International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 7; Issue 5(E); May 2018; Page No. 12518-12521
DOI: http://dx.doi.org/10.24327/ijcar.2018.12521.2203

Research Article

AN ADAPTIVE FEATURE EXTRACTION APPROACH FOR IMPROVING HANDWRITTEN CHARACTER RECOGNITION

Krishnendu K R[1]* and Nikhila T Bhuvan[2]

Department of Information Technology, Rajagiri School of Engineering and Technology, Kakkanad, Kerala, India

A R T I C L E I N F O

Article History:

Received 6th February, 2018
Received in revised form 20th March, 2018 Accepted 8th April, 2018
Published online 28th May, 2018

Key words:

HoG; SIFT; handwritten recognition; machine learning; SVM

A B S T R A C T

Handwritten character recognition, one of the challenging problem in machine learning field and it still has got so many scopes to work on because of its huge applications. Some of that application includes, Postal address and zip code verification, writer identification, bank cheque processing and so on. Even though there are many handwritten recognition systems are developed, none gives better results when it faces more than one line of handwritten text. This work proposes a technique to improve the accuracy of handwritten recognition system. The proposed approach basically focuses on an adaptive feature extraction techniques based on HoG (Histogram of oriented gradients) and SIFT(Scale invariant feature transform).

INTRODUCTION

Handwritten recognition is a challenging problem in machine learning field. Machine learning field evolved through lots of challenging problems like this. The ability of a machine to learn itself and then take appropriate decisions based on learned information changed the perspective of future world. In future, each and every object will be an automated one. As these technology grows, maintenance and improvements of the developed systems take a major role.

Why a handwritten recognition system is needed, in order to save the time and to reduce the work load of humans. That is the major concern of all the machine learning related problems. Handwritten recognition becomes one of the challenging problem because, different people have different handwriting styles. Basically handwritten recognition system can be develop in two ways that are off-line recognition systems and on-line recognition systems. Off-line handwriting recognition involves detection and conversion of text in an image into letters which are usable within computer and text-processing applications[7]. Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles. On-line handwriting recognition involves the automatic conversion of text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. Handwritten recognition involves processes such as pre-processing of captures image, segmentation of characters, feature extraction from detected character, classification and recognition. Role of feature extraction is really big in recognition, because choosing the right features from the image can improve the accuracy of the recognition system. This proposed work mainly focuses in the area feature extraction techniques to improve the recognition rate. Two feature extraction techniques are using in this, HoG and SIFT. Both feature extractions got little disadvantages such as HoG is suitable to recognize characters that have sharp and pointed edges where as SIFT suitable to recognize characters with loops and smooth curves. So, combination of these two feature extraction will result in a better handwritten recognition system.

Corresponding author:* **Krishnendu K R
Department of Information Technology, Rajagiri School of Engineering and Technology, Kakkanad, Kerala, India

Related Work

Madhuri Yadav and Dr. Ravindra Purwar did Hindi handwritten recognition using multiple classifiers[1]. They created a database consisting of 4,428 handwritten character samples written by different individuals and 108 samples have been taken for each individual character to ensure different orientations and size. For feature extraction they used histogram of oriented gradients as one feature and profile projection histogram as another feature. The performance of various classifiers has been evaluated using these features experimentally. The multiple classifiers namely Quadratic SVM, k-NN, weighted k-NN, Ensemble Subspace Discriminant and Bagged Trees are used and quadratic SVM has been found to produce better results.

Gauri Katiyar and Shabana Mehfuz has been developed SVM based offline handwritten digit recognition system[3]. Experiments have been performed using well known standard database acquired from CEDAR, and also they proposed four

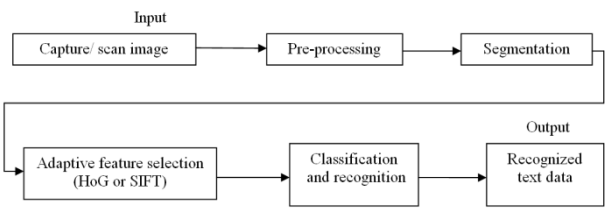
different techniques of feature extraction to construct the final feature vector. The four feature extraction methods they used are box approach, diagonal distance approach, mean and gradient operations. In their proposed experiment they have used SVM library known as LIBSVM running under WEKA tool. WEKA is universally used tool in the field of machine learning. LIBSVM is integrated software generally used for support vector classification and regression. They have used SVM as the classifier with polynomial function as the kernel type and have set the values of other parameters to their default values. The networks are trained and tested on the CEDAR CDROM-1 dataset. They achieved 97.16% accuracy. According to them the superior performance of SVM is because of superior generalisation capability of support vector machine in high dimensional space.

