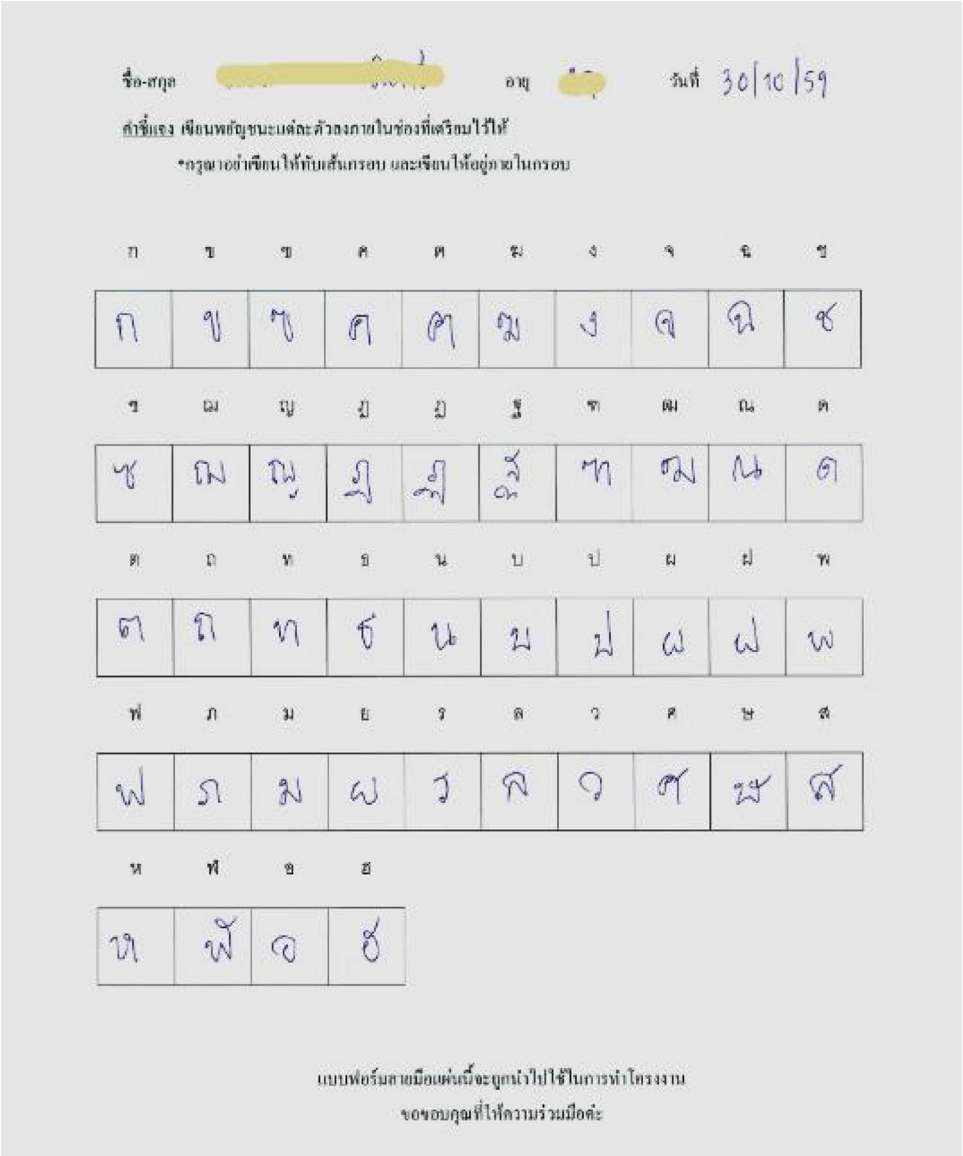

Fig. 4 Sample questionnaire form

As of now, this is the only dataset available online and available to download for research purposes.

4.2 Support vector machine (SVM)

The test images are subjected to feature extraction and then subjected to classification using SVM with the database of feature sets of training and labels. The similarity index of test image is calculated against all the feature sets stored in the database according to the kernel function calculated using the radial basis function (rbf). Euclidean distance [20] between the vectors taken for our earlier method of query matching is not sufficient enough for inter class distance maximization. The class label of feature set corresponding to the highest recorded similarity index is taken as resultant recognition of the character present in the input image. Support Vector Machine (SVM) [21] was considered to use with single class and multiclass SVM. Single class SVM holds better than the multi class SVM as a result of preliminary experimentation. The preliminary experimentation used LIBSVM [22] as a tool to validate SVM. SVM is chosen based on the preliminary experimentation of SVM against various other machine learning classifiers like K-NN, Naïve Bayes and Decision trees.

