EasyChair Preprint

№ 5612

Automatic Document Verification and Government Policy Recommendation System.

Spoorti Kulkarni, Shreya Madge, Tejaswi Madhave and S.P. Kosbatwar

May 26, 2021

Automatic Document Verification and Government Policy Recommendation System

Spoorti Kulkarni
BE Computer, SKNCOE
spoo.123k@gmail.com

Shreya Madge
BE Computer, SKNCOE
shreya.madge8@gmail.com

Tejaswi Madhave
BE Computer, SKNCOE
tejaswim19@gmail.com

DR. S.P.Kosbatwar
Professor, SKNCOE
spkosbatwar@sinhgad.edu

ABSTRACT
The ministries of the Government of India have come up with various government programs called schemes (Yojana) from time to time for the welfare and development of the citizen of India. These schemes require a plethora of identity documents. When these documents go for verification in the government offices, they are verified manually for their genuineness. This process consumes a lot of time and even manpower. Then some people are even unaware of the schemes they are eligible for. Hence in this paper, a model has been proposed that will verify the authenticity of the documents and will recommend all the schemes for which the person is eligible. This uses an image processing algorithm for the verification of documents, along with that a machine learning algorithm for the recommendation of policy.

Keywords
Image processing, Machine Learning(ML), OCR, DAR.

1. INTRODUCTION
There are many documents available for an individual, it helps in the identification of that person as a citizen of that country. The government of India has designed many schemes (Yojanas) for the benefit of the citizen. Depending on the financial condition, caste, gender, etc, there are different schemes available. To avail of the benefits from these schemes, an individual has to have the related documents. In India, we have a very little automation system, as most of the work is done manually. Hence to verify the authenticity of these documents we use manual verification.

The system proposed in this paper enables automatic verification of the documents using image processing. There is a vast range of applications of image processing. We can extract suitable information from the image using image processing algorithms. In this system, we extract the symbols which prove the document to be true with IPA (image processing algorithm). Then we can compare the symbols with the original symbol and prove the document to be real or fake.

Once the document is verified we move to the next part that is the recommendation of the schemes for which the user is eligible. The machine learning algorithm is used for the same. A content-based recommendation model of machine learning is used which is a self-learning system. This suggests the best schemes available for the user.

2. RELATED WORKS
The proposed system requires work related to document analysis and image segmentation. Related work has been done on an image by detecting skews in images. This skew detection technique only considers printed text on an image by eliminating noise [1].

Other work in the area of web information [2] processing which extracts meaningful information from documents. Semantic pattern approach and ontology-based approach is used for this purpose.

In case of document verification process the verification of these documents hosted on third party server. In such cases the trustworthiness of document s hosted on such server is questionable. To overcome these issues and design a foolproof system, this paper [3] illustrate an online document verification system based on Attribute Based Encryption (ABE)

The paper [4] proposes a standard of unique document identifier (UID) to every key document that is issued by government and improves the verification mechanism with security, Verifiability, usability. These can be referred as digitized key documents which are made available for reuse anytime and anywhere

From character recognition point of view only 78 character classes are sufficient for the identification of these characters. But in Devnagari the characters fuse with each other, which result in segmentation errors. Therefore to avoid such errors there is need of such a compound characters as separate recognizable units. But it is very difficult to handle such a large number of classes, therefore it has further optimized the character class count. Thus this research work illustrate that the first 100 classes can contribute to 98.0898 of the overall recognition.[5]

Another work in this area of work is to make segmentation accurate and faster for processing of large number of

Devnagari document image using parallel implementation of algorithm on Graph Processing Unit (GPU).This algorithm is necessary for Optical Character Recognition(OCR) system to perform operations on document images such as pre-processing, segmentation, extracting features, training-testing of classifiers and post processing.[6]

WISDOM++ is an intelligent document processing system that transforms a paper document into HTML/XML format. This paper illustrate the pattern recognition technique used in document analysis and recognition (DAR)[7]

Using image processing in that extraction of interested region is possible using neural network and pattern recognition technique of machine learning that is used for document verification. In this image obtained is preprocessed and by cropping and detecting edges the gray scale image is obtained and result is compared with original image for its verification.[8]

In content based recommendation system there is use of addition information about the user or item to make prediction. The content based methods are similar to classical machine learning, in the sense that system is build based upon features in which users are interested and use of that to make prediction. In this system input will be features of user and items. Thus output of system will be prediction of whether or not user would like or dislike the item.[9]

3. PROPOSED APPROACH

The system has a two-stepped approach consisting of Image Processing and Machine Learning.