Mohamed Elleuch *et al.*[2], proposed offline arabic handwritten recognition system with dropout applied in deep networks based SVM. This work presented two models for the recognition of the handwritten Arabic characters. Number one is based on deep learning approach, called Deep networks using SVM (DSVM), and dropout technique while number two insists on hand-crafted feature extraction as SIFT descriptor. In this study, a deep learning model based on Support Vector Machine (SVM) named Deep SVM (DSVM) is represented and applied a dropout technique on DSVM. The deep SVM is built by a stack of SVMs permitting to extracting/learning automatically features from the raw images and to realize classification. A Scale Invariant Feature Transform (SIFT) descriptor was utilized for SVM classifier. LIBSVM tool used for SVM implementation. The outcomes of their experiments proved that, DSVM model, which takes raw data as input, produced strong features automatically without extra-feature engineering stages. Even more, improvements were obtained by using dropout technique.

Dewi Nasien *et al.*[6], implemented English handwritten character recognition using SVM. For feature extraction they used Freeman chain code (FCC) as the representation technique of an image character. The boundary of a character image get from chain code representation in which the codes represent the direction of where is the location of the next pixel. An FCC method that uses 8-neighbourhood that starts from direction labelled as 1 to 8 is used. Support vector machine (SVM) classifier is chosen for the classification purpose. The data used in their experiment is NIST dataset of handwritten digits . Results after test show that by applying the proposed model, the system could reached a relatively high accuracy for the problem of English handwritten recognition.

Proposed Approach

Design - Architecture Diagram



Description of the Proposed Approach

Pre-processing

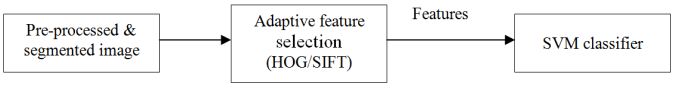
It is an important step in any image processing based application. It enhances the input image and make it suitable for segmentation. Pre-processing involves resizing, noise removal, binarization, thinning, edge detection, skew detection etc to enhance the quality of images and to correct distortion.

Segmentation

Segmentation is an important process in character recognition. Segmentation separates the characters in an input image, make separation between each individual character. Segmentation is difficult mainly because of variability in inter-character distance, skew, slant, size and curved like handwriting[5]. Sometimes components of two consecutive characters may be touched or overlapped and this situation complicates the segmentation task greatly. There are two types of segmentation, external and internal segmentation. External segmentation is the isolation of various writing units, such as paragraphs, sentences or words. Internal Segmentation decompose an image of a sequence of characters into sub images of individual symbols or characters.

Adaptive Feature Selection

The proposed approach mainly focuses on feature extraction. Two feature extraction techniques are considered here, are HoG and SIFT. HoG feature can detect some of the characters like the characters that have sharp edges. So the characters that cannot recognize with HoG, can be recognize by using SIFT feature extraction. The proposed recognition system will itself decide which feature extraction need to apply for an input image based on the few characteristics of that input image such as loops, pointed edges and smooth curves. Both this HoG and SIFT use image gradients, and divide the image into spatial bins and forms a histogram for an image. HoG and SIFT has got better recognition rate as its own. So, combination of these feature extraction technique might yield good results and increase accuracy of the handwritten recognition system. SIFT feature extraction need a lot of processing time compared to all other feature extraction techniques. HoG need very less processing time than SIFT. So, in the proposed method most of the characters can be recognized with HoG and the characters which have smooth edges need SIFT feature extraction. For example characters like 2,5,A,E etc are suitable for HoG, whereas S,C,O,D,8,6 like characters are suitable for SIFT. So, overall the accuracy can be increased, because the proposed system can adaptively choose the right feature extraction (either HoG or SIFT) based on the input image.



Feature Extraction

Based on the previous stage(Adaptive feature selection) of proposed method either HoG or SIFT feature extraction on pre-processed image need to carry out. HoG use the direction of intensity of the gradients and edge directions. HoG feature extraction goes through steps such as, divide the image into small cells, compute histogram for each cell, combine these histograms to compute one single descriptor. HoG works well with geometric and illumination transformations. SIFT feature extraction has following steps such as, key point detection, local descriptor extraction, matching and model verification. SIFT descriptors match local regions using key point locations to detect objects. SIFT works well with changes in orientations.

Classification and Recognition

The classification stage is the decision making part of a recognition system and it uses the features extracted in the previous stage. The classifier using in this proposed work is SVM. Compared to other machine learning algorithms for handwritten character recognition, SVM has got highest recognition rate and accuracy. The features extracted in the feature extraction stage fed as input to the classifier. SVM is implemented using the sklearn library and linear kernel is used. The trained classifier recognize the character if the features extracted form the scanned image matches with features of any classified class.

Implementation

For implementing the proposed work the chars74k handwritten dataset is used, which contains digits, capital and small letters. Each class contains 55 images. Each of these images resized and thinned using opencv library. Opencv is used for pre-processing of images. OpenCV (*Open Source Computer Vision*) is a library of programming functions mainly aimed at real-time computer vision. From the preprocessed image, either HoG or SIFT features are extracted based on the character in the input image. The Histogram of Oriented Gradient (HOG) feature descriptor is popular for object detection. Compute a Histogram of Oriented Gradients (HOG) by global image normalization (optional), computing the gradient image in x and y, computing gradient histograms, normalizing across blocks and flattening into a feature vector[8]. By applying HoG in each of the dataset images, returns a single list descriptor as the image identity of one image. All these single list vector from all the images are combined which produce an array of vectors and this array of vector is then gave to the SVM classifier. The scale-invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. There are mainly four steps involved in SIFT algorithm: Scale-space Extrema Detection, Keypoint Localization, Orientation Assignment, Keypoint Descriptor and Keypoint Matching[9]. Using Opencv SIFT function on each of the images, returns an array of descriptors, ie: key points and feature sets at each key point. After combining the SIFT descriptors, it fed to SVM classifier.

