Image Classification and Text Extraction using Machine Learning

R. Deepa, Associate Professor
Department of Information Technology
Loyola-ICAM College of Engineering and Technology
Chennai, India

Kiran N Lalwani, Student
Department of Information Technology
Loyola-ICAM College of Engineering and Technology
Chennai, India
kiran98780@gmail.com

Abstract— **Machine Learning is a branch of Artificial Intelligence in which a system is capable of learning by itself without explicit programming or human assistance based on its prior knowledge and experience. It is used to predict or make decisions to perform certain task based on the training set that is provided. In the proposed system, image classification is implemented using Convolutional Neural Network (CNN). The text is then extracted from the classified image using Tesseract, which has implemented a Long Short-Term Memory (LSTM) based recognition engine. The LSTM networks are the units of Recurrent Neural Network. The CNN performs better on very large datasets, by overcoming the problem of overfitting. Also, single line text extraction is replaced by multiple line text extraction. Thus, the accuracy of this system can be improved by incorporating a large dataset and increasing the number of epochs. In addition, a trial-and-error methodology is used to determine the number of convolution and pooling layers with the number of nodes in each layer. Finally, CNNs use relatively few preprocessing compared to other image classification algorithms.**

Keywords— text extraction; image classification; Tesseract OCR; Convolution Neural Networks

I. INTRODUCTION

Machine learning, a subdivision of Artificial Intelligence comprises of algorithms and statistical framework which helps the system to learn by itself and make predictions about certain functions [1]. Image classification and text extraction are some of the applications of machine learning. Image classification is the process of feature extraction and pattern recognition from the images and classifying them. Whereas, text extraction is the process of generating text from electronic documents.

Image classification can be implemented using various supervised techniques such as Naive Bayes [2], K-Nearest Neighbor (KNN) [3], Support vector machines (SVM) [4], Decision trees [5], Random forests [6] and Convolutional Neural Network (CNN) [7]. These techniques process and classify images into various classes.

One of the classification techniques called Naïve Bayes, is based on the Bayes theorem of conditional probability. In this technique, the presence of a particular feature in a class does not linked to any other features. The major disadvantage of Naive Bayes is, it considers that the features are independent of each other. However, in the real world, features depend on each other. The naïve-Bayes assumption needs to be relaxed, which is too "naive" for overall image recognition. [2].

In K-Nearest Neighbor, the Euclidean distance is used to find the closest class. In this technique, the target class is predicted using the neighbor point's information. Then in training procedure, feature vectors are stored and the training images are labelled. In the classification process, then the unlabeled point is basically allotted to the label of its k-nearest neighbors [3]. The major disadvantage of KNN is that, it is a slow algorithm, which has curse of dimensionality and outlier sensitivity, thus it needs the homogeneous features and an optimal number of neighbors. In this method, imbalanced data cause problems and has no capability in dealing with missing value problem.

In Support vector machines [4], the entire plane is divided into distinct classes on the hyperplane by plotting the dataset on a graph. This is because, the distance between the separating plane and the adjacent training data points of any class is large and thus the generalization error of the overall classifier is descended. These points are predicted by plotting them into the same space. SVM can be made to handle the multiple classification tasks in scene classification. It takes a long time for training large datasets. The small calibrations cannot be made to the model and it is difficult to include the business logic. It is also hard to understand and infer the final model, variable weights and individual impact.

Decision Trees are a flowchart like tree structure, where each internal node represents a test on an attribute, each branch represents a test outcome, and each leaf node represents a class label [5]. The uppermost node in a tree is the root node. Some of the drawbacks of decision trees are: performance doesn't generally contend with the most supervised learning techniques; it can easily overfit the training data as shown in Figure 1, thus requires tuning; slight differences in the data can result in an entirely different tree; recursive binary splitting creates "locally optimal" decisions that may not cause a globally optimal tree; doesn't tend to work well with highly unstable classes and very small datasets.

