Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Document Image Classification using Visual and Textual Features

Sammed S. Admuthe
Research Scholar, Dept. of Computer Engineering
Pune Institute of Computer Technology
Pune, India

Hemlata P. Channe
Professor, Dept. of Computer Engineering
Pune Institute of Computer Technology
Pune, India

***Abstract*: A large number of photographs, signatures, and documents are produced, processed, and stored in the form of digital images. Pan Card, Aadhar Card, Passport, Voter Id are the most authentic documents. Classification of these documents is an important step in office automation, digital libraries, and other document image analysis applications. Document classification generally focuses on extracting textual data and using that for feature engineering. Many document image classification methods exists but they are generally used for photographic image classification. In the present work to classify document images, two different techniques are used. The first one is based on textual feature extraction methods TF-IDF (Term Frequency Inverse Term Frequency). The second is visual classification methods by convolution neural network. The dataset contains 600 documented images belonging to 4 different categories. Classification of textual feature extraction methods has given an average accuracy of 85.5% while visual feature classification has given an average accuracy 93%.**

Keywords: Classification algorithm; Term frequency; Inverse term frequency; Deep learning.

I. INTRODUCTION

Document image classification is a vital step in document retrieval task in document processing. It is the key element in a wide range of contexts such as automated archiving of documents, digital library construction, and other general-purpose document image analysis applications. Passport, Aadhar Card, Voter Id, Pan Card are the major authenticated documents required in many important applications. Each document is having substantially different visual and textual layouts from one another and thus can be accurately classified by comparing their visual and textual characteristics. The whole process of document image classification aims to recognize the text and graphical components in images and to extract the intended information as a human would. In many documents, textual features are important while other visual features can convey more information. For few documents both are important. In this paper, classification models are developed by using two feature extraction techniques which are textual features extraction and visual features extraction.

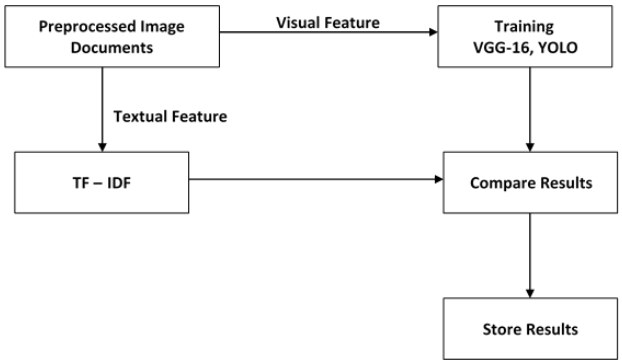

Fig. 1. System workflow for Document Image Classification

In the first Classification model, textual features are extracted from document images by using OCR (Optical Character Recognition). These textual contents are used to develop a classification model. The most commonly used text classification techniques are the TF-IDF technique, text rank, and rule-based classification [1-2]. Literature review shows TF-IDF gives good results so in this paper TF-IDF technique is implemented. The second classification model is based on visual features. Convolution neural network (CNN) techniques are used to capture the most important visual features which are then used for classification. Two CNN nets - VGG-16 and YOLO (You Only Look Once) are used [4-5]. Results of TF-IDF, VGG-16, and YOLO are compared.

II. RELATED WORK

Classification of Documents is an essential task for authentication in many of the areas. Researchers have developed different models based on text and images. For text-based documents, OCR with TF-IDF is used. These methods are used for a large number of text documents [1-2]. Different machine learning algorithms have been implemented for documents containing text and images. Le Kang, Jayant Kumar et.al.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

[3] presented a learning-based approach for computing structural similarities among document images for unsupervised exploration in large document collections Structural similarity is computed using a random forest classifier trained with these histogram features. Lucia Noce et al. [4] proposed an innovative document image classification approach based on a combination of visual and textual data. To extract and alter relevant text concepts from document images, OCR and NLP algorithms are employed. Alessandro Zamberletti et.al. [5] combined OCR and NLP algorithms and are used to extract and manipulate relevant text concepts from document images. They significantly outperformed the state of the art at that time, with their approach. Fusheng [6] uses Deep Learning for the classification of text in legal documents and compared the result to those obtained via Logistic Regression and Support Vector Machines. CNN outperformed Logistic Regression and Support Vector Machine for larger volumes of the training dataset. Kamran Kowsari [7] implements a stack of deep learning architectures to provide understanding at each level of the document hierarchy. The combination of RNN at the higher level and DNN or CNN at the lower level produced consistently high accuracies. Jian Zhang [8] adopted a constrained Boltzmann Machine and the kernel-target alignment subset selection approach to identify a collection of hierarchical features from data sources. The proposed technique improved classification accuracy by more than 10%. Andreas Kolsch [9] proposed approach uses two-staged architectures for document image classification. The first stage uses a deep network to extract features while the Extreme Learning Machines (ELMs) are used for classification. The accuracy of 83.24% is achieved on the standard Tobacco-3482 dataset with a minimal training time of 1.176 sec. Muhammad [10] author has used.