3.1 Image Processing

Document verification is traditionally done by government officials. They look for certain official signatures and stamps to infer the truthfulness of the document. Government documents also have specific boundaries, font, and patterns. These all features are very useful in the verification process.

3.1.1 Brief of the algorithm:

a. Get the image of the document using any possible method (e.g Camera, Scanner etc)
b. Use pre-processing to change the nature of the image and extract the required information.
c. Use cropping to detect the boundaries and extracts the ROI (Region of Interest).Convert image to grayscale and apply simple edge filter, average filter, Laplacian filter.
d. Compare the converted image with the original image by using Neural Networks and Pattern Recognition Tool in Mat lab.

3.1.2 Detailed description of Algorithm:

The aim is to have such an algorithm that will have good efficiency and fewer steps.

1) Image obtaining process:

Image of the document can be procured in many ways such as basic image capturing in the camera. Other ways are by scanning the document using a scanner or the scanning apps present in the mobile phones.

2) Pre-processing operations:

Pre-processing operations alter the nature of the image, which helps in the extraction of features in an easier way. We will be using, blurring, grayscale conversion, thresholding, noise removal using filters. This helps in boundary detection and cropping of the ROI.

3) Detection of Boundary:

To detect boundary, we require a black and white image i.e binary image. We simply separate the background and foreground of the image and separate the ROI(Region of Interest).

4) Extraction of Important Features:

From the binary image, we find out the dimensions of the symbols on documents and find out the aspect ratio. Then we compare the aspect ratio of the image with the original aspect ratio of the original document. We then use signature verification, this compares the signature of both the documents.

5) Comparing the result:

After obtaining the value, it is compared with the value of the original image by using Neural Networks and Pattern Recognition Tool in Mat lab.

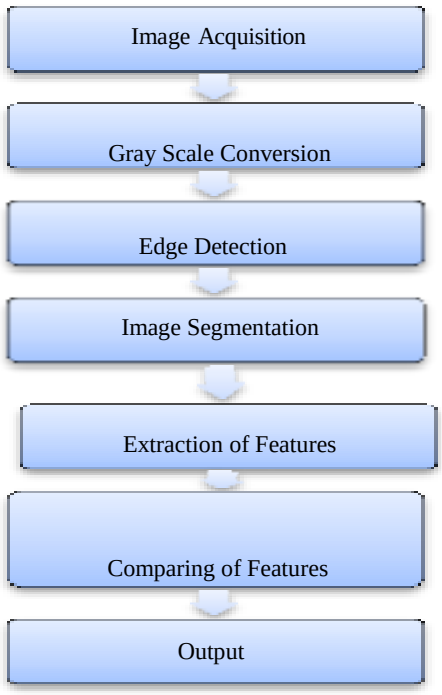

Fig 1: Flow diagram of Image processing algorithm

3.2 Machine Learning

In the proposed system we will be using a content-based recommendation system that recommends the policy to the user based upon a description of policies and profile of the user's interest.

3.2.1 Data Representation:

The data used for prediction need to be stored in a database table which contains columns of attributes/characteristics/features of policies or user. This data used for prediction is structured data that can be stored in the table.

3.2.2 User Profile:

A profile of the user's interest or information is used by most of the recommendation system. This profile consists of many different types of information.

This user's profile can be built by using:

1) Model of user's preferences: Description of the type of policy that interest the user

2) Customization: In user's customization, a system provides an interface that allows user to choose explicitly their interest area. Often checkboxes are provided to select the preferences.

3) Rule-based recommendation: system will have rules based on which recommendation can be done, e.g. policy with annual income value criteria, a policy with age limit, gender specific policy, cast oriented policy.

3.2.3 Identifying Attributes:

Attributes of user and policies are identified which can be served as a piece of additional information required for the content-based recommendation system.

1) User's Attribute: gender, age, cast, annual income, kinds of certificate available etc.

2) Policy attribute: farmer related policy, medical related policy, girls related, cast oriented, education related, annual income limit based policy, etc.