Random forest is a machine learning algorithm used for classification and regression. The algorithm creates multiple decision trees and finally merges them all together to get a more accurate prediction [6]. The major disadvantage of the

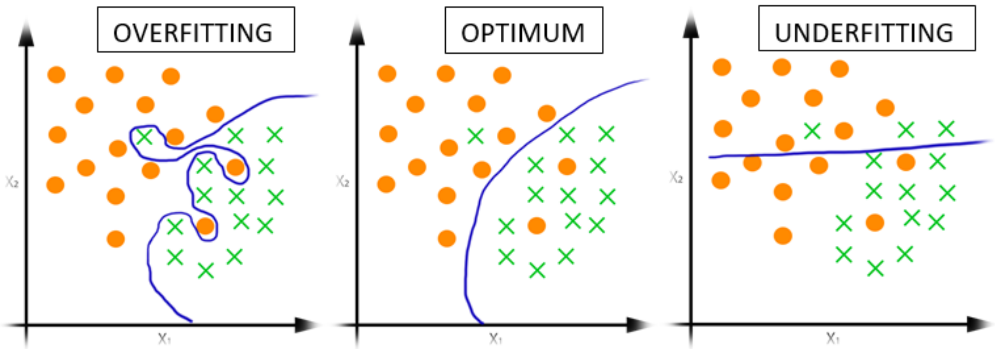

Fig. 1. Comparison of overfitting, underfitting and optimum condition
(Source: https://machinelearningmedium.com/2017/09/08/overfitting-and-regularization/)

random forest algorithm is that, since a greater number of decision trees are used to increase accuracy, it slows down the speed of prediction, thus, being ineffective in making predictions compared to other algorithms.

These problems can be reduced using convolutional neural network. CNN is a machine learning technique for image classification. It makes use of deep learning concepts that often uses machine vision for various functions. In CNN, the data is divided into training, test and validation sets. It reduces the problem of underfitting and overfitting, the features must be increased, as it expands the hypothesis space. The multiple lines of text are then extracted from the classified image using Tesseract-OCR [8]. This is implemented using Tesseract OCR package which contains an OCR engine - libtesseract and a command line program - tesseract. Then, the extracted text is filtered and stored in the database.

Section II explains existing techniques in image classification and text extraction. Section III explains the proposed techniques and Section IV discusses performance analysis. Section V concludes with future work.

II. EXISTING WORK

Image classification is done by assigning pixels in the image into different categories or classes of interest and it is a widespread research area in the field of Deep learning, Pattern recognition, Human Computer Interaction. In [2], a pairwise local observation based Naive Bayes (NBPLO) classifier is proposed for image classification. First, the salient regions and the Key points are found as the local observations. Second, the discriminative pairwise local observations are described using Bag-of-features histogram. Third, the object class models are trained by using random forest to develop the NBPLO classifier for image classification. The memory space problem prevents from doing experimentations on huge database.

In [13], the KNN classification model is used to detect and recognize varieties of Indian food representing an automatic food detection system. A combined color and shape feature are used. The KNN classification model is used to classify the feature. The food images are taken using the high-resolution portable cameras attached to a wearable glass, cap or hat. Eventually, this system can also support the vision impairment in recognizing the food on the platter with its features like color, texture and shape.

The real-world scenes are classified into four semantic groups, namely coast, forest, highways and street using the Support Vector Machines (SVMs) [6]. First, the features are extracted and scene images are normalized. Then, they are fed into the SVM with various kinds of kernel functions. The various kernels are linear, polynomial and Radial Basis Function. Any error in the classification, the accuracy and the performance of these kernels are recorded. Finally, the feature dimension reduction methods are considered.

The work in [11] proposes a random forest-based face image classification method. The random forest, an ensemble learning method, grows many classification trees, where each tree provides a classification. The classification that has the maximum votes is selected from the forest. It performs three experiments. The several existing approaches are trained and evaluated with the random forest-based method.

The work proposed in [12] uses logistic regression and neural network methods for a face recognition technique based on binary images. These methods convert a color image to gray image and then denoised using a low pass filter. Then, the local intensity variations around eyebrows, eyelids, nose and mouth are captured by applying local window standard deviation to the denoised image. Then, the adaptive thresholding is used to binarize the image to get a good quality binary image. Then, the image size is normalized and reduced to 50%, 30%, 20% and 10% of its original size using nearest neighbor interpolation method. For each reduced size image, a corresponding face database is created. In this technique, computational space and training time are minimized.

