Author Identification of Handwritten Text

Prof. Deepa Bendigeri [1]
Dept of ISE
SDMCET
Deepabendigeri12@gmail.com

Adarsh A Shanbhag [2]
Dept of ISE
SDMCET
adarshshanbhag31@gmail.com

Rakshit S Handral [3]
Dept of ISE
SDMCET
rakkiran22@gmail.com

Akshay K [4]
Dept of ISE
SDMCET
akshaykagatur@gmail.com

Manushree N Bhat [5]
Dept of ISE
SDMCET
manushreenb6@gmail.com

Abstract— **Computer and phones may be more ubiquitous than ever, but many people still prefer the traditional feeling of writing with ink on paper. Ultimately, this method set out well for decades in human history. Despite the availability of various technological writing tools, many people still choose to take their notes traditionally: with pen and paper. However, there are certain pitfalls in traditional way of handwritten text. It is tedious job to store and retrieve the documents in an organized manner, search through them efficiently and to share them with others. Thus, handwritten Identification is the propensity to interpret scanned handwritten document as input from external sources such as notes, scripts, pdfs etc into digital form. A handwriting recognition system handles pre-processing, performs accurate segmentation into characters images, and finds the most feasible words. Hence, translating the handwritten characters to the digital format also saying the author who wrote the text. With time the text on the paper will fade away but a file stored on a computer will be lost only if it is deleted. Storing any handwritten document in a digital format has gained prime importance as a way of preserving history.**

Index Terms – CNN, IAM Dataset, Feature Vector, Binarization.

INTRODUCTION

Identifying an author is highly essential in many areas like forensic expert decision-making systems, network security, biometric authentication in information etc. In forensic science author identification is used to authenticate documents such as records, signatures and also in criminal justice. Author identification of handwritten text can be generally classified into two types as online mode and offline mode. In online mode, sequence of signals using a transducer device are used for identifying author, but when it comes to offline mode handwritten text is used to identify scanned images.

A published author usually has a unique writing style in any of the works. If there exists a few authors then manually a human can identify the style and predict the author. The process of identifying the author from a group of candidates according to his writing samples is author identification. Author Identification plays an important role in day today life. It also helps to recommend new author of similar handwriting style. The main and important requirement is huge Dataset for training and testing.

1. LITERATURE SURVEY

I Khandokar, Md M Hasan, F Ernawan, Md S Islam, M N Kabir, "Handwritten character recognition using convolutional neural network", Journal of Physics: Conference Series, doi:10.1088/1742-6596/1918/4/042152

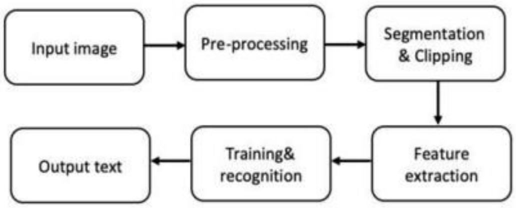
Author identification can be defined as recognising or identifying the author of his or her respective handwritten script. Author identification can be done in two modes online and offline. Online mode, where author is identified during, he or she is writing on piece of paper, the application recognises the handwritten text through pen movement and written text. Offline mode, where the written text document is scanned and author is identified. The aim of the project was to develop an efficient HCR application using CNN in offline mode. The CNN algorithm was implemented in MATLAB R2015a under Windows 7 operating system. The implemented program was run on intel core i7-2640 CPU with 4GB RAM. To train and test the model NIST dataset was used. The handwritten text was scanned copy in form of image. Training was carried out with various images, then testing was conducted to find the accuracy of the CNN algorithm. To examine the performance of CNN algorithm, they experimented with the NIST dataset. After several train and test of the model they found the accuracy of handwritten characters through NIST dataset.

PROPOSED METHODOLOGY

The proposed model shown in Figure has four stages for the classification and detection purpose: Pre-processing, Segmentation & Clipping, Feature extraction and Classification & Recognition.

Pre-Processing: in this stage input taken is image to perform enhancement of image by removing noise. Image may require in form of greyscale or binary form, which are achieved in this stage.

Segmentation: Once the input images are pre-processed, the further step is to separate individual characters using a segmentation method. The characters derived are then stored into a sequence of images. If extreme boarders are detected then the boarders in each character image are removed. further, individual characters are elevated to some specific size.