4.3 Validation

Tenfold cross validation is used to validate the accuracy obtained in the classification. This validation is done for each character in the dataset. For the images obtained for each character 10% of the images are taken for testing and remaining 90% are taken for training. This is done ten

times until all images are subjected to testing. The resultant accuracy from each fold is recorded and the mean is taken as the accuracy of the particular character as a class. This process is done for all the characters. So the individual accuracy is obtained for all 44 consonants. This accuracy from all characters are taken a mean to quote the accuracy of this proposed methodology.

4.4 Implementation

Matlab is used to implement the proposed framework on KVIS Thai OCR Dataset. The experiments are carried out in a standalone machine with a normal configuration. The performance evaluated in existing method and the proposed methodology used KVIS Thai OCR Dataset [15]. Since the other existing methodologies use various datasets which are not practically accessible as there is no standard procedure in data collection, availability or procedure, the performance claimed in their papers are directly quoted in the existing methodology section in this paper.

5 Results

The above screenshot in Fig. 5 is obtained by testing sample images using a GUI developed using MATLAB. For performance evaluation, the dataset need to be refined by removing unclear samples.

Qualitative performance of the proposed methodology is listed in Table 1. Methodologies of Ant Miner and Single stage classifier in [13] and [14] looks like giving better performance than the proposed methodology. Ant-Miner Algorithm proposed in [13] is time consuming than our proposed method. The base for classification of characters is entirely different from the one proposed using local features. The method proposed in [14] used both global and local features. Our proposed method is focused only on local features. The dataset used [12–14] are not available currently for experimental validation online or for free.

The dataset taken for experimentation of the proposed methodology is from real users and it is observed that, most of the users write few of the characters either similar or wrong as other characters. The results of the proposed methodologies are validated using tenfold cross validation. The accuracy of some characters like ก, ป, ฑ and อ are above 80% but characters พ, ผ and ฟ are like similarly looking characters but different classes reported less performance. These characters are grouped as single class and reported higher performance in most of the existing methodologies. The performance of the proposed methodologies looks significant because, these methodologies use individual characters as classes. So, we have 44 classes for this dataset. The existing methodologies used grouping of similarly looking characters as a class. Grouping is mentioned in most of the existing methodologies and the proposed method uses individual character class. Those methods with grouping methods use a set of similarly looking characters into one single class. If the existing methodologies use individual character class, the accuracy of those methods will drop below the proposed

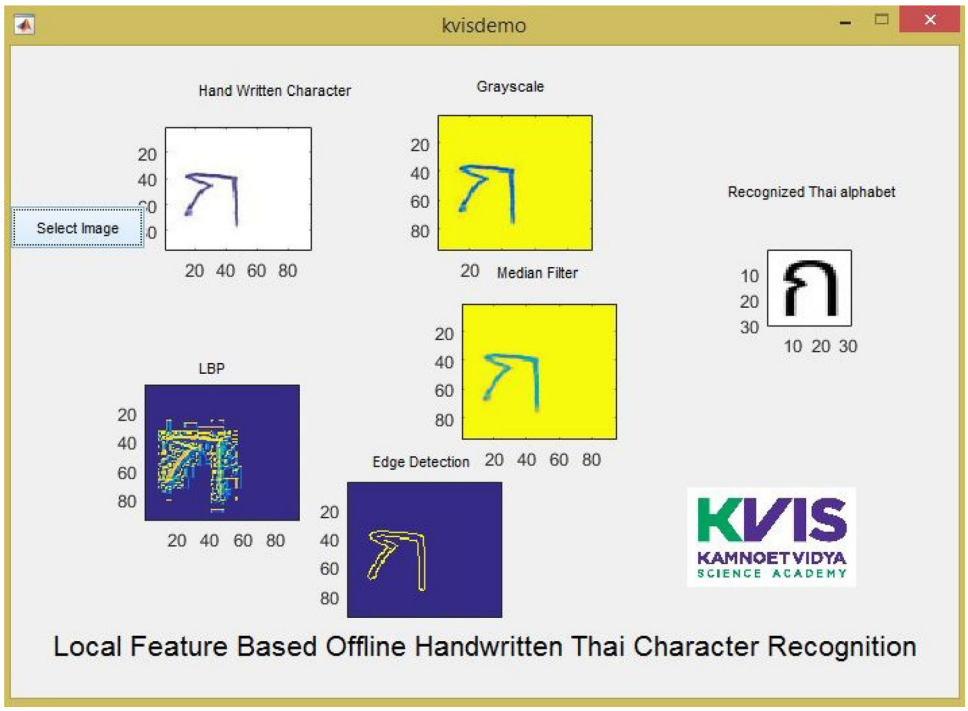

Fig. 5 Implementation in matlab

Table 1 Performance of proposed methodologies over various methodologies

S. no	Methodology	Classification accuracy	Classification basis of characters
1	Ant miner algorithm [13]	82.7%	Grouping
2	Genetic algorithm [12]	NA	Grouping
3	Single stage classifier with global features alone [14]	54.61%	Grouping
4	Single stage classifier with global and local features [14]	78.89%	Grouping
5	Query matching using 64 bin LBP [15]	68.82%	Individual
6	Query Matching using 59 bin LBP [15]	68.96%	Individual
7	Single class SVM with rbf kernel using 59 bin LBP (proposed)	74.32%	Individual

methodologies. Since the datasets used for those methods and the experimental parameters are missing in those papers, those methodologies using individual character classes are not possible to produce. The number of bins shown in the proposed methodology is 59. This value is chosen based on the preliminary experimentation with various values possible to the number of bins. During the experimentation, the accuracy was higher than any for when the number of bins is set to 59. This is due to the fact that the visual words created in this configuration suits well with the radial basis function based calculation.