Due to increase in number of documents, the requirements of information, identification, indexing and retrieval, text extraction is implemented by many researchers. Due to the variation of size, font, style, orientation, alignment, contrast, complex colored, textured background in the document, the text characters are tough to be identified and recognized. Hence, several techniques have been developed for extracting the text from images. Text Extraction plays a key role to find out essential and esteemed information.

Kouzani et al. describe about superpixel image classification based on the CNN concept [11]. It is known that high resolution pictures are hard to process. But they give good accuracy. High resolution images can be classified in two ways. One of the concepts of feature extraction and the other by sharpening the image and classifying them. In both ways, it uses CNN for classification. In this case, low resolution is difficult to sharpen and classify.

In the Gamma Correction method [13] non-text background details from the image are suppressed by applying appropriate gamma value to remove non-text regions. The gamma values are estimated using the texture measure without any prior details of the imaging device. The estimated gamma value is then applied to an input image to achieve the background suppressed image. A range of gamma values from 0.1 to 10, with an interval of 0.1 is used to find a proper gamma value, applied to the image resulting in 100 images. After converting into a gray image, the textural features are extracted using Gray Level co-occurrence matrix.

In [14], a method to localize the text data in both image and video files is proposed. It is easy to recognize and extract text from images, but difficult to do so in case of a playing video. Once the system locates the text files on the multimedia file, it is easier to extract them. The drawback of this system is that it takes a very long time in processing long videos.

In [15], TravelMate, the real time application is designed and developed, where text extraction is done using the stroke width transform and connected component-based approach. In this application the tourists are assisted, when they roam in foreign countries. The extraction rate is used to test the performance of this system. This application extracts almost all texts in horizontal orientation appropriately. The lighting condition and camera resolution vary with the performance of real time images.

The work in [16], investigates critical features such as natural scene text identification and extraction due to cluttered background, unstructured scenes, orientations, ambiguities and much more. The LUV channel on an input image for image enhancement is applied to an input image to get perfect stable regions. Then, the standard segmentation technique MSER is used to select L-Channel for region segmentation. The various geometrical properties are also considered to differentiate between text/non-text regions.

III. PROPOSED WORK

In the devised model as shown in Fig.2, the idea proposed is to take in a number of images of documents like identity proofs of individuals and classify them into classes, such as passport and license. Once the images are classified, they are subjected to the text extraction module. The text data are extracted from the classified images. The extracted credentials from the images are then stored in the database.

A. Image Classification

The image classification is done using the Convolutional Neural Network as shown in Figure 2. In CNN, pre-processing comprises the convolutional layer, pooling layer and flattening. The convolutional layer maps the image to a matrix format, the pooling layer reduces the dimensions for faster processing time and the flattening converts the image into a linear array format for feeding it into the neural network. Then, the features are extracted from the image. Then, the system is supplied with a training set to learn from it. Finally, the extracted features of the image are compared to that of the training set images and is classified into various layers that goes in sequence such as the input, fully connected, hidden