As a summary, the large number of text documents are classified with good accuracy using various text-based classification methods where characters are extracted using OCR. Accuracy is based on extracted characters. Machine learning approaches such as Logistic regression, Decision tree, Support vector machine, Random Forest, etc. are used to develop a model. Drawback is training time is large and if documents contain images, then need to do feature extraction. Pre-processing is the essential task in these machine learning methods. For Deep learning, direct image is used to classify. But the drawback is it is time-consuming and needs hyper parameter tunning.

III. METHODOLOGY FOR DOCUMENT CLASSIFICATION

A. *Data collection: -*

Four different documents Voter Id, Aadhar Card, Pan Card, and Passport are used for document classification. Total 600 samples with 150 from each category are used for experimentation. These documents are collected by scraping images from the web and then are subjected to preprocessing. In the image preprocessing part augmenting scrapped images position, color, scaling, cropping, flipping, and rotation augmentation was performed on the document images. Preprocessed images are sent for document classification.70% of data is used for training and 30% is used for testing.

B. *Classification based on textual features: -*

OCR (Optical Character Recognition) is used to extract textual information from the document images. The extracted text is then used for determining TF-IDF score for individual word.

TF-IDF (Term Frequency Inverse Document Frequency) is determined by the product of the relative frequency of words in a specific document and the inverse document frequency of that word over the entire document corpus. Terms that are used frequently in a single or small set of documents have a higher TF-IDF score than common words like articles and prepositions. Calculate TF (Term Frequency), which is the ratio of the number of times a word appears in a document to the total number of words, given a document collection D, a word w, and an individual document d ϵ D.

$$f_{w,d} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad [1]$$

IDF (Inverse Document Frequency) is the logarithm of total number of documents (D) divided by number of documents that contain the word w. Thus TF-IDF is given by: -

$$IDF = \log\left(\frac{|D|}{f_{w,D}}\right) \quad [2]$$

$$TF - IDF = W_d = f_{w,d} * \log\left(\frac{|D|}{f_{w,D}}\right) \quad [3]$$

Where, d represents a single document, fw,d equals the number of times w appears in d, |D| is the size of the corpus and W_d is TF-IDF score of a word in a document.

If W_d is smaller but still positive then the word is relatively common over the entire corpus but still holds some importance throughout D. Common words like articles, prepositions receive a very low TF-IDF score. Suppose $f_{w,D}$ is large and $f_{w,D}$ is small, then log ($|D|/ f_{w,D}$) will be rather large, and so Wd will likewise be large. Words with high W_d imply that W is an important word in d but not common in D. This word is said to have a large discriminatory power.

Words with higher TF-IDF score are compared to the keys present in the dictionary. A dictionary can be a simple Python dictionary with key-value pairs. Corresponding values of the key(words with higher TF-IDF score) denote the class to which document belongs. A mean of category to which 10 keywords from the document belongs is a class of given document.

C. *Classification based on visual features:*

Some documents contain visual features that convey a lot of information. Such visual features cannot be covered by text classification methods. Convolution models are used to deal with such visual information. Training time and Hyperparameter tuning of CNN are high. In this work pre-trained networks VGG-16 and YOLO are used.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

a) VGG-16:

VGG-16 is a 16-layer Convolution Neural Network. Document classification refers to classifying thousands of images. The input image to the VGG-16 model is 224x224 RGB image. Therefore, an image of any random size is first converted into an image of size 224x224. Fixed size 224 x 224 RGB image is given as input to cov1 layer. The image is processed through a series of convolutional layers (CLs), with the filters set to a very narrow receptive fields. In one of the setups, it additionally uses 1x1 convolution filters. This is equivalent to a linear transformation of the input channels (followed by non-linearity). The convolution stride is set to 1 pixel, and the spatial padding of CL input is set to 1 pixel for 3x3 CLs so that the spatial resolution is kept after convolution. Five max-pooling layers, which follow part of the CLs, do spatial pooling (not all the CLs are followed by max-pooling). Max-pooling is accomplished with stride 2 over a 2x2 pixel window. A stack of CLs is followed by three Fully-Connected (FC) layers. Softmax layer is the final layer of the network. VGG-16 has been used in conjunction with the soft-max layer in this paper. One of the four groups is the Softmax layer outputs.