3.2.4 Term frequency-Inverse Document Frequency:

TF-IDF algorithm is used in information retrieval for feature extraction purposes.

Term frequency: The frequency of the word in the current document to the total words in the document. It signifies occurrence of word in a document and gives higher weight when the frequency is more, so for normalization it is divided by document length.

$$\mathrm{Tf(t)} = \frac{\text{Frequency occurrence ofterm t in document}}{\text{Total number of terms in document}}$$

Inverse Document Frequency: Total number of documents to the frequency occurrence of documents containing the word. It signifies the rarity of the word. It helps in giving a higher score to the rare words

$$\mathrm{Idf(t)} = \log_{10}\left(\frac{\text{Total Number of documents}}{\text{Number of documents containing term t}}\right)$$

$$\mathrm{TF-IDF} = \mathrm{TF}(+,\mathrm{d}) \times \mathrm{IDF}(+)$$

TF(+,d): is number of times + appears in document.

IDF(+): is logarithm of documents in which + appears.

TF-IDF is a measure used to evaluate how important a word is to a document in document corpus.

4. CONCLUSION

The government identity document is very important for a person to have his/her identity in the country. There is a chance of fraud in these documents. It can affect the identity of the individual. Due to this, the government schemes can be easily misused. This system helps to avoid this fraud and the citizen to attain the complete advantage of the government scheme.

This system is making use of the strong Image processing algorithm to detect the authenticity of the document along with neural networking and pattern recognition. Documents have specific features that can prove it's authenticity, these features are extracted with the image processing algorithm.

The Machine Learning algorithm is used to recommend the policy for which the user is eligible. Out of thousands of policies, the user is mostly never able to find the policy for which he is eligible. Also due to the negligence of the user, he may not get the benefits. Due to this algorithm, the right policies are recommended to benefit the people. A content-based recommendation model of machine learning is used to recommend government policies.

5. REFERENCES

[1] New Fast Content Based Skew Detection Algorithm for Document Images,Mohd Amir and Abhishek Jindal,Newgen Software Technologies Ltd., A-6 Satsang Vihar Marg, Qutab Institutional Area, New Delhi, 110067, India

[2] An Approach to Web Information Processing Anatoly Bobkov1, Sergey Gafurov2, Viktor Krasnoproshin1, and Herman Vissia2 1 ,Belarusian State University, Minsk, Republic of Belarus ,by Byelex Multimedia Products BV, Oud Gastel, The Netherlands

[3] Access Control and Data Security in Online Document Verification System by Ravinder Reddy B, Pavan Kumar C, Rajrupa Singh and Selvakumar R,Department of Computer Science and Engineering,Anurag Group of Institutions, Hyderabad, Telangana ravinderreddycse@cvsr.ac.in ,School of Computer Science and Engineering (SCOPE), VIT University, Vellore pavankumarc@ieee.org, School of Advanced Sciences (SAS), VIT University, Vellore rajrupa.singh,rselvakumar@vit.ac.in

[4] Optimization of Digitalized Document Verification Using e-Governance Service Delivery Platform (e-SDP) by Raghunathan.VS,Dr. V.Cyril Raj,Dr. Sumathy Eswaran,Ambika.A ,Department of Computer Science and Engineering, Dr. M.G.R

Educational Research Institute, Chennai and Sangeetha.R.U,Department of Information Technology, Easwari Engineering College, Chennai

[5] Optimizing Character Class Count for Devanagari Optical Character Recognition Jasbir Singh and Gurpreet Singh Lehal Department of Computer Science, Punjabi University, Patiala, Indiajbs.5@rediffmail.com

[6] Parallel Implementation of Devanagari Document Image Segmentation Approach on GPU 92 Brijmohan Singh, Nitin Gupta, Rashi Tyagi, Ankush Mittal, and Debashish Ghosh

[7] Symbolic Learning Techniques in Paper Document Processing

[8] Detection of counterfeit Indian Passport using image processing,Younus Ahmad Dar,"electronics and communication departs,PES college of engineering,Mandya"

[9] Content-Based Recommendation Systems,Michel J.Pazzani and Daniel Billsus,Rutgers University,ASBIII,3Rutgers Plaza New Brunswick,NJ 0890