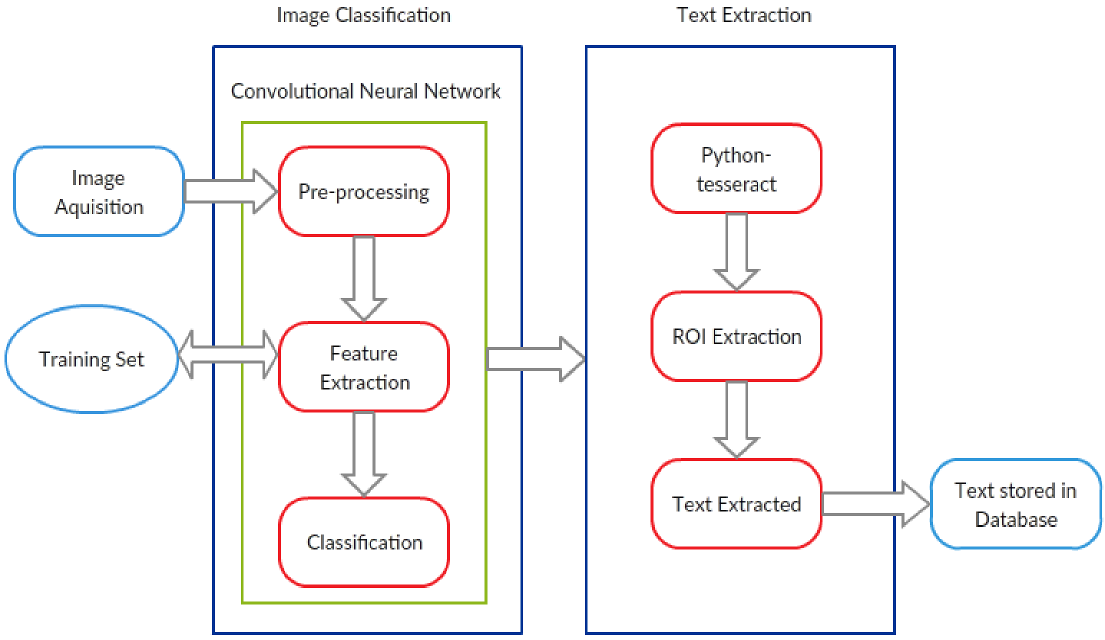

Fig.2. Process of Image Classification and Text Extraction

and output based on the match percentage. The use of CNN is, to overcome the problem of overfitting as in the case of Support Vector Machine.

B. Text Extraction

Text extraction is implemented using Tesseract OCR package which contains an optical character recognition (OCR) engine - libtesseract and a command line program – Tesseract. Tesseract includes a new neural net called Long Short-Term Memory (LSTM) based OCR engine, which focuses on line recognition and also recognizes character pattern. The LSTM network is the units of Recurrent Neural Networks. The Python-Tesseract is an optical character recognition (OCR) tool in python used for text extraction. This tool recognizes and "read" the text embedded in images. This tool is a wrapper for Google's Tesseract-OCR Engine, where it reads all image types including jpeg, png, gif, bmp, tiff, and others. These images are supported by the Python Imaging Library. Once the text data from the image files is extracted using Python-Tesseract, the required credentials are then stored in the database for easy access.

IV. PERFORMANCE ANALYSIS

The image classification of the proposed system using CNN is compared with that of the existing system using SVM [4], as shown in Fig. 3. The accuracy of the CNN is greater than that of SVM because, it overcomes the problem of overfitting. Overfitting occurs, when the system is trained with large dataset and learns from the noise. The CNN resolves the overfitting problem by using various datasets such as training set, test set and validation set to train the system. Therefore, the accuracy in prediction is greater in CNN compared to SVM.

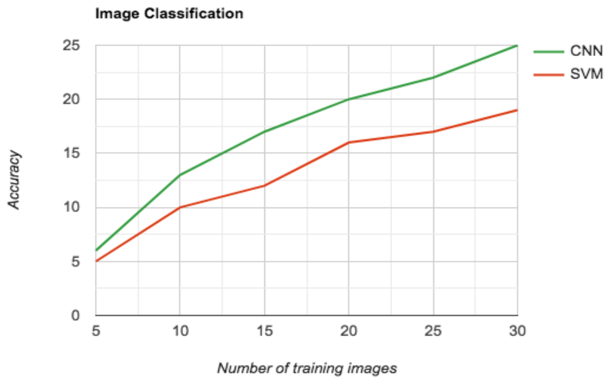

Fig. 3. Accuracy of CNN and SVM

Similarly, the text extraction of the proposed system using tesseract OCR is compared with that of the existing system sing OCRopus OCR, as shown in Fig. 4. The proposed system uses a pytesseract package of Tesseract OCR. Generally, in case of text extracted from image files using any OCR, the accuracy of extraction decreases with increasing number of letters per word. But, when compared to other OCRs, Tesseract OCR exhibits a better performance. Thus, the efficiency and performance of the proposed system are enhanced.