Feature Extraction: it's operated on the segmented characters from the previous step. the features from segmented characters are extracted using CNN with ReLU activation function. CNN is traced on each character image to form a matrix of reduced size using convolution and pooling. Finally, the reduced matrix is compacted to a vector form using the ReLU function. The obtained vector is known as feature vector.

Classification and Recognition: The obtained feature vector is inputted to formulate corresponding class. During the training phase, the parameters, biases, and weights are calculated. The calculated parameters, biases, and weights are used in the testing phase for classification and recognition purposes.

RESULTS AND CONCLUSION

After implementing, training the model, we can observe that the average accuracy increases with higher number of training images, parallelly more accurate information on the training parameters also, which automatically improves the accuracy in classification during testing phase. The accuracy obtained from 200 training images as 65.32% is improved gradually with increasing training images. The accuracy reaches to 92.91% with the 1000 training images. Thus, further increment of training images will continue to enhance the accuracy towards to certain limit – which cannot be exceeded due to numerical errors, and the constraints on the CNN capability of image differentiability for labels.

2. LITERATURE SURVEY

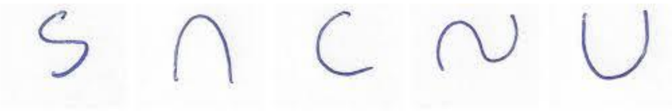
Marco Mora, José Naranjo-Torres, Verónica Aubin, "Convolutional Neural Networks for Off-Line Writer Identification Based on Simple Graphemes", Appl. Sci.,

Volume 10, Issue 22 (November-2 2020). RESULTS AND DISCUSSION

There are different biometric features that can be used for verification or identification of people, one of them is writing. Everyone has unique handwriting which cannot be copied, the rhythm of writing captures certain graphic characteristics in the text which is referred for the identification of author. Here firstly the written documents must be stored in a database and the authors of the documents must be known in advance for identification.

PROPOSED METHODOLOGY

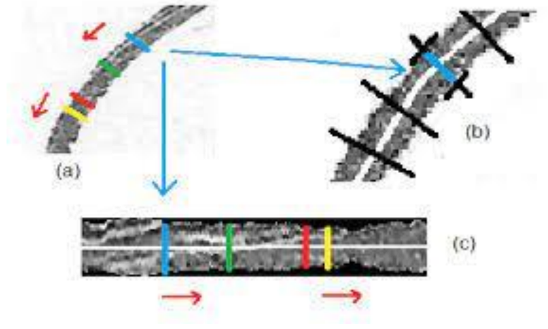
A database which contains 5 simple grapheme types (“S”, “∩”, “C”, “~”, “U”) is used. They have also introduced a new descriptor to represent the texture of the handwritten strokes and verification tests are performed with a SVM based classifier. The repository used contains the above 5 simple graphemes for 50 writers, with 100 samples of each simple grapheme per writer.



(a)Grapheme “S” (b)Grapheme “∩” (c)Grapheme “C” (d)Grapheme“~” (e)Grapheme “U”

As the original and rectified grapheme are greatly different from the images included in Imagenet database, they cannot be directly classified using the pretrained CNNs. So, a learning transfer process takes place, in this process proper adjustments and training of previously trained CNN with new images takes place. They have used AlexNet, ResNet and VGG (versions VGG-16 and VGG-19) CNN models. They have conducted several experiments with simple graphemes and the pretrained CNNs while performing learning transfer modality. They have used two variants of the grapheme images for this experiment. First one consists of the most commonly known approach namely rectified grapheme. The second one uses RGB image of the original grapheme. All the images used in this composition make up the LITRP- SGDB database.

For rectification of graphemes, they firstly carry out Segment of Original Simple Grapheme (Fig a) then Construction of Rectified Image (Fig b) and lastly Resulting Rectified Image is obtained (Fig c).



RESULT AND CONCLUSION

In this paper a way for processing simple graphemes for writer identification is defined. This approach is based on the use of CNN. In this experimentation AlextNet, VGG-16, VGG-19 and ResNet-18 models are used as they have a reasonable compromise between accuracy and training time. In this experiment the best result was obtained while using original grapheme image and ResNet-18 CNN model.