For implementing HoG, skimage library is used. Scikit-image is an open source image processing library for the Python programming language. It includes algorithms for segmentation, geometric transformations, color space manipulation, analysis, filtering, morphology, feature detection, and more [10]. The features extracted are then fed to the linear SVM classifier using sklearn library. Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels[11]. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. The SVM algorithm is implemented in practice using a kernel. The kernel types are liner, polynomial, radial and sigmoid. The proposed work uses the linear SVM of sklearn library. The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The SVM classifier gives better recognition rate compared to other classifiers for handwritten character recognition. The sklern's linear SVm is capable of performing multi-class classification on a dataset. The sklearn's linear SVM take as input two arrays: an array X of size [n_samples, n_features] holding the training samples, and an array y of class labels (strings or integers), size [n_samples]. After being fitted, the model can then be used to predict new values.

HoG features gave better recognition on characters with sharp edges like 4, A, E, 5 etc. So, the characters that cannot be recognized using the HoG feature extraction technique can be recognized using SIFT feature extraction. SIFT is more good for recognition, but it is used less when compared with HoG because SIFT need much more processing time[4]. Thus it will reduce the recognition time of the handwritten recognition system. The combination of these HoG and SIFT would reduce the recognition time as well as increase the recognition rate. We can conclude that the overall accuracy of proposed approach is better than other handwritten recognition approaches.

CONCLUSION AND FUTURE WORK

Handwritten character recognition is still a challenging task in the field of machine learning. There are two important factors involved in order to increase the recognition accuracy are the size of dataset and the feature extraction technique using. Handwritten character recognition has got various applications and these applications are greatly needed in so many domains such as bank cheque processing, signature verification, postal address and zip code recognition, reading commercial forms and writer identification. So, improving handwritten recognition is really beneficial for our future. The proposed work can improve the recognition rate, because the system adaptively chose the right feature extraction technique and train an SVM classifier. SVM gives better results compared to other classifiers. The overall system can give better accuracy. In future the system can be modified to accept input image which contains multi language textual information. Both the off-line and online recognition can be combined into a single system. Based on the captured image (images of paper, white board etc), pre-processing needs also varies, an apt pre-processing also yields good results.

References

1. Madhuri Yadav and Dr. Ravindra Purwar (2017),'Hindi handwritten character recognition using multiple classifiers', IEEE 7th International Conferenceon Cloud Computing, Data Science & Engineering.

2. Mohamed Elleuch, Raouia Mokni and Monji Kherallah (2016), 'Offline Arabic Handwritten Recognition System with Dropout applied in Deep Networks based-SVMs', IEEE International Joint Conference on Neural Networks (IJCNN).
3. Gauri Katiyar and Shabana Mehfuz (2015), 'SVM based offline handwritten digit recognition', IEEE INDICON.
4. Gooitzen van der Wal, David Zhang, Indu Kandaswamy, James Marakowitz, Kevin Kaighn, Joe Zhang, Sek Chai(2015),'FPGA Acceleration for Feature Based Processing Applications', IEEE.
5. Shivangkumar R Patel and Ms. Jasmine Jha (2015), 'Handwritten character recognition using machine learning approach', International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO).
6. Dewi Nasien, Habibollah Haron and Siti Sophiayati Yuhaniz (2010), 'support vector machine (svm) for english handwritten character recognition', IEEE Second International Conference on Computer Engineering and Applications.
7. Handwritten character recognition. Online: http : "// www. cvisiontech. com/ /library/ocr/ocr-pdf/handwriting-character-recognition.html". Accessed at August, 2017.
8. Histogram of oriented gradients. Online: 'http://scikit-image.org/docs/dev/auto_examples/features_detection/plot_hog.html". Accessed at August, 2017.
9. Introduction to SIFT (scale invariant feature transform): "http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_sift_intro/py_sift_intro.html". Accessed at September,2017.
10. Scikit-image: "https://en.wikipedia.org/wiki/Scikit-image". Accessed at September, 2017.
11. Support vector machines for machine learning: " https://machinelearningmastery.com/support-vector-machines-for-machine-learning/". Accessed at September, 2017.

How to cite this article:

Krishnendu K R and Nikhila T Bhuvan (2018) 'An Adaptive Feature Extraction Approach for Improving Handwritten Character Recognition', *International Journal of Current Advanced Research,* 07(5), pp. 12518-12521.
DOI: http://dx.doi.org/10.24327/ijcar.2018.12521.2203

****