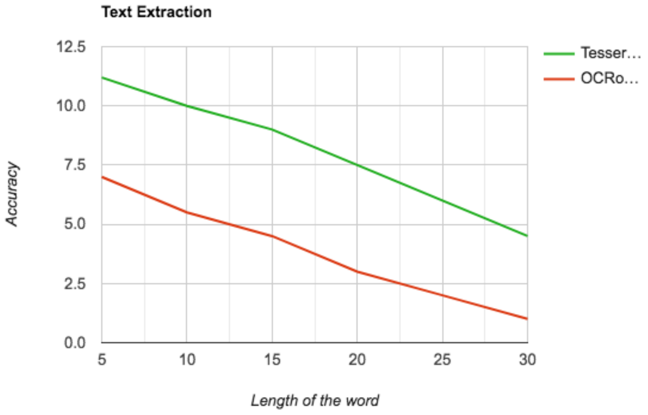

Fig. 4. Accuracy of tesseract OCR and OCRopus OCR

V. CONCLUSION

The proposed system classifies the image and then extracts the text data from them. Image Classification is done using the Convolutional Neural Network, which classifies images based on feature extraction. Text extraction is done using Python-Tesseract OCR package. The extracted text is then stored in the database. The proposed system gives better performance compared to the existing systems, based on accuracy, since CNN overcomes the problem of overfitting. The proposed system can be further improved by developing it into a full stack application, where an user interface is designed for the user to upload the image files into the system.

REFERENCES

[1] Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (March 2002), 1-47. DOI=http://dx.doi.org/10.1145/505282.505283

[2] Shih-chung Hsu1, I-chieh Chen And Chung-lin Huang (2017) "Image Classification Using Naive Bayes Classifier with Pairwise Local Observations" - Journal of information science and engineering.

[3] Aistė Štulienė, Agnė Paulauskaitė-Tarasevičienė (2017) "Research on human activity recognition based on image classification methods".

[4] Chunhua Qian, Hequn Qiang and Shengrong Gong (2015) "An Image Classification Algorithm based on SVM" in Applied Mechanics and Materials Trans Tech Publications, Switzerland Vols. 738-739 pp 542-545.

[5] Bhaskar N. Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtari (2012) "Efficient Classification of Data Using Decision Tree" - Bonfring International Journal of Data Mining, Vol. 2, No. 1.

[6] Ned Horning "Random Forests: An algorithm for image classification and generation of continuous fields data sets".

[7] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2014, pp. 806-813

[8] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 629-633. doi: 10.1109/ICDAR.2007.4376991

[9] Pathanjali C, Vimuktha E Salis, Jalaja G, Latha A (2018) "A Comparative Study of Indian Food Image Classification Using K-

Nearest-Neighbour and Support-Vector-Machines" - International Journal of Engineering & Technology, 7 (3.12) 521-525.

[10] Venkata Naresh Mandhalai, V.Sujatha, B.Renuka Devi (2014) "Scene Classification Using Support Vector Machines" - IEEE International Conference on Advanced Communication Control and Computing Technologies (lCACCCT).

[11] Kouzani, A.Z., Nahavandi, S., Khoshmanesh: Face classification by a random forest. In: K. TENCON 2007 - 2007 IEEE Region 10 Conference. Deakin Univ., Geelong (October 30-November 2, 2007)

[12] Debachandra Singh. N (2017) "Binary Face Image Recognition using Logistic Regression and Neural Network" - International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).

[13] Gayathri Devi. G, Dr.C.P.Sumathi (2014) "Text Extraction from Images using Gamma Correction Method and different Text Extraction Methods – A Comparative Analysis" - International Conference on Information Communication and Embedded Systems (ICICES)

[14] Anubhav Kumar, Neeta Awasthi (2013) "An efficient algorithm for text localization and extraction in complex video text images" - 2nd International Conference on Information Management in the Knowledge Economy.

[15] Pooja Chavre, Dr. Archana Ghotkar (2016) "Scene Text Extraction using Stroke Width Transform for Tourist Translator on Android Platform" - International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) International Institute of Information Technology (IIT), Pune.

[16] Ghulam Jillani Ansari, Jamal Hussain Shah, Mussarat Yasmin, Muhammad Sharif, Steven Lawrence Fernandes (2018) "A novel machine learning approach for scene text extraction" - Future Generation Computer Systems 87 328-340.