b) YOLO V3:

YOLO, "You Look Only Once" is a neural network capable of detecting the important features of an image. Yolo requires only one pass to detect these features. It provides bounding boxes for the identified objects and can detect numerous objects at once. The capacity of YOLO to do all of the detections in one shot is its most notable feature, which explains why it is so quick and efficient. It functions by conducting a regression, in which it predicts the bounding boxes and confidence score for each using only one network session (hence the name). Other methods typically include a pipeline of activities, such as sending the image through classifiers to detect features of various locations and/or incorporating additional procedures. YOLOv2 adds a few new features, the most notable of which are anchor boxes (pre-determined sets of boxes such that the network moves from predicting the bounding boxes to predicting the offsets from these.) and the use of finer-grained characteristics to better predict smaller things. Furthermore, YOLOv2 generalizes better across image sizes because it employs a method that resizes images at random. When utilizing a neural network like YOLO to predict several items in a photo, the network makes hundreds of guesses and only shows the ones that have a higher level of confidence in the location of the object. Calculate which object's bounding box has the biggest overlap divided by non-overlap for each anchor box to get Intersection Over Union (IOU).

The anchor box recognizes the object that provided the highest IOU. An anchor is assigned a positive label if it has the highest IOU or an IOU over 0.7. On the other hand, if IOU is lower than 0.3 then an anchor is assigned a negative label.

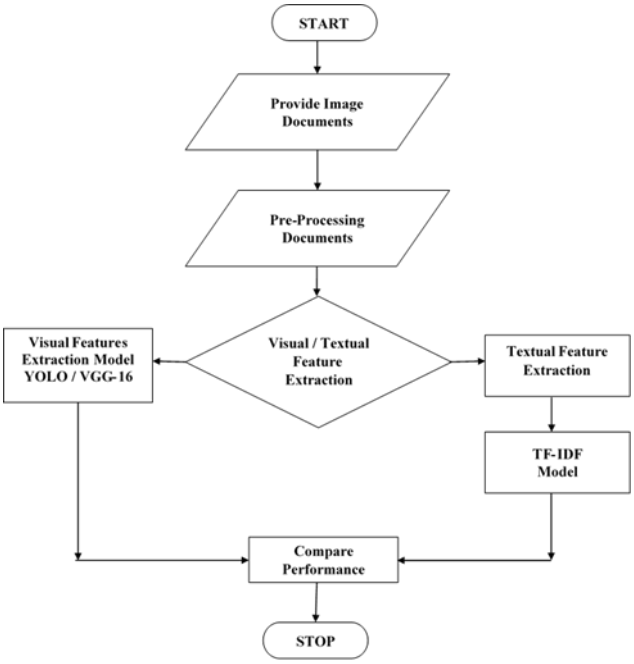

Fig. 2. Detailed Flow Chart of the Implemented System

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In the present work, TF-IDF and Deep Convolution Network has been implemented for document image classification. We tested document image classification methods on the collection of 600 documents. These documents are preprocessed and then converted into the required format. For document image classification, textual classification method (TF-IDF) and visual classification models (VGG-16 and YOLO) are implemented and compared to find out the best suitable one. The classification performance is evaluated using two majors, accuracy and confusion matrix.

TABLE 1: RESULT OF TF-IDF, YOLO AND VGG-16

Document Image	Accuracy (%)		
	TF-IDF	YOLO	VGG-16
Aadhar Card	88	88	90
Pan Card	84	92	94
Voter Id	86	92	94
Passport	84	90	96
Average Accuracy (%)	**85.5**	**90.5**	**93**