3. LITERATURE SURVEY

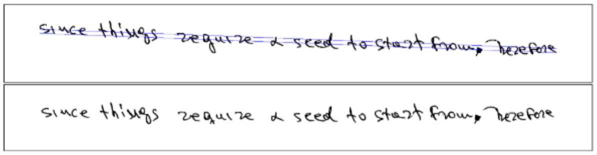
Stefan Fiel, Robert Sablatnig, “Writer Identification and Retrieval using a Convolutional Neural Network”, Conference: International Conference on Computer Analysis of Images and Patterns, DOI:10.1007/978-3-319-23117-4_3 (September 2015)

Author identification is the task of identifying an author of a handwritten document by comparing the writing with the ones that are stored in a database. The authors of the documents must be known in advance for identification.

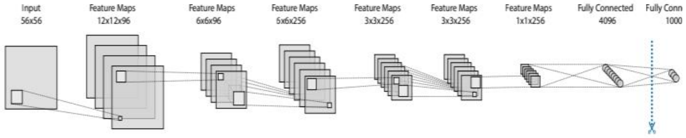
PROPOSED METHODOLOGY

In this paper they use CNN for writer identification and writer retrieval. Here the output of the second last fully connected layer is used as a feature vector is needed for every document image to have a comparison with precalculated features in the dataset to identify the writer. The preprocessing of the document image is necessary as CNN requires an image as input with fixed size. The preprocessing includes binarization, text line segmentation and sliding windows. Generation of the feature vector using the CNN is the next step. The obtained vectors are then used for writer identification and writer retrieval using the nearest neighbour approach.

Preprocessing:


since things require a seed to start from, therefore
since things require a seed to start from, therefore

Generation of the Feature Vector:



RESULT AND CONCLUSION

This method uses CNN for generating a feature vector which is then used to compare using the χ 2 -distance. The proposed method has been evaluated on three different datasets namely ICDAR 2011 and 2013 writer identification contests and the CVL dataset. This experiment shows that the proposed method shows slightly better results on two of three datasets but worse on the remaining dataset which is originated mainly from the preprocessing steps.

4. LITERATURE SURVEY

Text-independent Writer Identification via CNN Features and Joint Bayesian.Tang, Y., & Wu, X. (2016). Text-Independent Writer Identification via CNN Features and Joint Bayesian. 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR).

The proposed system aims to spot the author of a text from a variety of known writers by using their handwriting images. Convolutional neural network (CNN) has powerful ability to learn deep features and has numerous applications in computer vision domain. CNN is employed to perform the feature extraction out of handwriting images for author identification. To extract global and discriminative features from the complete handwriting image, an oversized number of handwriting images for every writer are the first obtained for CNN model training. So, a data augmentation technique is proposed to get the handwriting images, & then a CNN model is trained for global feature extraction. Finally, the joint Bayesian is employed for the author identification.

PROPOSED METHODOLOGY

The process of finding the author of the given handwriting starts by extracting the features of the image which contains handwriting of an author. Below are the subsequent steps that were dispensed to attain the result:

a) Data Augmentation

To train a CNN model a very large set of data is very essential. However, the number of handwriting is very small for each writer in the current dataset for writer identification. As an example, 4 samples for every writer in ICDAR2013 dataset and 5 samples for every writer in CVL dataset. That’s why we perform data augmentation to unravel this problem. For the given handwriting of a specific writer, we obtain four different images by using the proposed data augmentation technique.

b) CNN based feature extraction

After doing data augmentation, a large number of images are generated from the training dataset, supported which, the subsequent step is carried out. Which is to extract some discriminative features from the images to represent their properties. The CNN framework consists of four convolutional layers. Out of which two are fully connected layers, one is output layer, and one is a softmax layer.Each convolutional layer is followed one local response normalization (LRN) normalization layer, one rectified linear units (ReLU) nonlinear activation layer and one max pooling layer. The output layer is a fully connected layer and the number of its nodes is decided by the number of classes of training dataset. The last layer is softmax layer. Here ICDAR2013 training dataset is employed which consists of 100 writers and every writer with 4 handwriting images to train the CNN model for feature extraction. Therefore, the output layer and softmax layer both have 100 nodes.

III. Author Identification
After CNN based feature extraction, next step is to identify the author. The joint Bayesian technique which has been successfully applied in face verification is employed for writer identification based on the extracted features. In joint Bayesian, an extracted CNN based feature is represented by the sum of two independent Gaussian variables as x=μ+€.For two feature vectors x1 and x2 if they are of same person, r(x1,x2) are larger. Since for every handwriting image, 20 constructed images are generated for writer identification during test, voting result of 20 feature vectors is employed to make the ultimate decision.