6 Conclusion

A framework using local features is proposed for classifying handwritten Offline Thai Character Recognition and evaluated the performance using SVM. This framework uses both texture and shape feature combined and performs better recognition of handwritten offline Thai characters while each Thai character is taken as an individual class. As far as local features and single character-based classifier level, the proposed methodology holds significant performance. There are some criteria which can be explored with different methodologies and improve the performance. It is suggested that collecting more sample images for training will increase recognition efficiency as we know that the more training, the more capability in classification. Large amount of samples also elevates reliability of results since it is an indicator of people's handwriting tendency. The dataset can be improved by adding vowels and combination of vowels and consonants and thereby increasing the number of classes in the dataset.

References

1. Tembe AU, Thombre SS (2017) Survey of copy-paste forgery detection in digital image forensic. In 2017 international conference on innovative mechanisms for industry applications (ICIMIA), pp 248–252
2. Singh S (2013) Optical character recognition techniques: a survey. J Emerg Trends Comput Inf Sci 4(6):545–550
3. Kannan RJ, Prabhakar R (2009) A comparative study of optical character recognition for tamil script. Eur J Sci Res 35(4):570–582
4. Ravi FJJT, Velayutham PR (2010) Effective tamil character recognition in tablet PCs using pattern recognition. In Tamil Internet Conference
5. Hildebrandt TH, Liu W (1993) Optical recognition of handwritten Chinese characters: advances since 1980. Pattern Recognit 26 (2):205–225
6. Raouf AMA (2012) Offline printed Arabic character recognition. University of Nottingham, Nottingham
7. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on feature distributions. Pattern Recognit 29(1):51–59
8. Thailand Population (2018) Worldometers. http://www.worldometers.info/world-population/thailand-population/. Accessed 04 June 2018
9. Thai Language", Wikipedia. https://en.wikipedia.org/wiki/Thai_language. Accessed 04 Nov 2016
10. Chandra MA, Bedi SS (2018) Survey on SVM and their application in image classification. Int J Inf Technol. https://doi.org/10.1007/s41870-017-0080-1
11. Dubey P (2019) The Hindi to Dogri machine translation system: grammatical perspective. Int J Inf Technol 11(1):171–182
12. Pornpanomchai C, Wongsawangtham V, Jeungudomporn S, Chatsumpun N (2011) Thai handwritten character recognition by genetic algorithm (THCRGA). Int J Eng Technol 3(2):148–153
13. Phokharatkul P, Sankhuangaw K, Somkuarnpanit S, Phaiboon S, Kimpan C (2005) Off-line hand written Thai character recognition using ant-miner algorithm. Int J Comput Electron Autom Control Inf Eng 8(1):276–281
14. Methasate I, Marukatat S, Sae-Tang S, Theeramunkong T (2005) The feature combination technique for off-line Thai character recognition system. In: Proceedings of the international conference on document analysis and recognition, ICDAR, vol 2005: 1006–1009
15. Joseph FJJ, Anantaprayoon P (2018) Offline handwritten Thai character recognition using single tier classifier and local features. In: 2018 international conference on information technology (InCIT), pp 1–4
16. Ahmad K, Sahu M, Shrivastava M, Rizvi MA, Jain V (2018) An efficient image retrieval tool: query based image management system. Int J Inf Technol. https://doi.org/10.1007/s41870-018-0198-9
17. Joseph FJJ, Auwatanamongkol S (2016) A crowding multi-objective genetic algorithm for image parsing. Neural Comput Appl 27(8):2217–2227
18. Joseph FJJ, Ravi T, Justus C (2011) Classification of correlated subspaces using HoVer representation of Census Data. In: 2011

international conference on emerging trends in electrical and computer technology, pp 906–911
19. Joseph FJJ (2019) Empirical dominance of features for predictive analytics of particulate matter pollution in Thailand. In: 5th Thai-Nichi Institute of Technology Academic Conference TNIAC 2019, pp 385–388
20. Deza MM, Deza E (eds) (2009) Encyclopedia of distances. Springer, Berlin, Heidelberg, pp 1–583
21. Hsu C, Chang C, Lin C (2010) A practical guide to support vector classification
22. Chang C, Lin C (2013) LIBSVM: a library of support vector machines