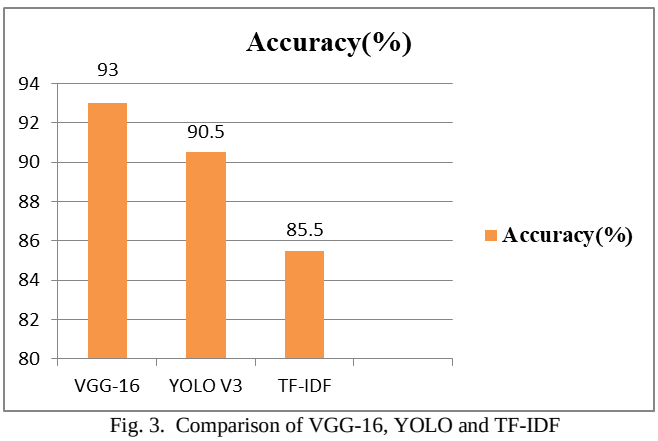

Fig. 3. Comparison of VGG-16, YOLO and TF-IDF

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Comparison of accuracy for three models is depicted in the Figure 3. Also due to the blurriness of images, there is a high probability that OCR might extract vague keywords and eventually misclassify images. Thus, for unclear and blurry images model did not give expected results. In addition, the word limit may be a barrier to text classification. If the same word that classifies a given document appears in any other document type, then it wrongly classifies the latter document which contributes to the false-positive rate. But the important merit of Image classification using the textual feature is its low computational complexity and minimal memory requirement. For document images containing limited textual information such as Passport and Voter id, the classification accuracy drops significantly. The presence of abundant common words and the absence of unique words reduces the TF-IDF score. If an image is unclear in the area where unique words are present then the classification accuracy scores are affected. This resulted in less average accuracy of 85.5%.

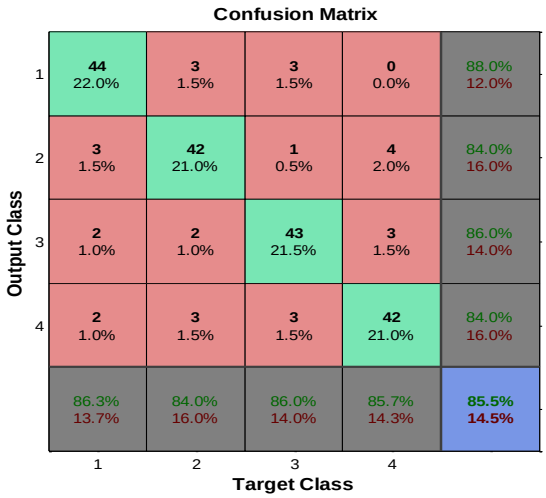

Fig. 4. Confusion matrix of TF-IDF

In the confusion matrix of Figure 4 column 1-Adhar Card, 2-PAN Card, 3-Voter ID, 4-Passport.

Figure 4 shows the confusion matrix of TF-IDF. This technique gives average accuracy of 85.5%. Aadhar card classification accuracy is 88%, Pan Card classification accuracy is 84%, Voter Id classification accuracy is 86%, and Passport classification accuracy is 84%. Although the textual feature classification model's true positive rate is good but at the same time, the false positive rate is very high.

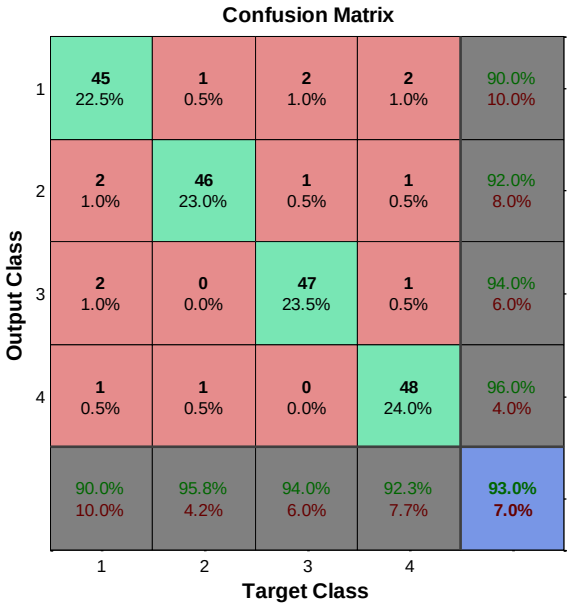

Fig. 5. shows the confusion matrix of YOLO Net.

YOLO gives an overall accuracy of 90.5%. In addition, Aadhar card classification accuracy of 88%, Pan Card classification accuracy of 92%, Voter Id classification accuracy of 92%, and Passport classification accuracy of 90% is achieved

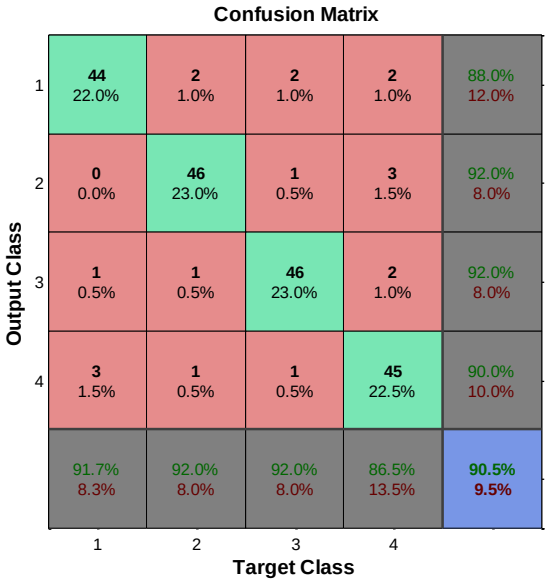

Fig. 6. Confusion matrix of YOLO

Figure 6 shows confusion matrix of VGG-16. This technique gives overall accuracy of 93%. Whilst Passport has the highest classification accuracy of 96%, Aadhar card has a classification accuracy of 90%, Pan Card has a classification accuracy of 92%, and Voter Id has a classification accuracy of 94%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