5. LITERATURE SURVEY

Parveen Kumar ,Ambalika Sharma "Segmentation-free writer identification based on Convolutional Neural Network",July 2020 via sciencedirect.com on Elsevier journal.

Writer identification is employed to identify the writer of a given handwritten text sample.Over the previous couple of decades, the applications of writer identification have widely varied from ancient document analysis to the modern forensic document analysis and the list is continuously growing. The offline data acquisition process saves data as physical documents. The image enhancement techniques are required to enhance the visual information of the document. The text document images are a group of text-blocks,words, text-lines, and backgrounds. A segmentation method is required to divide the document image into its characteristic regions. The segmentation of image text documents could be of text, character, word, a line, or an entire block of text. The pre-processing of the documents plays a significant role in the performance of the WIV system.

PROPOSED METHODOLOGY

A segmentation free model is proposed to identify the writer which utilizes the offline images. The features extracted by the proposed SEG-WI model are often used for distance-based verification system. The key features of the proposed work is as follows:

- A Segmentation free Writer Identification (SEG-WI) model based on a convolutional neural network (CNN) is proposed to identify the writer.
- Region selection mechanism is also developed to enhance the overall performance of the model.
- The similarity between two documents of various writers makes the training difficult, therefore, a brand new training strategy is recommended to train the model.

The advent of deep neural networks (DNNs) extends the possibilities in machine learning applications. The performance of CNN in vision-based problems is astonishing and provides opportunities for further enhancements. In this research, the powerful abilities of deep CNN is utilized for writer identification of handwritten text images. The pre-processing steps to any pattern recognition system are critical as the error introduced in these steps severely affects the result. The performance of the proposed model is evaluated on various datasets from different languages including kannada,devangiri script etc

CONCLUSION

A segmentation free deep convolution neural network (DCNN) for offline text independent writer identification is presented in this research. A brand new method which does not has segmentation and pre-processing is the highlight of this research,so as to enhance the performance of proposed model,training scheme and region selection were introduced. A comparative evaluation of the SEG-WI model is performed for various different datasets of various languages.

FUTURE SCOPE

The deployed projects shows the identified author,accuracy and graph as output.However,there is numerous feautures that can be used to improve the project.Some are as follows:

- In the current implementation of the project only English is supported.There is a scope to support multiple languages like kannada,Sanskrit,hindi,Arabic etc.

- The dataset used in this project has only limited number of authors and their handwritten images.This can be scaled to a larger number.
- A graphical user interface(GUI) can be added to interact in a better way with the project and its different components.
- An android application can be developed.
- There lies a scope for increasing the accuracy of the output.

CONCLUSION

Automatic writer identification system helps in determining and identifying whether the given handwriting is truly matched and assigned to the claimed writer of handwriting. The proposed system helps in various areas to identify the author of a handwritten text. Here the handwritten text is converted to digitized format and identification of the author is done. As this can be used in day to day life for various purposes, we aim at reducing the efforts and time of the users.

REFERENCES

[1] I Khandokar, Md M Hasan, F Ernawan, Md S Islam, M N Kabir, “Handwritten character recognition using convolutional neural network”, Journal of Physics: Conference Series, doi:10.1088/1742-6596/1918/4/042152

[2] Marco Mora, José Naranjo-Torres, Verónica Aubin, “Convolutional Neural Networks for Off-Line Writer Identification Based on Simple Graphemes”, Appl. Sci., Volume 10, Issue 22 (November-2 2020). RESULTS AND DISCUSSIONS.

[3] Stefan Fiel, Robert Sablatnig, “Writer Identification and Retrieval using a Convolutional Neural Network”, Conference: International Conference on Computer Analysis of Images and Patterns, DOI:10.1007/978-3-319-23117-4_3 (September 2015).

[4] A. Arora, A. Kaul and V. Mittal, "Mood Based Music Player," 2019 International Conference on Signal Processing and Communication (ICSC), 2019, pp. 333-337, doi: 10.1109/ICSC45622.2019.8938384.

[5] K. Chankuptarat, R. Sriwatanaworachai and S. Chotipant, "Emotion-Based Music Player," 2019 5th International Conference on Engineering, Applied Sciences and Technology

[6] (ICEAST), 2019, pp. 1-4, doi: 10.1109/ICEAST.2019.8802550.