V. CONCLUSION

Different methods have been implemented in this paper for document image classification. The traditional TF-IDF classification is based on words and sentences present in document images and if all text is not extracted properly then it results in misclassification of the documents. This leads to less classification accuracy. Average classification accuracy of 85.5 % is achieved via TF-IDF. However, textual clues are equally important as compared to visual clues. In this paper CNN based architectures namely VGG-16 and YOLO are used for document image classification. It has been observed that CNN models like VGG-16 and YOLO provides better results than TF-IDF for a given dataset. In the case of document image classification using visual features, VGG-16 gives overall accuracy of 93% and the YOLO algorithm gives 90.5%. In addition, these CNN models can perform document classification tasks well even with incomplete information. Further the accuracy of VGG-16 and YOLO models can be improved by increasing dataset size. Besides their advantages, VGG-16 and YOLO do possess some demerits in the form of long training time, high-end GPU requirements, and the requirement of a large dataset to increase training accuracy.

REFERENCE

[1] Sango-Woon Kim and Joon-Min Gil, “Research paper classification systems based on TF-IDF and LDF schemes”, in Human-centric Computing and Information Sciences, 2019

[2] Shahzad Qaiser and Ramsha Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents”, in International Journal of Computer Applications, 2018.

[3] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann, “Convolutional Neural Networks for Document Image Classification”, in International Conference on Pattern Recognition (ICPR), 2014.

[4] Lucia Noce and Ignazio Gallo, “Document Image Classification Combining Textual and Visual Features”, in University of Insubria, Department of Theoretical and Applied Science (DiSTA), December 2016.

[5] Alessandro Zamberletti, Alessandro Calefati, Lucia Noce and Ignazio Gallo, “Embedded Textual Content for Document Image Classification with Convolutional Neural Networks”.

[6] Fusheng Wei, Han Qin, Shi Ye, Haozhen Zhao, “Empirical Study of Deep Learning for Text Classification in Legal Document Review”, in IEEE International Conference on Big Data (Big Data), 2018

[7] Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. erber, Laura E. Barnes, “HDLTex: Hierarchical Deep Learning for Text Classification”, in 16th IEEE International Conference on Machine Learning and Applications, 2017

[8] Jian Zhang, “Deep Transfer Learning via Restricted Boltzmann Machine for Document Classification”, in 10th International Conference on Machine Learning and Applications and Workshops, 2011

[9] Andreas Kölsch, Muhammad Zeshan Afzal , Markus Ebbecke , Marcus Liwicki, “Real-Time Document Image Classification Using Deep CNN and Extreme Learning Machines”, in 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017

[10] Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, Marcus Liwicki, “Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification”, in 14th IAPR International Conference on Document Analysis and Recognition, 2017

[11] C. Shin and D. S. Doermann, “Document image retrieval based on layout structural similarity”, In International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), 2006

[12] R. Johnson and T. Zhang, “Effective use of word order for text categorization with convolutional Neural networks”, in Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2015.

[13] L. Noce, I. Gallo, and A. Zamberletti, “Combining textual and visual features to identify anomalous user-generated content”, in International Journal of Computational Linguistics and Applications (IJCLA), 2015.

[14] A. W. Harley, A. Ufkes, and K. G. Derpanis, “Evaluation of deep convolutional nets for document image classification and retrieval”, in International Conference on Document Analysis and Recognition (ICDAR), 2015.

[15] S. Liu and W. Deng, “Very deep convolutional neural network-based image classification using small training sample size”, in *3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur*, 2015.

[16] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI*, 2017.

17] F. Cesarini, M. Lastri, S. Marinai, and G. Soda, “Encoding of modified x-y trees for document Classification”, in *International Conference Document Analysis and Recognition (ICDAR)*, 2001.

[18] Narit Hnoohom, Sumeth Yuenyong, “Classification of Dhamma Esan Characters by Transfer Learning of a Deep Neural Network”, in *15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2018